Bernoulli **25**(3), 2019, 1603–1639
https://doi.org/10.3150/17-BEJ978

Sparse Hanson–Wright inequalities for subgaussian quadratic forms

SHUHENG ZHOU

Department of Statistics, University of California, Riverside, CA 92521, USA. E-mail: szhou@ucr.edu

In this paper, we provide a proof for the Hanson–Wright inequalities for sparse quadratic forms in subgaussian random variables. This provides useful concentration inequalities for sparse subgaussian random vectors in two ways. Let $X = (X_1, \dots, X_m) \in \mathbf{R}^m$ be a random vector with independent subgaussian components, and $\xi = (\xi_1, \dots, \xi_m) \in \{0, 1\}^m$ be independent Bernoulli random variables. We prove the large deviation bound for a sparse quadratic form of $(X \circ \xi)^T A(X \circ \xi)$, where $A \in \mathbf{R}^{m\times m}$ is an $m \times m$ matrix, and random vector $X \circ \xi$ denotes the Hadamard product of an isotropic subgaussian random vector $X \in \mathbf{R}^m$ and a random vector $\xi \in \{0, 1\}^m$ such that $(X \circ \xi)_i = X_i\xi_i$, where $\xi_1, \dots, \xi_m$ are independent Bernoulli random variables. The second type of sparsity in a quadratic form comes from the setting where we randomly sample the elements of an anisotropic subgaussian vector $Y = HX$ where $H \in \mathbf{R}^{m\times m}$ is an $m \times m$ symmetric matrix; we study the large deviation bound on the ℓ_2-norm $\|D_\xi Y\|_2^2$ from its expected value, where for a given vector $x \in \mathbf{R}^m$, $D_x = \mathrm{diag}(x)$ denotes the diagonal matrix whose main diagonal entries are the entries of x. This form arises naturally from the context of covariance estimation.

Keywords: Hanson–Wright inequality; sparse quadratic forms; subgaussian concentration

1. Introduction

In this paper, we explore the concentration of measure results for quadratic forms involving a sparse subgaussian random vector $X \in \mathbf{R}^m$. Sparsity can naturally come from the fact that the high dimensional vector $X \in \mathbf{R}^m$ is sparse, for example, when the elements of X are missing at random, or when we intentionally sparsify the vector X to speed up computation. The purpose of the paper is to prove the Hanson–Wright type of large deviation bounds for sparse quadratic forms in Theorems 1.1 and 1.2.

Sparsity comes in two forms. In Theorem 1.1, we randomly sparsify the subgaussian vector X involved in the quadratic form $X^T AX$, where $X = (X_1, \dots, X_m) \in \mathbf{R}^m$ is a random vector with independent subgaussian components, and $\xi = (\xi_1, \dots, \xi_m) \in \{0, 1\}^m$ consists of independent Bernoulli random variables. In particular, we first consider $(X \circ \xi)^T A(X \circ \xi)$, where $X \circ \xi \in \mathbf{R}^m$ denotes the Hadamard product of random vectors X and ξ such that $(X \circ \xi)_i = X_i\xi_i$ and A is an $m \times m$ matrix. The second type of sparsity comes into play when we sample the elements of an anisotropic subgaussian random vector $Y = D_0X$ where $X \in \mathbf{R}^m$ is as defined in Theorem 1.1 and $D_0 \in \mathbf{R}^{m\times m}$ is an $m \times m$ symmetric matrix.

The bound in Theorem 1.2 allows the second type of sparsity in a quadratic form in the following sense. Suppose A_0 is an $m \times m$ symmetric positive semidefinite matrix and $A_0^{1/2}$ is the unique square root of A_0. Suppose we randomly sample the rows or columns of $A_0^{1/2}$ to construct

1350-7265

a quadratic form as follows,

$$X^T A_0^{1/2} A_0^{1/2} X \to X^T A_0^{1/2} D_\xi A_0^{1/2} X. \tag{1}$$

We state in Theorem 1.2, where we replace $A_0^{1/2}$ with D_0, a symmetric $m \times m$ matrix, the large deviation bound for the sparse quadratic form on the right-hand side of (1). These questions arise naturally in the context of covariance estimation problems, where we naturally take A_0 and D_0 as symmetric positive (semi)definite matrices.

The following definitions correspond to Definitions 5.7 and 5.13 in [17]. For a random variable Z, the subgaussian (or ψ_2) norm of Z denoted by $\|Z\|_{\psi_2}$ is defined to be [17]:

$$\|Z\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\big(\mathbb{E}|Z|^p\big)^{1/p} \qquad \text{which is the smallest } K_2$$

$$\text{which satisfies} \qquad \big(\mathbb{E}|Z|^p\big)^{1/p} \leq K_2\sqrt{p} \qquad \forall p \geq 1;$$

$$\text{if } \mathbb{E}[Z]=0, \qquad \text{then } \mathbb{E}\exp(tZ) \leq \exp\big(Ct^2\|Z\|_{\psi_2}^2\big) \qquad \text{for all } t \in \mathbf{R}.$$

We use $X' \sim X$, where $X, X' \in \mathbf{R}^m$, to denote that two random vectors follow the same distribution. For a symmetric matrix $A = (a_{ij}) \in \mathbf{R}^{m\times m}$, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and the smallest eigenvalue of A, respectively. Moreover, we order the m eigenvalues algebraically and denote them by

$$\lambda_{\min}(A) = \lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_m(A) = \lambda_{\max}(A).$$

For a matrix A, the operator norm $\|A\|_2$ is defined to be $\sqrt{\lambda_{\max}(A^T A)}$. In particular, we prove the following theorem.

Theorem 1.1. *Let $X = (X_1, \dots, X_m) \in \mathbf{R}^m$ be a random vector with independent components X_i which satisfy $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let $\xi = (\xi_1, \dots, \xi_m) \in \{0, 1\}^m$ be a random vector independent of X, with independent Bernoulli random variables ξ_i such that $\mathbb{E}(\xi_i) = p_i$. Let $A = (a_{ij})$ be an $m \times m$ matrix. Then, for every $t > 0$,*

$$\begin{aligned} &\mathbb{P}\big(\big|(X\circ\xi)^T A(X\circ\xi) - \mathbb{E}(X\circ\xi)^T A(X\circ\xi)\big| > t\big) \\ &\quad \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4(\sum_{k=1}^m p_k a_{kk}^2 + \sum_{i\neq j} a_{ij}^2 p_i p_j)}, \frac{t}{K^2\|A\|_2}\right)\right), \end{aligned} \tag{2}$$

where $X\circ\xi$ denotes the Hadamard product of random vectors X and ξ such that $(X\circ\xi)_i = X_i\xi_i$.

Let ξ be as defined in Theorem 1.1. We now randomly sample entries of a correlated subgaussian random vector $Y = D_0 X$ and study the large deviation bound on the norm of $\|D_\xi Y\|_2^2$ from its expected value in Theorem 1.2, where for a given $x \in \mathbf{R}^m$, $D_x = \operatorname{diag}(x)$ denotes the diagonal matrix whose main diagonal entries are the elements of x. And we write $D_x := \operatorname{diag}(x)$

interchangeably. Partition a symmetric matrix $D_0 \in \mathbf{R}^{m\times m}$ according to its columns as $D_0 = [d_1, d_2, \ldots, d_m]$. Denote by

$$A_0 := D_0^2 = \sum_{i=1}^{m} d_i d_i^T = (a_{ij}) \succeq 0. \tag{3}$$

The bounds in Theorem 1.1 and Theorem 1.2 reduce to essentially the same type.

Theorem 1.2. *Let D_ξ be a diagonal matrix with independent elements from the random vector $\xi \in \{0,1\}^m$, where $\mathbb{E}\xi_j = p_j$, for $0 \le p_j \le 1$. Let X be as defined in Theorem 1.1, independent of ξ. Let $A_0 = (a_{ij}) = D_0^2$. Let $Y = D_0 X$. Then, for every $t > 0$,*

$$\begin{aligned}&\mathbb{P}\big(\big|Y^T D_\xi Y - \mathbb{E}Y^T D_\xi Y\big| > t\big)\\ &\quad =: \mathbb{P}\big(|S| > t\big) \le 2\exp\left(-c_2 \min\left(\frac{t^2}{K^4(\sum_{i=1}^m p_i a_{ii}^2 + \sum_{i\neq j} a_{ij}^2 p_i p_j)}, \frac{t}{K^2\|A_0\|_2}\right)\right),\end{aligned}$$

where c_2, C are some absolute constants.

To illustrate the sparse Hanson–Wright inequalities, we will consider the covariance estimation problem in the matrix variate model which we now define. A positive semidefinite matrix Σ is said to be separable if it can be written as a Kronecker product of two positive semidefinite matrices $A \in \mathbf{R}^{m\times m}$ and $B \in \mathbf{R}^{n\times n}$, for which we denote by $\Sigma = A \otimes B = (a_{ij}B)$, where $\otimes$ denotes the Kronecker product. We first work with the separable covariance model, however, now under the much more general subgaussian distribution, where we also model the sparsity in data with a random mask. Let $B_0 = (b_{ij}) \in \mathbf{R}^{n\times n}$ and $A_0 = (a_{ij}) \in \mathbf{R}^{m\times m}$ be symmetric positive definite matrices, and $B_0^{1/2}$ and $A_0^{1/2}$ be the unique square root of B_0 and A_0 respectively. We denote the $n \times m$ data matrix by

$$\mathbb{X} = \big[x^1\ x^2\ \ldots\ x^m\big] = \big[y^1\ y^2\ \ldots\ y^n\big]^T$$

with column vectors $x^1, \ldots, x^m \in \mathbf{R}^n$ and row vectors $y^1, \ldots, y^n \in \mathbf{R}^m$. Consider an $n \times m$ data matrix $\mathbb{X}$ which is generated from a random matrix $\mathbb{Z}_{n\times m} = (Z_{ij})$ as follows:

$$\mathbb{X} = B_0^{1/2} \mathbb{Z} A_0^{1/2}, \tag{4}$$

where Z_{ij} are independent subgaussian random variables with

$$\mathbb{E}Z_{ij} = 0 \quad \text{and} \quad \|Z_{ij}\|_{\psi_2} \le K \quad \text{and} \quad \mathbb{E}Z_{ij}^2 = 1 \qquad \forall i, j.$$

Suppose that we now observe for $\mathbb{X}$ as defined in (4)

$$\mathcal{X} = \mathbb{U} \circ \mathbb{X} \qquad \text{where } \mathbb{U} = \big[v^1\ v^2\ \ldots\ v^n\big]^T \in \{0,1\}^{n\times m}, \tag{5}$$

where $v^1, \dots, v^n \sim \mathbf{v} \in \{0,1\}^m$ are independent random vectors such that $\mathbf{v}$ is composed of independent Bernoulli random variables with $\mathbb{E}v_k = \zeta_k, k = 1, \dots, m$. Hence, we observe for each row vector y^i of $\mathbb{X}$: $\forall i = 1, \dots, n$,

$$v^i \circ y^i \qquad \text{where } v^i_k \sim \text{Bernoulli}(\zeta_k), \qquad \forall k = 1, \dots, m. \tag{6}$$

When $\mathbb{Z}$ is a Gaussian random ensemble with i.i.d. $N(0,1)$ entries, we say that random matrix $\mathbb{X}$ follows the matrix-variate normal distribution with a separable covariance structure:

$$\mathbb{X}_{n\times m} \sim \mathcal{N}_{n,m}(0, A_{0,m\times m} \otimes B_{0,n\times n}). \tag{7}$$

See [3,8,19] for characterization and examples. When the data (7) is observed in full, the theory is already in place on estimating matrix variate Gaussian graphical models which encode the conditional dependency structures in the precision matrices [19]. In particular, sample and penalized correlation estimators for the correlation matrix $\rho(B_0)$ and $\rho(A_0)$ can be derived from the gram matrix $\mathbb{X}\mathbb{X}^T$ and $\mathbb{X}^T\mathbb{X}$ respectively.

We exploit such similar relationships in the present work, which leads to the consideration of a set of oracle estimators which we present in Section 5. The task we will focus on in the current paper is limited to presenting the concentration of measure bounds on entries of the gram matrices $\mathcal{X}\mathcal{X}^T$ and $\mathcal{X}^T\mathcal{X}$ for the subgaussian data matrix generated from the model (4) and (5). We will show that these estimators possess excellent statistical convergence properties once the sampling rate is above a certain threshold. We leave the full-fledged development of graphical model estimation with incomplete data to a follow-up paper [21]. Indeed, beyond the above mentioned similarities in terms of using the gram matrices as the input to our estimation procedures, the theory and estimation tasks will depart significantly from the baseline model in (4) where we observe the full data matrix.

We mention without a proof the following Theorem 1.3, which is a variation upon Theorem 1.2. We use this theorem in the proof of Theorems 5.1 and 5.3. Formally, we have the following theorem.

Theorem 1.3. *Let $X = (X_1, \dots, X_m) \in \mathbf{R}^m$ be a random vector as defined in Theorem 1.1. Let $X' \sim X$, where X', X are independent. Let $\xi = (\xi_1, \dots, \xi_m) \in \{0,1\}^m$ be a random vector independent of X, X', with independent Bernoulli random variables ξ_i such that $\mathbb{E}(\xi_i) = p_i$ for $0 \le p_i \le 1$. Let D_ξ be a diagonal matrix with elements from the random vector $\xi \in \{0,1\}^m$. Partition an $m \times m$ symmetric matrix D_0 according to its columns as $D_0 = [d_1, d_2, \dots, d_m]$. Let $A_0 = (a_{ij}) = D_0^2$. Let $Y = D_0 X$ and $Y' = D_0 X'$. Then, for every $t > 0$,*

$$\mathbb{P}\left(\left|Y^T D_\xi Y'\right| > t\right) \le 2\exp\left(-c_2 \min\left(\frac{t^2}{K^4(\sum_{i=1}^m p_i a_{ii}^2 + \sum_{i\neq j} a_{ij}^2 p_i p_j)}, \frac{t}{K^2\|A_0\|_2}\right)\right),$$

where c_2, C are some absolute constants.

The proof follows from Theorem 1 in [14], where X, X' are independent and hence the intricate decoupling argument can be entirely avoided. Moreover, we will no longer bound the diagonal and the off-diagonal sums separately given that the sum is over decoupled random vectors X, X'. The part which deals with the randomness due to $\xi \in \mathbf{R}^m$ follows the same line of arguments as those in Theorem 1.2.

Before we leave this section, we also introduce the following notation. For a random variable Z, the sub-exponential (or ψ_1) norm of Z denoted by $\|Z\|_{\psi_1}$ is defined to be the smallest K_2 which satisfies

$$\big(\mathbb{E}|Z|^p\big)^{1/p} \le K_2 p \qquad \forall p \ge 1; \qquad \text{in other words}$$
$$\|Z\|_{\psi_1} = \sup_{p \ge 1} p^{-1}\big(\mathbb{E}|Z|^p\big)^{1/p}.$$

For two $m \times n$ matrices M_1, M_2, denote by $M_1 \circ M_2$ the Hadamard or Schur product, which is defined as follows:

$$(M_1 \circ M_2)_{ij} = (M_1)_{ij} \cdot (M_2)_{ij}.$$

For a matrix $A = (a_{ij})$ of size $m \times n$, let $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$ denote the maximum absolute row sum of the matrix A and $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ denote the maximum absolute column sum of the matrix. The matrix Frobenius norm is given by $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$. Let $|A|_{\max} = \max_{i,j} |a_{ij}|$ denote the elementwise max norm. Let $\operatorname{diag}(A)$ be the diagonal of A. Let $\operatorname{offd}(A)$ be the off-diagonal of A. Let $\|A\|_{\max,\mathrm{offd}} = \|\operatorname{offd}(A)\|_{\max} = \max_{i \ne j} |a_{ij}|$ denote the elementwise max norm on the off-diagonal of A, and $\|A\|_{\max,\mathrm{diag}} = \|\operatorname{diag}(A)\|_{\max} = \max_i |a_{ii}|$ denote that of the diagonal of A. Let $\operatorname{tr}(A)$ be the trace of A. For matrix A, $r(A)$ denotes the effective rank $\operatorname{tr}(A)/\|A\|_2$. We use A^{-T} as a shorthand notation for $(A^{-1})^T$. For two numbers a, b, $a \wedge b := \min(a, b)$, and $a \vee b := \max(a, b)$. We write $a \asymp b$ if $ca \le b \le Ca$ for some positive absolute constants c, C which are independent of n, m or sparsity and sampling parameters. Throughout this paper $C_0, C, C_1, c, c_1, \dots$ denote positive absolute constants whose value may change from line to line. For a vector $X \in \mathbf{R}^m$, let X_{Λ_δ} denote $(X_i)_{i \in \Lambda_\delta}$ for a set $\Lambda_\delta \subseteq [m]$.

2. Consequences and related work

In this section, we first compare with the following form of the Hanson–Wright inequality as recently derived in [14], as well as an even more closely related result in [13]. Such concentration of measure bounds were originally proved by [9,18]. The bound as stated in Theorem 2.1 is proved in [14].

Theorem 2.1 ([14]). *Let $X = (X_1, \dots, X_m) \in \mathbf{R}^m$ be a random vector with independent components X_i which satisfy $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_2} \le K$. Let A be an $m \times m$ matrix. Then, for every $t > 0$,*

$$\mathbb{P}\big(\big|X^T AX - \mathbb{E}X^T AX\big| > t\big) \le 2\exp\left(-c \min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right)\right).$$

When X is a vector whose coordinates are ± 1 Bernoulli random variables, the following lemma in the same spirit as in Theorem 1.1 is shown in [13].

Lemma 2.2 ([13]). *Let J be a random subset of $[m]$ of size $k < m$ uniformly chosen among all such subsets. Denote by $R_J = \sum_{j \in J} e_j e_j^T$ the coordinate projection on the set J. Let $Y = (\varepsilon_1, \dots, \varepsilon_m)$ be vector whose coordinates are ± 1 Bernoulli Random variables. Then for any $m \times m$ matrix A and any $t > 0$*

$$\mathbb{P}\big(\big|Y^T R_J A R_J Y - \mathbb{E} Y^T R_J A R_J Y\big| > t\big) \\ \leq 2 \exp\left(-c \min\left(\frac{t^2}{k\|A\|_2^2}, \frac{t}{\|A\|_2}\right)\right).$$

Other related results include [1,2,6,7,11,12]. We refer to [14] for a survey of these and other related results.

Clearly, the large deviation bounds in Theorems 1.1 and 1.2 are determined by the following quantity

$$\bar{M} := \sum_{i=1}^{m} p_i a_{ii}^2 + \sum_{i \neq j} a_{ij}^2 p_i p_j.$$

We now state some consequences of Theorems 1.1 and 1.2 in Corollaries 2.3 and 2.4.

Lemma 2.2 and Corollaries 2.3 and 2.4 show essentially a large deviation bound at roughly the same order given that

$$p\left\|\operatorname{diag}(A)\right\|_F^2 + p^2\left\|\operatorname{offd}(A)\right\|_F^2 \leq pm\|A\|_2^2$$

while $k\|A\|_2^2 = \frac{k}{m} m\|A\|_2^2$.

The following Corollary 2.3 follows from Theorem 1.1 immediately.

Corollary 2.3. *Let X, ξ be as defined in Theorem 1.1. Let $p_1 = p_2 = \cdots = p_m = p$. Let $A = (a_{ij})$ be an $m \times m$ matrix. Then, for every $t > 0$,*

$$\mathbb{P}\big(\big|X^T D_\xi A D_\xi X - \mathbb{E} X^T D_\xi A D_\xi X\big| > t\big) \\ \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^4(p\|\operatorname{diag}(A)\|_F^2 + p^2\|\operatorname{offd}(A)\|_F^2)}, \frac{t}{K^2\|A\|_2}\right)\right).$$

Corollary 2.4. *Let D_0, A_0, X, ξ, Y be as defined in Theorem 1.2. Let $p_1 = p_2 = \cdots = p_m = p$. Then, for every $t > 0$,*

$$\mathbb{P}\big(\big|Y^T D_\xi Y - \mathbb{E} Y^T D_\xi Y\big| > t\big) \\ = \mathbb{P}\big(\big|\|D_\xi D_0 X\|_2^2 - \mathbb{E}\|D_\xi D_0 X\|_2^2\big| > t\big) \\ \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^4(p\|\operatorname{diag}(A_0)\|_F^2 + p^2\|\operatorname{offd}(A_0)\|_F^2)}, \frac{t}{K^2\|A_0\|_2}\right)\right).$$

Corollary 2.5. *Suppose all conditions in Corollary 2.3 hold. Let $A \in \mathbf{R}^{m\times m}$ be positive semidefinite. Suppose $EX_i^2 = 1$ and*

$$\log m \|A\|_2 = o\big(p\,\mathrm{tr}(A)\big). \tag{8}$$

Then with probability at least $1 - 4/m^4$,

$$\big|X^T D_\xi A D_\xi X\big| \le p\,\mathrm{tr}(A)\big(1 + o(1)\big).$$

Proof. Define

$$S = \sum_{i,j} a_{ij}(X_i\xi_i X_j\xi_j - \mathbb{E}X_i\xi_i X_j\xi_j).$$

Thus $\mathbb{E}S = \sum_i a_{ii}\mathbb{E}X_i^2\mathbb{E}\xi_i = p\,\mathrm{tr}(A)$. We have under conditions of Theorem 1.1, with probability at least $1 - 4/m^4$, for some absolute constant C,

$$\begin{aligned}|S| &:= \big|X^T D_\xi A D_\xi X - p\,\mathrm{tr}(A)\big| \\ &\le CK^2 \log^{1/2} m\big(\log^{1/2} m\|A\|_2 + \sqrt{p}\big\|\mathrm{diag}(A)\big\|_F + p\big\|\mathrm{offd}(A)\big\|_F\big) =: t,\end{aligned}$$

where under condition (8), the deviation term is of a small order of the expected value $p\,\mathrm{tr}(A)$; that is,

$$t \asymp \log m\|A\|_2 + \log^{1/2} m\big(\sqrt{p}\big\|\mathrm{diag}(A)\big\|_F + p\|A\|_F\big) =: I + II = o\big(p\,\mathrm{tr}(A)\big).$$

To see this, notice that (8) immediately implies that the first term in t is of $o(p\,\mathrm{tr}(A))$. Now in order for the second and third term to be of $o(p\,\mathrm{tr}(A))$, we need that

$$\sqrt{p}\|A\|_F \log^{1/2} m \ll p\,\mathrm{tr}(A) \quad \text{and hence} \quad p \gg \log m\|A\|_F^2/\mathrm{tr}(A)^2$$

which is satisfied by (8) given that $\frac{\|A\|_2}{\mathrm{tr}(A)} \ge \frac{\|A\|_F^2}{\mathrm{tr}(A)^2}$, which in turn is due to $\|A\|_F^2 \le \mathrm{tr}(A)\|A\|_2$. $\square$

Corollary 2.6. *Suppose that (8) and all conditions in Corollary 2.4 hold. Assume $EX_i^2 = 1$. Then with probability at least $1 - \frac{4}{m^4}$, $|X^T D_0 D_\xi D_0 X| = p\,\mathrm{tr}(A_0)(1 + o(1))$.*

Proof. First by independence of X and ξ, we have for $\mathbb{E}X_i^2 = 1$,

$$\begin{aligned}\mathbb{E}X^T A_\xi X &= \mathbb{E}\sum_{k=1}^m X_k^2 A_{\xi,kk} = \sum_{k=1}^m \mathbb{E}\big(X_k^2\big)\mathbb{E}(A_{\xi,kk}) \\ &= \sum_{k=1}^m \mathbb{E}X_k^2\mathbb{E}\sum_{\ell=1}^m \xi_\ell d_{k\ell}^2 = \sum_{\ell=1}^m p_\ell \sum_{k=1}^m d_{k\ell}^2 = \sum_{\ell=1}^m p_\ell a_{\ell\ell}.\end{aligned}$$

We have by Corollary 2.4, with probability at least $1 - \frac{4}{m^4}$,

$$\begin{aligned}\left|X^T D_0 D_\xi D_0 X\right| &\leq \sum_{i=1}^m a_{ii} p_i + CK^2 \log^{1/2} m\left(\log^{1/2} m \|A_0\|_2 + \sqrt{\bar{M}}\right) \\ &\leq p\|D_0\|_F^2 \\ &\quad + CK^2 \log^{1/2} m\left(\log^{1/2} m\|A_0\|_2 + \sqrt{p}\left\|\mathrm{diag}(A_0)\right\|_F + p\left\|\mathrm{offd}(A_0)\right\|_F\right)\end{aligned}$$

for some absolute constants C, where $\sqrt{\bar{M}} \leq \sqrt{p}\|\mathrm{diag}(A_0)\|_F + p\|\mathrm{offd}(A_0)\|_F$. The rest of the proof for Corollary 2.6 follows from that of Corollary 2.5. □

2.1. Implications when $p_1, \ldots, p_m$ are not the same

We first need the following sharp statements about eigenvalues of a Hadamard product. See for example Theorem 5.3.4 [10].

Theorem 2.7. *Let $A, B \in \mathbf{R}^{m\times m}$ be positive semidefinite. Let $a_\infty := \max_{i=1}^m a_{ii}$ and $b_\infty := \max_{i=1}^m b_{ii}$. Any eigenvalue of $\lambda(A \circ B)$ satisfies*

$$\begin{aligned}\lambda_{\min}(A)\lambda_{\min}(B) &\leq \left(\min_{i=1}^m a_{ii}\right)\lambda_{\min}(B) \\ &\leq \lambda(A\circ B) \\ &\leq a_\infty \lambda_{\max}(B) \leq \lambda_{\max}(A)\lambda_{\max}(B).\end{aligned}$$

Corollary 2.8. *Suppose all conditions in Theorem 1.2 hold. Suppose $EX_i^2 = 1$. Let $\mathbf{p} = (p_1, \ldots, p_m)$. Let $|\mathbf{p}|_1 := \sum_{i=1}^m p_i$ and $\|\mathbf{p}\|_2^2 = \sum_{i=1}^m p_i^2$. Then with probability at lest $1 - 4/m^4$,*

$$\left|X^T D_0 D_\xi D_0 X\right| \leq \sum_{i=1}^m p_i\|d_i\|_2^2 + CK^2\log^{1/2} m\|D_0\|_2\left(\log^{1/2} m\|D_0\|_2 + 2\left(\max_i \|d_i\|_2\right)|\mathbf{p}|_1^{1/2}\right).$$

Proof. Recall $A_0 = (a_{ij}) = D_0^2 \succeq 0$. Let $a_\infty := \max_{i=1}^m a_{ii} = \max_i \|d_i\|_2^2$. Thus, we have $a_\infty \leq \|D_0\|_2^2$. Denote by $p = (p_1, \ldots, p_m)$. We have by Theorem 2.7,

$$\begin{aligned}\bar{M} = \sum_{i=1}^m p_i a_{ii}^2 + \sum_{i\neq j} a_{ij}^2 p_i p_j &\leq \sum_{i=1}^m p_i a_{ii}^2 + \mathbf{p}^T (A_0 \circ A_0)\mathbf{p} \\ &\leq a_\infty^2 |\mathbf{p}|_1 + \lambda_{\max}(A_0\circ A_0)\|\mathbf{p}\|_2^2 \\ &\leq a_\infty^2|\mathbf{p}|_1 + a_\infty\|A_0\|_2\|\mathbf{p}\|_2^2 \leq 2a_\infty\|A_0\|_2|\mathbf{p}|_1,\end{aligned}$$

where $\|\mathbf{p}\|_2^2 \leq |\mathbf{p}|_1$. The corollary thus follows immediately from Theorem 1.2. □

Remark 2.9. Assume that $p_i \geq \frac{\log m}{m}$ and hence $|\mathbf{p}|_1 \geq \log m$. Then we have $\|\mathbf{p}\|_2 \leq |\mathbf{p}|_1^{1/2} \leq |\mathbf{p}|_1$. Notice that the second term starts to dominate when $|\mathbf{p}|_1^{1/2} \gg \log m$ while the total deviation remains to be a small order of the mean $\sum_{i=1}^m p_i \|d_i\|_2^2$ so long as

$$|\mathbf{p}|_1 \gg \frac{\log m \max_k \|d_k\|_2^2 \|D_0\|_2^2}{\min_k \|d_k\|_2^4}.$$

We will use examples in Section 5 to elaborate on the lower bound immediately above.

2.2. Preliminary results

Before leaving this section, we provide some preliminary results which are used throughout the paper. We use the following properties of the Hadamard product [10],

$$A \circ xx^T = D_x A D_x \quad \text{and}$$
$$\operatorname{tr}\big(D_\xi A D_\xi A^T\big) = \xi^T (A \circ A)\xi$$

from which a simple consequence is $\operatorname{tr}(D_\xi A D_\xi) = \xi^T (A \circ I)\xi = \xi^T \operatorname{diag}(A)\xi$.

Theorem 2.10 shows a concentration of measure bound on a quadratic form with Bernoulli random variables where an explicit dependency on p_i, for all i, is shown. The setting here is different from Theorem 2.1 as we deal with a quadratic form which involves non-centered Bernoulli random variables. Theorem 2.10 is crucial in proving Theorem 1.2. The proof of Theorem 2.10 is deferred to Section 6.

Theorem 2.10. *Let $\xi = (\xi_1, \dots, \xi_m) \in \{0, 1\}^m$ be a random vector with independent Bernoulli random variables ξ_i such that $\xi_i = 1$ with probability p_i and 0 otherwise. Let $A = (a_{ij})$ be an $m \times m$ matrix. Then, for every $0 \leq \lambda \leq \frac{1}{104 \max(\|A\|_1, \|A\|_\infty)}$,*

$$\mathbb{E} \exp\left(\lambda \sum_{i,j} a_{ij}\xi_i\xi_j\right) \leq \exp\left(\lambda\left(\sum_{i=1}^m a_{ii} p_i + \sum_{i \neq j} a_{ij} p_i p_j\right)\right) * \exp\left(\frac{1}{3}\lambda \sum_{j \neq i} |a_{ij}| \sigma_i^2 \sigma_j^2\right)$$
$$* \exp\left(C_5 \lambda\left(\frac{1}{2}\sum_{i=1}^m |a_{ii}| p_i + \sum_{j \neq i} |a_{ij}| p_j p_j\right)\right),$$

where $\sigma_i^2 = p_i(1 - p_i)$ and $C_5 \leq 0.04$.

We use the following bounds throughout our paper. For any $x \in \mathbf{R}$,

$$e^x \leq 1 + x + \frac{1}{2}x^2 e^{|x|}. \tag{9}$$

We need the following result which follows from Proposition 3.4 in [15].

Lemma 2.11. *Let $A = (a_{ij})$ be an $m \times m$ matrix. Let $a_\infty := \max_i |a_{ii}|$. Let $\xi = (\xi_1, \dots, \xi_m) \in \{0,1\}^m$ be a random vector with independent Bernoulli random variables ξ_i such that $\xi_i = 1$ with probability p_i and 0 otherwise. Then for $|\lambda| \le \frac{1}{4a_\infty}$,*

$$\mathbb{E}\exp\left(\lambda \sum_{i=1}^m a_{ii}(\xi_i - p_i)\right) \le \exp\left(\frac{1}{2}\lambda^2 e^{|\lambda| a_\infty} \sum_{i=1}^m a_{ii}^2 \sigma_i^2\right),$$

where $\sigma_i^2 = p_i(1-p_i)$.

We need to state Lemma 2.12, which provides an estimate of the moment generating function for the centered sub-exponential random variable $Z_k := X_k^2 - \mathbb{E}X_k^2$ for X_k as defined in Theorem 1.1.

Lemma 2.12. *Let $X \in \mathbf{R}$ be a subgaussian random variable which satisfies $\mathbb{E}X = 0$ and $\|X\|_{\psi_2} \le K$. Let $|\tau| \le \frac{1}{23.5eK^2}$. Denote by $C_0 := 38.94$. Then*

$$\mathbb{E}\big(\exp\big(\tau\big(X^2 - \mathbb{E}X^2\big)\big)\big) \le 1 + 38.94\tau^2 K^4 \le \exp\big(C_0 \tau^2 K^4\big).$$

The proof follows essentially that of Lemma 5.15 in [17]; we provide here explicit constants.

The rest of the paper is organized as follows. In Section 2, we compare our results with those in the literature. We then prove Theorem 1.1 in Section 3 and Theorem 1.2 in Section 4. In Section 5, we provide a general theory on concentration inequalities under masks for entries of the gram matrix $\mathcal{X}^T\mathcal{X}$ and $\mathcal{X}\mathcal{X}^T$, where $\mathcal{X}$ is the observed data from the matrix variate model (cf. (5)). We prove Theorem 2.10 in Section 6. We leave certain calculations in Appendix A for the purpose of self-containment, namely, the proof of Lemmas 2.12 and 3.2.

3. Proof of Theorem 1.1

The structure of our proof follows that of Theorem 1.1 by [14]. The problem reduces to estimating the diagonal and the off-diagonal sums.

Part I: Diagonal sum. Define

$$S_0 := \sum_{k=1}^m a_{kk}\xi_k X_k^2 - \mathbb{E}\sum_{k=1}^m a_{kk}\xi_k X_k^2 \qquad \text{where } \mathbb{E}\sum_{k=1}^m a_{kk}\xi_k X_k^2 = \sum_{k=1}^m a_{kk} p_k \mathbb{E}X_k^2. \tag{10}$$

Lemma 3.1. *Let X and ξ be defined as in Theorem 1.1. Let A be an $m \times m$ matrix. Then, for every $t > 0$,*

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{k=1}^m a_{kk}\xi_k X_k^2 - \sum_{k=1}^m a_{kk} p_k \mathbb{E}X_k^2\right| > t\right) \\ &\quad \le \mathbb{P}(S_0 > t) + \mathbb{P}(S_0 < -t) \le 2\exp\left[-\frac{1}{4e}\min\left(\frac{t^2}{3K^4\sum_{k=1}^m a_{kk}^2 p_k}, \frac{t}{K^2 \max_k |a_{kk}|}\right)\right].\end{aligned}$$

We prove Lemma 3.1 after we state Lemma 3.2. For the general case where X_k are mean-zero independent sub-gaussian random variables with $\|X_k\|_{\psi_2} \le K$, we first state the following bound on the moment generating function of X_k^2.

Lemma 3.2. *Suppose that* $|\lambda| < 1/(4eK^2 \max_k |a_{kk}|)$. *Then for all* k, *we have for all* $a_{kk} \in \mathbf{R}$

$$\mathbb{E}\exp(\lambda a_{kk} X_k^2) - 1 \le \lambda a_{kk} \mathbb{E} X_k^2 + 16\lambda^2 a_{kk}^2 K^4. \tag{11}$$

Proof of Lemma 3.1. We first state some simple fact: $\max_{k=1}^m \mathbb{E}X_i^2 \le K^2$. By independence of $X_1, \dots, X_k$ and $\xi_1, \dots, \xi_k$, we bound the moment generating function of S_0 as follows: for $|\lambda| \le \frac{1}{4eK^2 \max_k |a_{kk}|}$

$$\begin{aligned}
\mathbb{E}\exp(\lambda S_0) &= \mathbb{E}\exp\left(\lambda \sum_{k=1}^m \xi_k X_k^2 a_{kk} - \lambda \sum_{k=1}^m p_k a_{kk} \mathbb{E}X_k^2\right) \\
&= \prod_{k=1}^m \left(\frac{\mathbb{E}\exp(\lambda a_{kk}\xi_k X_k^2)}{\exp(\lambda p_k a_{kk}\mathbb{E}X_k^2)}\right) = \prod_{k=1}^m \frac{\mathbb{E}_\xi \mathbb{E}_X \exp(\lambda a_{kk}\xi_k X_k^2)}{\exp(\lambda p_k a_{kk}\mathbb{E}X_k^2)} \\
&\le \prod_{k=1}^m \frac{1 + p_k(\lambda a_{kk}\mathbb{E}X_k^2 + 16\lambda^2 a_{kk}^2 K^4)}{\exp(\lambda p_k a_{kk}\mathbb{E}X_k^2)} \\
&\le \prod_{k=1}^m \frac{\exp(\lambda p_k a_{kk}\mathbb{E}X_k^2 + 16\lambda^2 p_k a_{kk}^2 K^4)}{\exp(\lambda p_k a_{kk}\mathbb{E}X_k^2)} = \exp\left(16\lambda^2 K^4 \sum_{k=1}^m p_k a_{kk}^2\right),
\end{aligned}$$

where we used (11) for the first inequality and the fact that $1 + x \le e^x$ for the second inequality. Hence for $0 < \lambda \le \frac{1}{4eK^2 \max_k |a_{kk}|}$, we have

$$\mathbb{P}(S_0 > t) \le \frac{\mathbb{E}\exp(\lambda S_0)}{e^{\lambda t}} \le \exp\left(-\lambda t + 16K^4\lambda^2 \sum_{k=1}^m p_k a_{kk}^2\right)$$

for which the optimal choice of λ is

$$\lambda = \min\left(\frac{t}{32K^4 \sum_{k=1}^m p_k a_{kk}^2}, \frac{1}{4eK^2 \max_k |a_{kk}|}\right).$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^m a_{kk}\xi_k X_k^2 - \sum_{k=1}^m a_{kk} p_k \mathbb{E}X_k^2 > t\right) \\
&\quad \le \exp\left[-\frac{1}{4e}\min\left(\frac{t^2}{3K^4 \sum_{k=1}^m p_k a_{kk}^2}, \frac{t}{K^2 \max_k |a_{kk}|}\right)\right].
\end{aligned}$$

We note that these constants have not been optimized. Repeating the arguments for $-A$ instead of A, we obtain for every $t > 0$, and for $S_0' := \sum_{k=1}^m (-a_{kk})\xi_k X_k^2 - \sum_{k=1}^m (-a_{kk}) p_k \mathbb{E} X_k^2$

$$\mathbb{P}\left(\sum_{k=1}^m a_{kk}\xi_k X_k^2 - \sum_{k=1}^m a_{kk} p_k \mathbb{E} X_k^2 < -t\right)$$
$$= \mathbb{P}\big(S_0' > t\big) \le \exp\left[-\frac{1}{4e}\min\left(\frac{t^2}{3K^4\sum_{k=1}^m p_k a_{kk}^2}, \frac{t}{K^2 \max_k |a_{kk}|}\right)\right].$$

The lemma thus holds. □

Part II: Off-diagonal sum. We now focus on bounding the off-diagonal part of the sum:

$$S_{\text{offd}} := \sum_{i\neq j}^m a_{ij} X_i X_j \xi_i \xi_j,$$

where by independence of X and ξ, $\mathbb{E} S_{\text{offd}} = \sum_{i\neq j}^m a_{ij} \mathbb{E} X_i \mathbb{E} X_j \mathbb{E}\xi_i \mathbb{E}\xi_j = 0$.

We will show that the following large deviation inequality holds for all $t > 0$,

$$\mathbb{P}\big(|S_{\text{offd}}| > t\big) \le 2\exp\left(-c\min\left(\frac{t^2}{K^4\sum_{i\neq j} a_{ij}^2 p_i p_j}, \frac{t}{K^2\|A\|_2}\right)\right). \tag{12}$$

First we prove a bound on the moment generating function for the off-diagonal sum S_{offd}. We assume without loss of generality that $K = 1$ by replacing X with X/K. Let C_4 be a constant to be specified. It holds that for all $|\lambda| \le \frac{1}{2\sqrt{C_4}\|A\|_2}$

$$\mathbb{E}\exp(\lambda S_{\text{offd}}) \le \exp\left(1.44 C_4 \lambda^2 \sum_{i\neq j} a_{ij}^2 p_i p_j\right). \tag{13}$$

Thus we have for $0 \le \lambda \le \frac{1}{2\sqrt{C_4}\|A\|_2}$ and $t > 0$,

$$\mathbb{P}(S_{\text{offd}} > t) \le \frac{\mathbb{E}\exp(\lambda S_{\text{offd}})}{e^{\lambda t}} \le \exp\left(-\lambda t + 1.44 C_4 \lambda^2 \sum_{i\neq j} p_i p_j a_{ij}^2\right).$$

Optimizing over λ, we conclude that

$$\mathbb{P}(S_{\text{offd}} > t) \le \exp\left(-c\min\left(\frac{t^2}{\sum_{i\neq j} a_{ij}^2 p_i p_j}, \frac{t}{\|A\|_2}\right)\right) =: q_1. \tag{14}$$

Repeating the arguments for $-A$ instead of A, we obtain for $S' := \sum_{i\neq j}^m (-a_{ij}) X_i X_j \xi_i \xi_j = -S_{\text{offd}}$, $0 \le \lambda \le \frac{1}{2\sqrt{C_4}\|A\|_2}$ and $t > 0$,

$$\mathbb{P}\big(S' > t\big) \le \frac{\mathbb{E}\exp(\lambda S')}{e^{\lambda t}} = \frac{\mathbb{E}\exp(-\lambda S_{\text{offd}})}{e^{\lambda t}} \le \exp\left(-\lambda t + 1.44 C_4 \lambda^2 \sum_{i\neq j} p_i p_j a_{ij}^2\right) \le q_1$$

by (13) and (14). Thus we have

$$\mathbb{P}\big(|S_{\text{offd}}| > t\big) = \mathbb{P}(S_{\text{offd}} > t) + \mathbb{P}(S_{\text{offd}} < -t) = \mathbb{P}(S_{\text{offd}} > t) + \mathbb{P}\big(S' > t\big) = 2q_1.$$

Thus (12) holds for all $t > 0$. The theorem is thus proved by summing up the bad events for diagonal sum and the non-diagonal sum while adjusting the constant c in (2).

The proof of (13) follows essentially from the decoupling and reduction arguments in [14] and thus omitted from the main body of the paper. For completeness, we include the full proof in Appendix C. See, for example, [4,5] for comprehensive discussions on modern decoupling methods. □

4. Proof of Theorem 1.2

Let X, ξ, D_0 and D_ξ be defined as in Theorem 1.2. We assume without loss of generality that $K = 1$ by replacing X with X/K. Denote by $\xi = (\xi_1, \dots, \xi_m) \in \{0,1\}^m$ a random vector with independent Bernoulli random variables ξ_i such that $\xi_i = 1$ with probability p_i and 0 otherwise.

We will bound the diagonal and the off-diagonal sums separately. Let $D_0 = [d_1, d_2, \dots, d_m]$ be a symmetric matrix. Recall that we need to estimate

$$q := \mathbb{P}\big(\big|X^T A_\xi X - \mathbb{E} X^T A_\xi X\big| > t\big) \qquad \text{where } A_\xi = D_0 D_\xi D_0 =: (\widetilde{a}_{ij}).$$

We first separate the diagonal sum from the off-diagonal sum as follows:

$$\begin{aligned}\big|X^T A_\xi X - \mathbb{E} X^T A_\xi X\big| &\le \Big|\sum_{i \ne j} X_i X_j A_{\xi,ij}\Big| + \Big|\sum_{k=1}^m X_k^2 A_{\xi,kk} - \mathbb{E}\big(X_k^2\big)\mathbb{E}(A_{\xi,kk})\Big| \\ &=: |S_{\text{offd}}| + |S_{\text{diag}}|,\end{aligned}$$

where S_{offd} and S_{diag} denote the following random variables:

$$\begin{aligned} S_{\text{offd}} &:= \sum_{i \ne j} X_i X_j A_{\xi,ij} = \sum_{i \ne j} X_i X_j \widetilde{a}_{ij} \quad \text{and} \\ S_{\text{diag}} &:= \sum_{k=1}^m X_k^2 A_{\xi,kk} - \mathbb{E}\big(X_k^2\big)\mathbb{E}(A_{\xi,kk}).\end{aligned}$$

To prove Lemma 4.5, we need the following bounds on moment generating functions for the diagonal sum in S_{diag} in Lemma 4.1 and the off-diagonal sum S_{offd} in Lemma 4.4. Let $A_0 = D_0^2 = (a_{ij}) \succeq 0$. The constants in the expression for N (and M) are not being optimized:

$$N = 82 \sum_{i=1}^m a_{ii}^2 p_i + 108 \sum_{i \ne j} a_{ij}^2 p_i p_j. \tag{15}$$

Lemma 4.1. *For all* $|\lambda| \le \frac{1}{128\|A_0\|_2}$,

$$\mathbb{E}\exp(\lambda S_{\mathrm{diag}}) \le \exp(\lambda^2 N) \quad \textit{and} \quad \mathbb{E}\exp(-\lambda S_{\mathrm{diag}}) \le \exp(\lambda^2 N).$$

To prove Lemma 4.1, first we write $S_{\mathrm{diag}} = S_0 + S_\star$ where

$$S_0 := \sum_{k=1}^m \big(X_k^2 - \mathbb{E}(X_k^2)\big) A_{\xi,kk} = \sum_{k=1}^m \big(X_k^2 - \mathbb{E}(X_k^2)\big)\left(\sum_{\ell=1}^m d_{k\ell}^2 \xi_\ell\right), \tag{16}$$

$$S_\star := \sum_{k=1}^m \mathbb{E}(X_k^2) A_{\xi,kk} - \mathbb{E}(X_k^2)\mathbb{E}(A_{\xi,kk}) = \sum_{k=1}^m \mathbb{E}(X_k^2)\left(\sum_{\ell=1}^m d_{k\ell}^2(\xi_\ell - \mathbb{E}\xi_\ell)\right), \tag{17}$$

where recall

$$A_\xi = D_0 D_\xi D_0 \qquad \text{where } D_0 = [d_1, \dots, d_m]. \tag{18}$$

We now state the following bounds on the moment generating functions of S_0 and $S_\star$ in Lemmas 4.2 and 4.3, respectively. The estimate on the moment generating function stated in Lemma 4.1 then follows immediately from the Cauchy–Schwarz inequality, in view of Lemmas 4.2 and 4.3.

Lemma 4.2. *Let* $a_{ii} = \|d_i\|_2^2$ *for* d_i *as defined in* (18). *Let* $a_\infty = \max_i \|d_i\|_2^2$. *Then for* $|\lambda| < \frac{1}{4a_\infty}$,

$$\mathbb{E}\exp(\lambda S_\star) \le \exp\left(\frac{1}{2}\lambda^2 e^{|\lambda| a_\infty} \sum_{i=1}^m a_{ii}^2 \sigma_i^2\right) \qquad \textit{where } \mathbb{E}(X_k^2) \le \|X_k\|_{\psi_2} = 1.$$

Proof. We have by independence of X and ξ and by definition of $S_\star$ in (17)

$$S_\star = \sum_{k=1}^m \mathbb{E}(X_k^2) \sum_{i=1}^m d_{ki}^2(\xi_i - p_i) = \sum_{i=1}^m \left(\sum_{k=1}^m \mathbb{E}(X_k^2) d_{ki}^2\right)(\xi_i - p_i) =: \sum_{i=1}^m a'_{ii}(\xi_i - p_i),$$

where by assumption, we have $\mathbb{E}(X_k^2) \le \|X_k\|_{\psi_2} \le K = 1$ and hence

$$0 \le a'_{ii} := \sum_{k=1}^m \mathbb{E}(X_k^2) d_{ki}^2 \le a_{ii} \quad \text{and thus} \quad \max_i |a'_{ii}| \le a_\infty.$$

The bound on the mgf of $S_\star$ follows from Lemma 2.11. For $|\lambda| < \frac{1}{4a_\infty}$, we have

$$\mathbb{E}\exp(\lambda S_\star) = \mathbb{E}\exp\left(\lambda \sum_{i=1}^m a'_{ii}(\xi_i - p_i)\right) \le \exp\left(\frac{1}{2}\lambda^2 e^{|\lambda| a_\infty} \sum_{i=1}^m (a'_{ii})^2 \sigma_i^2\right)$$
$$\le \exp\left(\frac{1}{2}\lambda^2 e^{|\lambda| a_\infty} \sum_{i=1}^m a_{ii}^2 \sigma_i^2\right).$$

□

Lemma 4.3. *Denote by* $a_{ij} = \langle d_i, d_j \rangle$ *for all* $i \neq j$ *and* $a_{ii} = \|d_i\|_2^2$ *for* d_i *as defined in* (18). *Denote by* $a_\infty := \max_i a_{ii}$. *Let* $C_0 = 38.94$. *Then for* $|\lambda| \leq \frac{1}{64\|A_0\|_2} \leq \frac{1}{64 a_\infty}$,

$$\mathbb{E}\exp(\lambda S_0) \leq \exp\left(\lambda^2\left(40\sum_{j=1}^m p_j a_{jj}^2 + 54\sum_{i\neq j} p_i p_j a_{ij}^2\right)\right). \tag{19}$$

Lemma 4.4. *Let* $A_0 = (a_{ij}) = D_0^2$. *For all* $|\lambda| \leq \frac{1}{58C\|A_0\|_2}$ *for some constant* C

$$\begin{aligned}\mathbb{E}\exp(\lambda S_{\mathrm{offd}}) &\leq \mathbb{E}\exp\big(\lambda^2 C_2 \xi^T (A_0 \circ A_0)\xi\big) \leq \exp\big(\lambda^2 M\big),\\ \mathbb{E}\exp(-\lambda S_{\mathrm{offd}}) &\leq \exp\big(\lambda^2 M\big),\end{aligned}$$

where $C_2 = 32C^2$ *and* $M = 11C^2(3\sum_{i=1}^m p_i a_{ii}^2 + 4\sum_{i\neq j} a_{ij}^2 p_i p_j)$.

We defer the proof of Lemma 4.4 to Section 4.2 and Lemma 4.3 to Section 4.1. We are now ready to state the large deviation inequalities for the diagonal sum S_{diag}, followed by that for the off-diagonal sum S_{offd}.

Lemma 4.5. *Let* $A_0 = (a_{ij}) = D_0^2$. *For all* $t > 0$ *and* N *as defined in* (15),

$$\mathbb{P}\big(|S_{\mathrm{diag}}| > t/2\big) \leq 2\exp\left(-\frac{1}{16}\min\left(\frac{t^2}{N}, \frac{t}{32\|A_0\|_2}\right)\right).$$

For the off-diagonal sum, we now state the following large deviation bound as in Lemma 4.6.

Lemma 4.6. *Suppose all conditions in Lemma* 4.4 *hold. For all* $t > 0$, *and some large enough absolute constant* C,

$$\mathbb{P}\big(|S_{\mathrm{offd}}| > t/2\big) \leq 2\exp\left(-\frac{1}{16}\min\left(\frac{t^2}{M}, \frac{t}{15C\|A_0\|_2}\right)\right),$$

where $M = 11C^2(3\sum_{i=1}^m p_i a_{ii}^2 + 4\sum_{i\neq j} a_{ij}^2 p_i p_j)$.

The theorem is thus proved by summing up the two bad events:

$$q = \mathbb{P}\big(|S_{\mathrm{diag}} + S_{\mathrm{offd}}| > t\big) \leq \mathbb{P}\big(|S_{\mathrm{offd}}| > t/2\big) + \mathbb{P}\big(|S_{\mathrm{diag}}| > t/2\big)$$

while adjusting the constant c in (2).

It remains to prove Lemmas 4.1, 4.5 and 4.6.

Proof of Lemma 4.1. Suppose that $|\lambda| \leq \frac{1}{128\|A_0\|_2}$. By Lemmas 4.3 and 4.2,

$$\mathbb{E}^{1/2}\exp(2\lambda S_\star) \leq \exp\left(\lambda^2 e^{2|\lambda| a_\infty}\sum_{j=1}^m \sigma_j^2 a_{jj}^2\right),$$

$$\mathbb{E}^{1/2}\exp(2\lambda S_0) \le \exp\left(80\lambda^2 \sum_{j=1}^{m} p_j a_{jj}^2\right)\exp\left(108\lambda^2 \sum_{i\neq j} a_{ij}^2 p_i p_j\right).$$

Now we have by the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathbb{E}\exp(\lambda S_{\text{diag}}) &= \mathbb{E}\exp\bigl(\lambda(S_0 + S_\star)\bigr) \le \mathbb{E}^{1/2}\exp(2\lambda S_0)\mathbb{E}^{1/2}\exp(2\lambda S_\star)\\ &\le \exp\left(82\lambda^2 \sum_{j=1}^{m} \sigma_j^2 a_{jj}^2\right)\exp\left(108\lambda^2 \sum_{i\neq j} a_{ij}^2 p_i p_j\right).\end{aligned}$$

□

Proof of Lemma 4.5. Lemma 4.5 follows from Lemma 4.1 immediately. Let $\mathbb{E}_X$ and $\mathbb{E}_\xi$ denote the expectation with respect to random variables in vectors X and ξ, respectively.

First, by the Markov's inequality, we have for $0 < \lambda \le \frac{1}{128\|A_0\|_2}$

$$\begin{aligned}\mathbb{P}(S_{\text{diag}} > t/2) &= \mathbb{P}(\lambda S_{\text{diag}} > \lambda t/2) = \mathbb{P}\bigl(\exp(\lambda S_{\text{diag}}) > \exp(\lambda t/2)\bigr)\\ &\le \frac{\mathbb{E}\exp(\lambda S_{\text{diag}})}{e^{\lambda t/2}} \le \exp\bigl(-\lambda t/2 + N\lambda^2\bigr).\end{aligned}$$

Optimizing over λ, for which the optimal choice of λ is $\lambda = \frac{t}{4N}$. Thus, we have for $t > 0$,

$$\begin{aligned}\mathbb{P}(S_{\text{diag}} > t/2) &\le \exp\left(-\min\left(\frac{t^2}{16N}, \frac{t}{4 * 128\|A_0\|_2}\right)\right)\\ &\le \exp\left(-\frac{1}{16}\min\left(\frac{t^2}{N}, \frac{t}{32\|A_0\|_2}\right)\right) =: q_d.\end{aligned}$$

Repeating the argument for $-A_\xi$ instead of A_ξ, we now consider

$$S'_{\text{diag}} := \sum_{k=1}^{m}\bigl(X_k^2(-A_{\xi,kk}) + \mathbb{E}\bigl(X_k^2\bigr)\mathbb{E}(A_{\xi,kk})\bigr) = -S_{\text{diag}}.$$

By Lemma 4.1, we have for all $|\lambda| \le \frac{1}{128\|A_0\|_2}$

$$\mathbb{E}\exp\bigl(\lambda S'_{\text{diag}}\bigr) = \mathbb{E}\exp(-\lambda S_{\text{diag}}) \le \exp\bigl(\lambda^2 N\bigr).$$

Thus, we have for $t > 0$ and $0 < \lambda \le \frac{1}{128\|A_0\|_2}$,

$$\mathbb{P}\bigl(S'_{\text{diag}} > t/2\bigr) \le \frac{\mathbb{E}\exp(\lambda S'_{\text{diag}})}{e^{\lambda t/2}} \le \exp\bigl(-\lambda t/2 + N\lambda^2\bigr) \le q_d.$$

The lemma is thus proved, given that for $t > 0$

$$\begin{aligned}&\mathbb{P}(S_{\text{diag}} < -t/2) = \mathbb{P}\bigl(S'_{\text{diag}} > t/2\bigr) \le q_d,\\ &\mathbb{P}\bigl(|S_{\text{diag}}| > t/2\bigr) = \mathbb{P}(S_{\text{diag}} > t/2) + \mathbb{P}(S_{\text{diag}} < -t/2) \le 2q_d.\end{aligned}$$

□

Proof. Proof of Lemma 4.6 Lemma 4.6 follows immediately from Lemma 4.4. We have for $0 < \lambda \le \frac{1}{58C\|A_0\|_2}$ and $S := S_{\mathrm{offd}}$,

$$\mathbb{P}(S > t/2) = \mathbb{P}\big(\exp(\lambda S) > \exp(\lambda t/2)\big) \le \frac{\mathbb{E}\exp(\lambda S)}{e^{\lambda t/2}} \le \exp\big(-\lambda t/2 + M\lambda^2\big)$$

for which the optimal choice of λ is $\lambda = \frac{t}{4M}$. Thus we have for $t > 0$,

$$\begin{aligned}\mathbb{P}(S > t/2) &\le \exp\big(-\lambda t/2 + M\lambda^2\big)\\ &\le \exp\left(-\frac{1}{16}\min\left(\frac{t^2}{M}, \frac{t}{15C\|A_0\|_2}\right)\right) =: q_{\mathrm{offd}}.\end{aligned}$$

Similarly, we have for $\lambda, t > 0$,

$$\begin{aligned}\mathbb{P}(S < -t/2) &= \mathbb{P}(-S > t/2) = \mathbb{P}\big(\exp\big(\lambda(-S)\big) > \exp(\lambda t/2)\big)\\ &\le \frac{\mathbb{E}\exp(\lambda(-S))}{e^{\lambda t/2}} \le \exp\big(-\lambda t/2 + M\lambda^2\big) \le q_{\mathrm{offd}}.\end{aligned}$$

The lemma is thus proved using the union bound. □

The theorem is thus proved. □

The plan is to first bound the moment generating function for the S_0 in the diagonal sum in Section 4.1. We then bound the moment generating function for the off-diagonal sum as stated in Lemma 4.4 in Section 4.2.

4.1. Proof of Lemma 4.3

Recall $A_\xi = D_0 D_\xi D_0 = (\widetilde{a}_{ij}) = (d_i^T D_\xi d_j)$. Then for $\widetilde{a}_{kk} = d_k^T D_\xi d_k = \sum_{i=1}^m d_{ki}^2 \xi_i$

$$S_0 := \sum_{k=1}^m \big(X_k^2 - \mathbb{E}X_k^2\big) A_{\xi,kk} = \sum_{k=1}^m \big(X_k^2 - \mathbb{E}X_k^2\big)\widetilde{a}_{kk}.$$

To estimate the moment generating function of S_0, we first consider ξ as being fixed and thus treat $\widetilde{a}_{ij}$ as fixed coefficients. The bound on the moment generating function of S_0 as in (16) will involve the following symmetric matrices A_1 and A_2 which we now define:

$$\begin{aligned}A_1 &:= D_0 \circ D_0 = [d_1 \circ d_1, \dots, d_m \circ d_m],\\ A_2 &= \big(a''_{ij}\big) = A_1^2 = (d_1 \circ d_1, \dots, d_m \circ d_m)(d_1 \circ d_1, \dots, d_m \circ d_m)^T \\ &= \sum_{k=1}^m \big(d_k d_k^T\big) \circ \big(d_k d_k^T\big) = \sum_{k=1}^m (d_k \circ d_k)(d_k \circ d_k)^T \succeq 0.\end{aligned} \tag{20}$$

Thus, we have both A_0, A_2 being positive semidefinite, while in general A_1 is not positive semidefinite unless $D_0 \succeq 0$ by the Schur Product theorem. See Theorem 5.2.1 [10].

Lemma 4.7. *Suppose all conditions in Lemma 4.3 hold. Let* $C_0 = 38.94$. *Then for* $|\lambda| \le \frac{1}{64\|A_0\|_2} \le \frac{1}{64 a_\infty}$,

$$\mathbb{E}\exp(\lambda S_0) \le \mathbb{E}\exp\big(C_0\lambda^2 \xi^T A_2 \xi\big) \le \mathbb{E}\exp\big(C_0\lambda^2 \big\|\operatorname{diag}(A_\xi)\big\|_F^2\big). \tag{21}$$

Proof. We first compute the moment generating function for S_0 when ξ is fixed. Conditioned on ξ, $\widetilde{a}_{kk}, \forall k$ are considered as fixed coefficients. Indeed, for $|\lambda| \le \frac{1}{64 a_\infty}$, by independence of X_i

$$\begin{aligned}\mathbb{E}\big(\exp(\lambda S_0)|\xi\big) &= \mathbb{E}_X \exp\left(\lambda \sum_{k=1}^m \widetilde{a}_{kk}\big(X_k^2 - \mathbb{E}X_k^2\big)\right) = \prod_{k=1}^m \mathbb{E}_X \exp\big(\lambda \widetilde{a}_{kk}\big(X_k^2 - \mathbb{E}X_k^2\big)\big) \\ &\le \prod_{k=1}^m \exp\big(38.94\lambda^2 \widetilde{a}_{kk}^2\big) = \exp\left(C_0\lambda^2 \sum_{k=1}^m \widetilde{a}_{kk}^2\right),\end{aligned}$$

where the inequality follows from Lemma 2.12 with $\tau := \lambda \widetilde{a}_{kk}$ in view of (22):

$$\forall k, \forall \xi, \qquad |\lambda \widetilde{a}_{kk}| \le \frac{1}{64} \le \frac{1}{23.5e} \qquad \text{where } |\widetilde{a}_{kk}| \le \langle d_k, d_k\rangle = a_{kk} \le a_\infty. \tag{22}$$

Now

$$\begin{aligned}\sum_{k=1}^m \widetilde{a}_{kk}^2 &= \sum_{k=1}^m \big(d_k^T D_\xi d_k\big)^2 = \sum_{k=1}^m \operatorname{tr}\big(d_k^T D_\xi d_k d_k^T D_\xi d_k\big) \\ &= \sum_{k=1}^m \operatorname{tr}\big(D_\xi d_k d_k^T D_\xi d_k d_k^T\big) = \sum_{k=1}^m \xi^T\big(\big(d_k d_k^T\big) \circ d_k d_k^T\big)\xi =: \xi^T A_2 \xi,\end{aligned}$$

where $A_2 = (a''_{ij}) = (D_0 \circ D_0)^2$ is as defined in (20). Thus

$$\mathbb{E}_X \exp(\lambda S_0) \le \exp\big(C_0\lambda^2 \xi^T A_2 \xi\big) = \exp\big(C_0\lambda^2 \big\|\operatorname{diag}(A_\xi)\big\|_F^2\big) \tag{23}$$

and (21) is thus proved by taking expectation on both sides of (23) with respect to random variables in vector ξ. □

To prove (19) in the lemma statement, notice that for all $\xi \in \{0,1\}^m$,

$$\sum_{k=1}^m \widetilde{a}_{kk}^2 = \big\|\operatorname{diag}(A_\xi)\big\|_F^2 \le \|A_\xi\|_F^2.$$

Thus, we have for $|\lambda| \le \frac{1}{64\|A_0\|_2}$,

$$\mathbb{E}\exp(\lambda S_0) \le \mathbb{E}\exp\big(C_0\lambda^2 \big\|\operatorname{diag}(A_\xi)\big\|_F^2\big) \le \mathbb{E}\exp\big(C_0\lambda^2 \|A_\xi\|_F^2\big) = \mathbb{E}\exp\big(C_0\lambda^2 \xi^T (A_0 \circ A_0)\xi\big),$$

where $A_0 = (a_{ij})$. Finally, we invoke Corollary 4.8 to finish the proof of Lemma 4.3.

Corollary 4.8. *Let A_0, ξ be as defined in Theorem 1.2. Then for $|\lambda| \le \frac{1}{64\|A_0\|_2}$ and $C_0 \le 38.94$*

$$\mathbb{E}\exp\left(C_0\lambda^2 \sum_{i,j} a_{ij}^2 \xi_i \xi_j\right) \le \exp\big(\lambda^2 N\big),$$

where $N = (40\sum_{j=1}^m p_j a_{jj}^2 + 54\sum_{i \ne j} p_i p_j a_{ij}^2)$.

The proof of Corollary 4.8 follows exactly that of Corollary 4.11 in view of Theorem 2.10 and is thus omitted. The lemma is thus proved. □

Remark 4.9. An alternative bound can be stated as follows: for $\lambda \le \frac{1}{64 a_\infty}$,

$$\mathbb{E}\exp(\lambda S_0) \le \exp\left(41\lambda^2 \sum_{j=1}^m \sigma_j^2 a_{jj}^2 + 52\lambda^2 \|A_1 \mathbf{p}\|_2^2\right),$$

where $\mathbf{p} = [p_1, \dots, p_m]$ and $\sigma_j^2 = p_j(1 - p_j)$. The proof follows from a direct analysis based on the quadratic form $\xi^T A_2 \xi$ on the RHS of (21), which is omitted from the present paper. This bound may lead to a slight improvement upon the final bound in (19). We do not pursue this improvement here because the bound in (19) is sufficient for us to obtain the final large deviation bound as stated in Theorem 1.2.

4.2. Proof of Lemma 4.4

Let $\mathbb{E}_X$ and $\mathbb{E}_\xi$ denote the expectation with respect to random variables in vectors X and ξ, respectively. Recall

$$S_{\text{offd}} = \sum_{i \ne j} X_i X_j (A_{\xi,ij}) =: \sum_{i \ne j} \widetilde{a}_{ij} X_i X_j \qquad \text{where } \widetilde{a}_{ij} = d_i^T D_\xi d_j = \sum_{k=1}^m d_{ik} \xi_k d_{jk}.$$

To estimate the moment generating function of S_{offd}, we first consider ξ as being fixed and thus treat $\widetilde{a}_{ij}$ as fixed coefficients. Lemma 4.10 reduces the original problem of estimating the moment generating function of S_{offd} to the new problem of estimating the moment generating function of $S := \xi^T (A_0 \circ A_0)\xi$, which involves a new quadratic form with independent non-centered random variables $\xi_1, \dots, \xi_m \in \{0, 1\}$ and the symmetric matrix $(A_0 \circ A_0)$ as shown in (24). Lemma 4.10 follows from the proof of Theorem 1 [14] directly. We omit the proof in this paper.

Lemma 4.10. *Consider $\xi \in \{0, 1\}^m$ as being fixed and denote by $A_\xi = D_0 D_\xi D_0$ and $A_0 = D_0^2 = (a_{ij})$. Then, for some constant C and $|\lambda| \le \frac{1}{12C\|A_0\|_2}$ and $C_2 = 32C^2$,*

$$\mathbb{E}_X \exp(\lambda S_{\text{offd}}) \le \exp\big(C_2\lambda^2 \|A_\xi\|_F^2\big) = \exp\big(C_2\lambda^2 \xi^T (A_0 \circ A_0)\xi\big). \tag{24}$$

Note that $\|D_\xi D_0\|_2 \le \|D_0\|_2$ and hence by symmetry

$$\|A_\xi\|_2 = \|D_0 D_\xi D_\xi D_0\|_2 = \|D_\xi D_0\|_2^2 \le \|D_\xi\|_2^2 \|D_0\|_2^2 = \|A_0\|_2,$$
$$\|A_\xi\|_F^2 = \mathrm{tr}(A_0 D_\xi A_0 D_\xi) = \xi^T (A_0 \circ A_0)\xi.$$

In order to estimate the moment generating function S_{offd}, we now take expectation with respect to ξ on both sides of (24). Thus, we have for $|\lambda| \le \frac{1}{12C\|A_0\|_2}$

$$\mathbb{E}_\xi \mathbb{E}_X \exp(\lambda S_{\mathrm{offd}}) \le \mathbb{E}\exp\bigl(C_2\lambda^2 \|A_\xi\|_F^2\bigr) = \mathbb{E}\exp\bigl(C_2\lambda^2 \xi^T (A_0 \circ A_0)\xi\bigr). \tag{25}$$

Corollary 4.11. *Then for $|\lambda| \le \frac{1}{58C\|A_0\|_2}$ and $t := C_2\lambda^2$, where $C_2 = 32C^2$ and C is a large enough absolute constant,*

$$\mathbb{E}\exp\left(t\sum_{i,j} a_{ij}^2 \xi_i \xi_j\right) \le \exp\bigl(\lambda^2 M\bigr),$$

where $M := C^2(33\sum_{i=1}^m a_{ii}^2 p_i + 44\sum_{i\ne j} a_{ij}^2 p_i p_j)$.

Combining Lemma 4.10, (25) and Corollary 4.11, we have for $|\lambda| \le \frac{1}{58C\|A_0\|_2}$

$$\mathbb{E}(\lambda S_{\mathrm{offd}}) \le \mathbb{E}\exp\bigl(C_2\lambda^2 \|A_\xi\|_F^2\bigr) = \mathbb{E}\exp\left(t\sum_{i,j} a_{ij}^2 \xi_i \xi_j\right) \le \exp\bigl(\lambda^2 M\bigr).$$

Lemma 4.4 thus holds. □

Corollary 4.11 follows from Theorem 2.10 immediately, which is derived in the current paper for estimating the moment generating function of $S' := \xi^T A\xi$ where A is an arbitrary matrix and ξ is a Bernoulli random vector with independent elements as defined in Theorem 1.2.

Proof of Corollary 4.11. Clearly for the choices of t and λ,

$$t = C_2\lambda^2 \le \frac{32C^2}{58^2 C^2 \|A_0\|_2^2} \le \frac{1}{104\|A_0\|_2^2} \le \frac{1}{104\|A_0 \circ A_0\|_1 \vee \|A_0 \circ A_0\|_\infty},$$

where we use the fact that for symmetric A_0,

$$\begin{aligned}\|A_0 \circ A_0\|_1 &= \|A_0 \circ A_0\|_\infty \\ &= \max_{1\le i\le m} \sum_{j=1}^m a_{ij}^2 = \max_{1\le i\le m} \|A_0 e_i\|_2^2 \le \|A_0\|_2^2.\end{aligned}$$

Thus we can apply Theorem 2.10 with $B := (A_0 \circ A_0)$ to obtain for $0 < t \le \frac{1}{104(\|A\|_1 \vee \|A\|_\infty)}$,

$$\mathbb{E}\exp\Big(t\sum_{i,j} a_{ij}^2 \xi_i \xi_j\Big) \le \exp\Big(1.02t\sum_{j=1}^m p_j a_{jj}^2\Big)\exp\Big(1.373t\sum_{i\neq j} a_{ij}^2 p_i p_j\Big)$$
$$\le \exp\Big(C^2\lambda^2\Big(33\sum_{j=1}^m p_j a_{jj}^2 + 44\sum_{i\neq j} p_i p_j a_{ij}^2\Big)\Big). \qquad \square$$

5. Application to covariance estimation in a matrix variate model

In the current paper, we focus on presenting the concentration of measure bounds for entries in the gram matrices for (4) and (5) rather than estimators for $A_0 \succ 0 \in \mathbf{R}^{m\times m}$ and $B_0 \succ 0 \in \mathbf{R}^{n\times n}$. In particular, the large deviation bounds in Theorems 5.1 and 5.3 can be used to design a set of entrywise unbiased estimators for A_0 and B_0, up to a scaling factor, as well as penalized estimators which achieve convergence in the operator and the Frobenius norm in the spirit of [19]. In this section, we narrowly focus on the baseline concentration of measure bounds on gram matrices $\mathcal{X}\mathcal{X}^T$ and $\mathcal{X}^T\mathcal{X}$ evolving around the relationship (29) and (30). Without loss of generality, we assume that $n \le m$ and $n/m \to r$ for some $r \in (0, 1]$.

Recall that we observe the matrix variate data under a mask:

$$\mathcal{X} = \mathbb{U} \circ \mathbb{X} \qquad \text{where } \mathbb{X} = B_0^{1/2}\mathbb{Z}A_0^{1/2} \text{ is as defined in (4)},$$

and $\mathbb{U}$ is a mask with entries being either zero or 1. We denote $\mathbb{U} \in \{0, 1\}^{n\times m}$ by

$$\mathbb{U} = \begin{bmatrix} u^1 & u^2 & \dots & u^m \end{bmatrix} = \begin{bmatrix} v^1 & v^2 & \dots & v^n \end{bmatrix}^T \qquad \text{where } \forall i,\ v^1, \dots, v^n \sim \mathbf{v} \in \{0, 1\}^m$$

are independent random vectors such that $\mathbf{v}$ is composed of independent Bernoulli random variables and

$$\mathbb{E}\mathbf{v} =: \zeta = (\zeta_1, \dots, \zeta_m), \text{ the vector of sampling probabilities.} \tag{26}$$

Theorem 5.1 justifies the consideration of (29) as an entrywise unbiased estimator of A_0 and $\rho(A_0)$; for the sake of proper normalization, we present our bounds using entries of $\rho_{ij}(A_0)$.

Theorem 5.1. *Consider the data generating model as in* (4) *and* (5)*. Then for $t > 0$, for each i,*

$$\mathbb{P}\Big(\frac{1}{a_{ii}}\big|\langle u^i \circ x^i, u^i \circ x^i\rangle - \zeta_i \operatorname{tr}(B_0)\big| > \tau\Big)$$
$$\le 2\exp\Big(-c_2 \min\Big(\frac{\tau^2}{4K^4(\zeta_i\|\operatorname{diag}(B_0)\|_F^2 + \zeta_i^2\|\operatorname{offd}(B_0)\|_F^2)}, \frac{\tau}{2K^2\|B_0\|_2}\Big)\Big),$$

and for $i \neq j$,

$$\mathbb{P}\left(\left|\frac{\langle u^i \circ x^i, u^j \circ x^j\rangle}{\sqrt{a_{ii}a_{jj}}} - \rho_{ij}(A_0)\operatorname{tr}(B_0)\zeta_i\zeta_j\right| > \tau\right) \leq 6\exp\left(-c_2 \min\left(\frac{\tau^2}{4K^4(\zeta_i\zeta_j\|\operatorname{diag}(B_0)\|_F^2 + \zeta_i^2\zeta_j^2\|\operatorname{offd}(B_0)\|_F^2)}, \frac{\tau}{2K^2\|B_0\|_2}\right)\right).$$

Theorem 5.2 follows from Theorem 5.1 and the analysis of Corollary 2.5. The proof is thus omitted.

Theorem 5.2. *Consider the data generating model as in* (4) *and* (5). *Let* $\mathcal{N}_{jj} = \operatorname{tr}(B_0)\zeta_j, \forall j$ *and* $\mathcal{N}_{ij} := \zeta_i\zeta_j \operatorname{tr}(B_0)$ *for all* $i \neq j$. *Then, with probability at least* $1 - \frac{2}{m^2}$, *we have*

$$\forall j, \qquad \left|\frac{\|u^j \circ x^j\|_2^2}{\mathcal{N}_{jj}} - a_{jj}\right| \leq C_1 K^2 a_{jj} \log m \frac{\|B_0\|_2}{\operatorname{tr}(B_0)\zeta_j} + C_3 \log^{1/2} m a_{jj}\left(\frac{\|\operatorname{diag}(B_0)\|_F}{\operatorname{tr}(B_0)\zeta_j^{1/2}} + \frac{\|\operatorname{offd}(B_0)\|_F}{\operatorname{tr}(B_0)}\right), \tag{27}$$

and for all $i \neq j$,

$$\left|\frac{1}{\sqrt{a_{ii}a_{jj}}\mathcal{N}_{ij}}\langle u^i \circ x^i, u^j \circ x^j\rangle - \rho_{ij}(A_0)\right| \leq C_2 K^2 \frac{\log m \|B_0\|_2}{\zeta_i\zeta_j \operatorname{tr}(B_0)} + C_4 K^2 \log^{1/2} m \left(\frac{\|\operatorname{diag}(B_0)\|_F}{\sqrt{\zeta_i\zeta_j}\operatorname{tr}(B_0)} + \frac{\|\operatorname{offd}(B_0)\|_F}{\operatorname{tr}(B_0)}\right), \tag{28}$$

where C_1, C_2, C_3 *and* C_4 *are some absolute constants chosen so that the probability holds.*

Some consequences on correlation estimation. For $\rho(B_0)$, we have a rather nice matrix entry-wise max norm bound as we will show in Theorem 5.4; For $\rho(A_0)$, this bound very much depends on the sampling probabilities in $\zeta = (\zeta_1, \dots, \zeta_m)$ as shown in Theorem 5.2. In particular, in order for both terms on the RHS (27) to be of $o(a_{jj})$, we require that for all j,

$$\zeta_j = \Omega\left(\frac{\log m \|B_0\|_2}{\operatorname{tr}(B_0)}\right), \quad \text{and similarly} \quad \forall i \neq j \qquad \zeta_i\zeta_j = \Omega\left(\frac{\log m \|B_0\|_2}{\operatorname{tr}(B_0)}\right)$$

is needed so that both terms on the RHS of (28) will be of $o(1)$. In the context of estimating $\rho(B_0)$, we will discuss what happens when the sampling rate is below a certain threshold.

First, we assume that we know the parameters $\operatorname{tr}(B_0)$ and ζ as defined in (26), Theorems 5.1 and 5.2 show that in order to estimate A_0, we may consider the following oracle estimators for entries of A_0 and $\rho(A_0)$ with the gram matrix $\mathcal{X}^T\mathcal{X}$:

$$\widetilde{A}_0 = \mathcal{X}^T\mathcal{X} \oslash \mathcal{N} \qquad \text{where } \mathcal{N} := \operatorname{tr}(B_0)\mathbb{E}v^i \otimes v^i \tag{29}$$

and

$$\mathcal{N}_{ij} = \operatorname{tr}(B_0) \begin{cases} \zeta_i & \text{if } i = j, \\ \zeta_i \zeta_j & \text{if } i \neq j, \end{cases}$$

where $\oslash$ denotes entrywise division. Clearly, one can take advantage of the bounds as derived in (27) and (28) and consider $\widetilde{A}_{0,ij}/(\widetilde{A}_{0,ii}\widetilde{A}_{0,jj})^{1/2}$ in order to estimate $\rho_{ij}(A_0)$ for each $i \neq j$. We leave the presentation of such estimators and their statistical properties for future work [21], where we will discuss the estimation of elements in $\mathcal{M}$, ζ and their concentration of measure properties. In order to estimate B_0, we first exploit the following relationship

$$\widetilde{B}_0 = \mathcal{X}\mathcal{X}^T \oslash \mathcal{M} \qquad \text{where } \mathcal{M}_{ij} = \begin{cases} \sum_{k=1}^m a_{kk}\zeta_k & \text{if } i = j, \\ \sum_{k=1}^m a_{kk}\zeta_k^2 & \text{if } i \neq j. \end{cases} \tag{30}$$

Theorem 5.3. *Consider the data generating random matrices as in* (4) *and* (5). *Then for* $t > 0$, *for each* i,

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{b_{ii}}\left|\langle v^i \circ y^i, v^i \circ y^i\rangle - \sum_{k=1} \zeta_k a_{kk}\right| > t\right) \\ &\quad \le 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\sum_{k=1}^m \zeta_k a_{kk}^2 + \sum_{k\neq \ell} a_{k\ell}^2 \zeta_k\zeta_\ell)}, \frac{t}{2K^2\|A_0\|_2}\right)\right), \end{aligned} \tag{31}$$

and

$$\begin{aligned} \forall i \neq j, \qquad &\mathbb{P}\left(\left|\frac{\langle v^i \circ y^i, v^j \circ y^j\rangle}{\sqrt{b_{ii}b_{jj}}} - \rho_{ij}(B_0)\sum_{k=1}^m a_{kk}\zeta_k^2\right| > t\right) \\ &\quad \le 6\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\sum_{k=1}^m a_{kk}^2\zeta_k^2 + \sum_{k\neq \ell} a_{k\ell}^2 \zeta_k^2\zeta_\ell^2)}, \frac{t}{2K^2\|A_0\|_2}\right)\right). \end{aligned} \tag{32}$$

Theorem 5.4 follows from Theorem 5.3 and the analysis of Corollary 2.8. The proof is thus omitted.

Theorem 5.4. *Consider the data generating random matrices as in* (4) *and* (5). *Let* $\mathcal{M}_{ii} = \sum_{k=1}^m a_{kk}\zeta_k$ *for all* i. *We have with probability at least* $1 - \frac{1}{m^2}$, *for all* i,

$$\left|\frac{\langle v^i \circ y^i, v^i \circ y^i\rangle}{\mathcal{M}_{ii}} - b_{ii}\right| \le \frac{CK^2 b_{ii}}{\mathcal{M}_{ii}}\left(\log m\|A_0\|_2 + \log^{1/2} m\sqrt{a_\infty\|A_0\|_2}\sqrt{\sum_{k=1}^m \zeta_k}\right), \tag{33}$$

and for all $i \neq j$ *and* $\mathcal{M}_{ij} = \sum_{k=1}^{m} a_{kk}\zeta_k^2$,

$$\left|\frac{\langle v^i \circ y^i, v^j \circ y^j\rangle}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B_0)\right| \leq \frac{C'K^2}{\mathcal{M}_{ij}}\left(\log m\|A_0\|_2 + \log^{1/2} m\sqrt{a_\infty\|A_0\|_2}\sqrt{\sum_{i=1}^{m}\zeta_k^2}\right), \quad (34)$$

where C, C' *are chosen so that the probability holds.*

Remarks. Let us now elaborate on the choices of $\zeta = [\zeta_1, \ldots, \zeta_m]$, which are the sampling probabilities as defined in (26) to make sense of the relative errors in estimating entries of the covariance matrix B_0. Denote by $\zeta_{\min} = \min_k \zeta_k$ and $\zeta_{\max} = \max_k \zeta_k$. Let $A_0 = (a_{ij})$ and $a_\infty = \max_{i=1}^n a_{ii}$ and $a_{\min} = \min_{i=1}^n a_{ii}$. To ease the discussion, we assume that $K = 1$ w.l.o.g. First, we focus on the diagonal entries. Recall that for all i, $\mathbb{E}\langle v^i \circ y^i, v^i \circ y^i\rangle = b_{ii}\langle D_\zeta, \operatorname{diag}(A_0)\rangle$, where $D_\zeta = \operatorname{diag}(\zeta_1, \ldots, \zeta_m)$.

Case 1: Suppose that $\sum_{k=1}^m \zeta_k = O(\log m)$. In this regime, the linear in t term in (31) would be smaller than the quadratic one for all non-trivial values of t: given that

$$M = \sum_{i=1}^{m}\zeta_i a_{ii}^2 + \sum_{i\neq j} a_{ij}^2\zeta_i\zeta_j \leq 2a_\infty\|A_0\|_2\sum_{k=1}^{m}\zeta_k$$

following the analysis in Corollary 2.8; and hence for $t \geq 4a_\infty\sum_{k=1}^m \zeta_k$,

$$\min\left(\frac{t^2}{4(\sum_{i=1}^m \zeta_i a_{ii}^2 + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k\zeta_\ell)}, \frac{t}{2\|A_0\|_2}\right)$$
$$\geq \min\left(\frac{t^2}{8a_\infty\|A_0\|_2\sum_{k=1}^m \zeta_k}, \frac{t}{2\|A_0\|_2}\right) = \frac{t}{2\|A_0\|_2}.$$

Thus we would not see the two-phase behavior of Hanson–Wright inequality when we set $t \geq 4\log m\|A_0\|_2$, which is necessary for us to obtain probability error bound in the order of $\frac{1}{m^d}$ for some $d \geq 2$. Moreover, the large deviation bound we obtain through (31) is not tight enough for our purpose, in the sense that the RHS of (33) is $\Omega(b_{ii})$ for all i, when we set $t \asymp \log m\|A_0\|_2$.

Case 2: Suppose that all sampling rates are at the same order and

$$\zeta_{\max} \asymp \zeta_{\min} \asymp p = \Omega\left(\frac{\log m\|A_0\|_2}{\operatorname{tr}(A_0)}\right).$$

Following the analysis of Corollary 2.5, we have with probability at least $1 - \frac{1}{m^2}$,

$$\left|\frac{\langle v^i \circ y^i, v^i \circ y^i\rangle}{\sum_{i=1}^m a_{ii}\zeta_i} - b_{ii}\right| \leq \frac{b_{ii}\log^{1/2} m}{\zeta_{\min}\operatorname{tr}(A_0)}\left(\log^{1/2} m\|A_0\|_2 + \zeta_{\max}^{1/2}\|A_0\|_F\right) = o(1).$$

Case 3: There is no reason to assume a limit on $\zeta_{\max}$. Suppose instead, we assume that

$$\sum_{k=1}^{m} \zeta_k = \Omega\left(\frac{a_\infty \|A_0\|_2 \log m}{a_{\min}^2}\right) \tag{35}$$

for $a_\infty := \max_i a_{ii}$ and $a_{\min} = (\min_i a_{ii})$, which would imply that $\zeta_{\min} = \Omega(\frac{a_\infty \|A_0\|_2 \log m}{m a_{\min}^2})$. This in turn is slightly stronger than the lower bound on $\zeta_{\min} = \Omega(\frac{\log m \|A_0\|_2}{\operatorname{tr}(A_0)})$ in Case 2, given that $\frac{a_\infty}{a_{\min}^2} \geq \frac{m}{\operatorname{tr}(A_0)}$ since $a_\infty \operatorname{tr}(A_0) \geq m a_{\min}^2$. However, when we assume that $a_{ii} \asymp 1$, for example, when we deal with a correlation matrix, then (35) is an overall weaker condition than that in Case 2. In general, condition (35) is needed for the following upper bound to go through. With probability at least $1 - \frac{1}{m^3}$, for all i, by (33),

$$\left|\frac{\langle v^i \circ y^i, v^i \circ y^i \rangle}{\sum_{i=1}^m a_{ii}\zeta_i} - b_{ii}\right| \leq O(b_{ii})\left(\frac{\log m \|A_0\|_2}{\zeta_{\min} \operatorname{tr}(A_0)} + \frac{\log^{1/2} m \sqrt{a_\infty \|A_0\|_2}}{a_{\min}\sqrt{\sum_{i=1}^m \zeta_k}}\right) = o(b_{ii}),$$

where the last step holds in view of (35).

Now we exam the rate of convergence for the off-diagonal entries.

Case 1: For $i \neq j$, assume that $\sum_{k=1}^m \zeta_k^2 = O(\log m)$; In this regime, the linear in t term in (32) would be smaller than the quadratic one for all non-trivial values of t, following the same reasoning for the diagonal case, except that the effective sampling rate becomes $\sum_{k=1}^m \zeta_k^2$. Hence, we would not see the two-phase behavior of Hanson–Wright when we set $t \geq 4 \log m \|A_0\|_2$. Moreover, the large deviation bound we obtain through the expression on the RHS of (34) is not tight enough for our purpose, as

$$\frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + \frac{\log^{1/2} m \sqrt{a_\infty \|A_0\|_2}\sqrt{\sum_{i=1}^m \zeta_k^2}}{\sum_{k=1}^m a_{kk}\zeta_k^2} = \Omega(1).$$

In order to obtain convergence in estimating $\rho_{ij}(B_0)$, we need to impose the following conditions.

Case 2: Suppose all sampling rate are at the same order:

$$\zeta_{\max}^2 \asymp \zeta_{\min}^2 \asymp p = \Omega\left(\frac{\log m \|A_0\|_2}{\operatorname{tr}(A_0)}\right). \tag{36}$$

Following the analysis of Corollary 2.5, we have with probability at least $1 - \frac{1}{m^2}$,

$$\begin{aligned}\left|\frac{\langle v^i \circ y^i, v^j \circ y^j \rangle}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B_0)\right| &\leq \frac{\log^{1/2} m}{\sum_{k=1}^m \zeta_k^2 a_{kk}} O\left(\log^{1/2} m \|A_0\|_2 + \zeta_{\max}\|A_0\|_F\right) \\ &= o(1).\end{aligned}$$

Case 3: In general, there is no reason to impose any condition on $\zeta_{\max}$ except that by definition, it is larger than $\zeta_{\min}$. Hence in general, we assume that

$$\sum_{k=1}^{m} \zeta_k^2 = \Omega\left(\frac{a_\infty \|A_0\|_2 \log m}{a_{\min}^2}\right) \tag{37}$$

which implies that $\zeta_{\min}^2 = \Omega(\frac{a_\infty \|A_0\|_2 \log m}{m a_{\min}^2})$ which is slightly stronger than the lower bound on $\zeta_{\min}^2$ as stated in (36). Finally, we have with probability at least $1 - \frac{1}{m^2}$, RHS of (34) $= o(1)$ for all $i \neq j$.

The proof of Theorem 5.1 is similar to the proof of Theorem 5.3 and thus is omitted. We now sketch the proof of Theorem 5.3.

Proof of Theorem 5.3. Recall that we observe for each row vector y^i of data matrix $\mathbb{X}$:

$$v^i \circ y^i, \qquad \text{where } v_k^i \sim \text{Bernoulli}(\zeta_k), \forall k = 1, \dots, m, \forall i = 1, \dots, n. \tag{38}$$

Let $\xi = (\xi_1, \dots, \xi_m)$, where $\xi_k := v_k^i v_k^j$ are independent Bernoulli random variables such that $\mathbb{E}\xi_k = \zeta_k$ if $i = j$ and else $\mathbb{E}\xi_k = \zeta_k^2$. First, observe that when $i = j$, the random vector $(y_k^j)_{k=1}^m$ involved in the sum is of size m, with covariance being $b_{jj} A_0$. Without loss of generality, we write $(y_k^j)_{k=1}^m = (b_{jj} A_0)^{1/2} (g_1, \dots, g_m)^T$, where $g_1, \dots, g_m$ i.i.d. $\sim Y$ where

$$\mathbb{E}Y = 0, \qquad \|Y\|_{\psi_2} \leq K \quad \text{and} \quad \mathbb{E}Y^2 = 1 \tag{39}$$

and replace the inner product with a random quadratic form

$$\mathbf{g}^T A_0^{1/2} D_\xi A_0^{1/2} \mathbf{g} - \mathbb{E}\mathbf{g}^T A_0^{1/2} D_\xi A_0^{1/2} \mathbf{g},$$

where D_ξ follows the same distribution as $\text{diag}(v^j)$ for $v^j \sim \mathbf{v}$ as defined in (26), with $\mathbb{E}v_k^j = \zeta_k$ for $k = 1, \dots, m$. The first inequality (31) in the theorem thus follows immediately from Theorem 1.2. For $i \neq j$, we exploit the decorrelation idea in Theorem 13.1 [19,20], and Theorems 1.2 and 1.3 in the present work, for which we now have $D_\xi = \text{diag}(v^i \otimes v^j)$ in the random quadratic forms, which explains the difference in the quadratic form in the second inequality (32) versus the first. Due to its significant length, we omit it from the current work and leave it in [21]. $\square$

In summary, we have shown that the entries of estimators $\widetilde{A}_0$ and $\widetilde{B}_0$ presented in this section are tightly concentrated around their mean while the diagonal entries have a tighter concentration than that for the off-diagonal entries of A_0 and B_0; the proof exploits the sparse Hanson–Wright type of inequalities, namely, Theorem 1.2 and its corollaries, as well as the decorrelation ideas in [19,20]. In an ongoing work [21], we consider fully automated estimators for A_0 and B_0 and their statistical convergence properties, where no population parameters are assumed to be known; indeed, the factor such as $\text{tr}(B_0)$ appearing in (29) should not matter as we only aim to estimate A_0 and B_0 up to a certain factor.

6. Proof of Theorem 2.10

We first state the following Theorem 6.1 from a note by Vershynin [16]; we state its consequence as follows.

Theorem 6.1. *Let A be an $m \times m$ matrix. Let $X = (X_1, \dots, X_m)$ be a random vector with independent mean zero coefficients. Then, for every convex function F,*

$$\mathbb{E}F\left(\sum_{i \neq j} a_{ij} X_i X_j\right) \leq \mathbb{E}F\left(4 \sum_{i \neq j} a_{ij} X_i X_j'\right), \tag{40}$$

where X' is an independent copy of X.

Let $Z_i := \xi_i - p_i$. Denote by $\sigma_i^2 = p_i(1 - p_i)$. For all Z_i, we have $|Z_i| \leq 1$, $\mathbb{E}Z_i = 0$ and

$$\mathbb{E}Z_i^2 = (1 - p_i)^2 p_i + p_i^2(1 - p_i) = p_i(1 - p_i) = \sigma_i^2, \tag{41}$$

$$\mathbb{E}|Z_i| = (1 - p_i)p_i + p_i(1 - p_i) = 2p_i(1 - p_i) = 2\sigma_i^2. \tag{42}$$

Proof of Theorem 2.10. Let $Z_i = \xi_i - p_i$. Denote by $\breve{a}_i := \sum_{j \neq i}(a_{ij} + a_{ji})p_j + a_{ii}$. We express the quadratic form as follows:

$$\sum_{i=1}^m a_{ii}(\xi_i - p_i) + \sum_{i \neq j} a_{ij}(\xi_i \xi_j - p_i p_j) = \sum_{i \neq j} a_{ij} Z_i Z_j + \sum_{j=1}^m Z_j \breve{a}_i =: S_1 + S_2.$$

We first state the following bounds on the moment generating functions of S_1 and S_2 in (45) and (46). The estimate on the moment generating function for $\sum_{i,j} a_{ij}\xi_i\xi_j$ then follows immediately from the Cauchy–Schwarz inequality in view of (45) and (46).

Bounding the moment generating function for S_1. In order to bound the moment generating function for S_1, we start by a decoupling step following Theorem 6.1. Let Z' be an independent copy of Z.

Decoupling. Now consider random variable $S_1 := \sum_{i \neq j} a_{ij}(\xi_i - p_i)(\xi_j - p_j) = \sum_{i \neq j} a_{ij} Z_i Z_j$ and

$$S_1' := \sum_{i \neq j} a_{ij} Z_i Z_j', \qquad \text{we have } \mathbb{E}\exp(2\lambda S_1) \leq \mathbb{E}\exp\bigl(8\lambda S_1'\bigr) =: f$$

by (40). Thus we have by independence of Z_i,

$$f := \mathbb{E}_{Z'}\mathbb{E}_Z \exp\left(8\lambda \sum_{i=1}^m Z_i \sum_{j \neq i} a_{ij} Z_j'\right) = \mathbb{E}_{Z'} \prod_{i=1}^m \mathbb{E}\bigl(\exp(8\lambda Z_i \widetilde{a}_i)\bigr). \tag{43}$$

First consider Z' being fixed. Let us define

$$t_i := 8\lambda \widetilde{a}_i \qquad \text{where } \widetilde{a}_i := \sum_{j \neq i} a_{ij} Z_j'.$$

Hence for all $0 \le \lambda \le \frac{1}{104\|A\|_\infty}$ and $C_4 := \frac{4}{13}e^{1/13}$, and any given fixed Z' by (9)

$$\begin{aligned}\mathbb{E}\exp(8\lambda\widetilde{a}_i Z_i) := \mathbb{E}\exp(t_i Z_i) \le 1 + \frac{1}{2}t_i^2\mathbb{E}Z_i^2 e^{|t_i|} \le \exp\left(\frac{1}{2}t_i^2\mathbb{E}Z_i^2 e^{|t_i|}\right) \\ \le \exp\left(\frac{4}{13}e^{1/13}\lambda|\widetilde{a}_i|\sigma_i^2\right) =: \exp\big(C_4\lambda|\widetilde{a}_i|\sigma_i^2\big),\end{aligned} \tag{44}$$

where $Z_i, \forall i$ satisfies: $|Z_i| \le 1$, $\mathbb{E}Z_i = 0$ and $\mathbb{E}Z_i^2 = \sigma_i^2$,

$$|t_i| = |8\lambda\widetilde{a}_i| \le 8\lambda\sum_{j\ne i}|a_{ij}|\big|Z'_j\big| \le 8\lambda\|A\|_\infty \le \frac{1}{13} \quad \text{and} \quad \frac{1}{2}t_i^2 \le \frac{4}{13}\lambda|\widetilde{a}_i|.$$

Denote by $|\bar{a}_j| := \sum_{i\ne j}|a_{ij}|\sigma_i^2$. Thus by (43) and (44)

$$\begin{aligned}f &\le \mathbb{E}_{Z'}\prod_{i=1}^m \exp\big(C_4\lambda|\widetilde{a}_i|\sigma_i^2\big) \le \mathbb{E}_{Z'}\exp\left(C_4\lambda\sum_{i=1}^m\sigma_i^2\sum_{j\ne i}|a_{ij}|\big|Z'_j\big|\right) \\ &= \prod_{j=1}^m \mathbb{E}\exp\left(C_4\lambda\big|Z'_j\big|\sum_{i\ne j}|a_{ij}|\sigma_i^2\right) =: \prod_{j=1}^m \mathbb{E}\exp\big(C_4\lambda|\bar{a}_j|\big|Z'_j\big|\big),\end{aligned}$$

where we have by the elementary approximation (9) and $\breve{t}_j := C_4\lambda|\bar{a}_j|$

$$\begin{aligned}\mathbb{E}\exp\big(C_4\lambda|\bar{a}_j|\big|Z'_j\big|\big) &=: \mathbb{E}\exp\big(\breve{t}_j\big|Z'_j\big|\big) \\ &\le 1 + \mathbb{E}\big(\breve{t}_j\big|Z'_j\big|\big) + \frac{1}{2}(\breve{t}_j)^2\mathbb{E}\big(Z'_j\big)^2 e^{|\breve{t}_j|} \\ &\le \exp\left(2\breve{t}_j\sigma_j^2 + \frac{1}{2}(\breve{t}_j)^2\sigma_j^2 e^{0.0008}\right) \le \exp\big(2.0005\breve{t}_j\sigma_j^2\big) \\ &\le \exp\big(2.0005C_4\lambda|\bar{a}_j|\sigma_j^2\big) \le \exp\left(\frac{2}{3}\lambda\sigma_j^2\sum_{i\ne j}|a_{ij}|\sigma_i^2\right),\end{aligned}$$

where $\mathbb{E}(Z'_i)^2 = \sigma_i^2$ and $\mathbb{E}|Z'_i| = 2\sigma_i^2$ following (41) and (42), and for $0 < \lambda \le \frac{1}{104\max(\|A\|_1, \|A\|_\infty)}$,

$$\begin{aligned}\breve{t}_j := C_4\lambda|\bar{a}_j| &= \frac{4}{13}e^{1/13}\lambda\sum_{i\ne j}|a_{ij}|\sigma_i^2 \\ &\le \frac{4}{13}e^{1/13}\frac{1}{4}\frac{\sum_i|a_{ij}|}{104\|A\|_1} \le \frac{1}{13}e^{1/13}\frac{1}{104} < 0.0008.\end{aligned}$$

Thus for every $0 \le \lambda \le \frac{1}{104 \max(\|A\|_1, \|A\|_\infty)}$,

$$\mathbb{E}\exp(\lambda 2 S_1) \le \exp\left(\frac{2}{3}\lambda \sum_{i \neq j} |a_{ij}| \sigma_i^2 \sigma_j^2\right). \tag{45}$$

Bounding the moment generating function for S_2. Recall

$$S_2 := \sum_{i=1}^m Z_i \left(\sum_{j \neq i} (a_{ij} + a_{ji}) p_j + a_{ii}\right) =: \sum_{i=1}^m Z_i \breve{a}_i.$$

Let $a_\infty := \max_i |\breve{a}_i| \le \|A\|_\infty + \|A\|_1$. Thus we have by Lemma 2.11

$$\begin{aligned} g &:= \mathbb{E}\exp(2\lambda S_2) = \mathbb{E}\exp\left(2\lambda \sum_{i=1}^m Z_i \breve{a}_i\right) \\ &\le \exp\left(2\lambda^2 e^{2|\lambda| a_\infty} \sum_{i=1}^m \breve{a}_i^2 \sigma_i^2\right) \le \exp\left(C_5 |\lambda| \sum_{i=1}^m |\breve{a}_i| p_i\right), \end{aligned}$$

where $e^{2\lambda a_\infty} 2\lambda |\breve{a}_i| \le \frac{1}{26} e^{1/26} =: C_5 \le 0.04$ given that for all $|\lambda| \le \frac{1}{52(\|A\|_\infty + \|A\|_1)}$

$$2\lambda |\breve{a}_i| \le \frac{2(\|A\|_\infty + \|A\|_1)}{52(\|A\|_\infty + \|A\|_1)} \le \frac{1}{26} \qquad \text{for all } i.$$

Thus we have for $0 < \lambda \le \frac{1}{52(\|A\|_\infty + \|A\|_1)}$,

$$\mathbb{E}\exp(\lambda 2 S_2) \le \exp\left(0.02 * 2\lambda \sum_{i=1}^m p_i |\breve{a}_i|\right). \tag{46}$$

Hence by the Cauchy–Schwarz inequality, for all $0 < \lambda \le \frac{1}{104(\|A\|_\infty \vee \|A\|_1)}$,

$$\begin{aligned} &\mathbb{E}\exp\left(\lambda\left(\sum_{i=1}^m a_{ii}(\xi_i - p_i) + \sum_{i \neq j} a_{ij}(\xi_i \xi_j - p_i p_j)\right)\right) \\ &\quad = \mathbb{E}\exp\big(\lambda(S_1 + S_2)\big) \le \mathbb{E}^{1/2}\exp(2\lambda S_1)\mathbb{E}^{1/2}\exp(2\lambda S_2). \end{aligned}$$

The theorem is thus proved by multiplying $\exp(\lambda(\sum_{i=1}^m a_{ii} p_i + \sum_{i \neq j} a_{ij} p_i p_j))$ on both sides of the above inequality. $\square$

Appendix A: Proof of Lemma 2.12

Let $Z := X^2 - \mathbb{E}X^2$ and $Y := Z/\|Z\|_{\psi_1}$. Then Y and Z are both centered sub-exponential random variables with $\|Y\|_{\psi_1} = 1$ and

$$\|Z\|_{\psi_1} = \left\|X^2 - \mathbb{E}X^2\right\|_{\psi_1} \leq 2\left\|X^2\right\|_{\psi_1} \leq 4\|X\|_{\psi_2}^2 \leq 4K^2$$

which follows from the triangle inequality and Lemma 5.14 of [17].

Now set $t := \tau\|X^2 - \mathbb{E}X^2\|_{\psi_1}$, where for $|\tau| \leq \frac{1}{23.5eK^2}$,

$$e|t| = e|\tau|\left\|X^2 - \mathbb{E}X^2\right\|_{\psi_1} \leq \frac{4K^2}{23.5K^2} < \frac{8}{47}$$

and $2(e|t|)^3 \leq (e|t|)^2$. By Lemma 5.15 of [17], we have for all k,

$$\begin{aligned}\mathbb{E}\exp(tY) &= 1 + t\mathbb{E}Y + \sum_{p=2}^{\infty}\frac{t^p\mathbb{E}Y^p}{p!} \leq 1 + \sum_{p=2}^{\infty}\frac{|t|^p\mathbb{E}|Y|^p}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty}\frac{|t|^p p^p}{p!} \leq 1 + \sum_{p=2}^{\infty}\frac{|t|^p e^p}{\sqrt{2\pi p}}.\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}\exp(tY) &\leq 1 + \frac{(te)^2}{2\sqrt{\pi}} + \frac{1}{\sqrt{6\pi}}\sum_{p=3}^{\infty}\big(e|t|\big)^p \\ &\leq 1 + \frac{|t|^2e^2}{2\sqrt{\pi}} + \frac{1}{\sqrt{6\pi}}\frac{(e|t|)^3}{1-e|t|} \leq 1 + \frac{|t|^2e^2}{2\sqrt{\pi}} + \frac{8(e|t|)^2}{39\sqrt{6\pi}} \\ &< 1 + e^2\tau^2\left\|X^2 - \mathbb{E}X^2\right\|_{\psi_1}^2\left(\frac{1}{2\sqrt{\pi}} + \frac{8}{39\sqrt{6\pi}}\right) \leq 1 + 38.94|\tau|^2K^4,\end{aligned}$$

where we used the following form of Stirling's approximation for all $p \geq 2$,

$$\frac{1}{p!} \leq \frac{e^p}{p^p}\frac{1}{\sqrt{2\pi p}} \leq \frac{e^p}{p^p}\frac{1}{2\sqrt{\pi}}.$$

The lemma is thus proved given that

$$\mathbb{E}\exp\tau\big(X^2 - \mathbb{E}X^2\big) = \mathbb{E}\exp\big(\tau\left\|X^2 - \mathbb{E}X^2\right\|_{\psi_1}Y\big) = \mathbb{E}\exp(tY). \qquad \square$$

Appendix B: Proof of Lemma 3.2

Note that the following holds by Lemma 5.14 [17],

$$\|X_i\|_{\psi_2}^2 \leq \left\|X_i^2\right\|_{\psi_1} \leq 2\|X_i\|_{\psi_2}^2 = 2K^2.$$

For all k, let $Y_k := X_k^2/\|X_k^2\|_{\psi_1}$. By definition, Y_k is a sub-exponential random variable with $\|Y_k\|_{\psi_1} = 1$. We now set $t_k := \lambda a_{kk}\|X_k^2\|_{\psi_1}$. Following the proof of Lemma 2.12, we first use the Taylor expansions to obtain for all k,

$$\begin{aligned}
\mathbb{E}\exp(t_k Y_k) := \mathbb{E}\exp\bigl(\lambda a_{kk} X_k^2\bigr) &= 1 + t_k \mathbb{E} Y_k + \sum_{p=2}^{\infty} \frac{t_k^p \mathbb{E} Y_k^p}{p!} \\
&\le 1 + t_k \mathbb{E} Y_k + \sum_{p=2}^{\infty} \frac{|t_k|^p \mathbb{E}|Y_k|^p}{p!} \\
&\le 1 + \lambda a_{kk}\mathbb{E}X_k^2 + \frac{|t_k|^2 e^2}{2\sqrt{\pi}} + \frac{1}{\sqrt{6\pi}} \sum_{p=3}^{\infty} \bigl(e|t_k|\bigr)^p \\
&\le 1 + \lambda a_{kk}\mathbb{E}X_k^2 + \frac{|t_k|^2 e^2}{2\sqrt{\pi}} + \frac{1}{\sqrt{6\pi}} 2\bigl(e|t_k|\bigr)^3 \\
&< 1 + \lambda a_{kk}\mathbb{E}X_k^2 + e^2\bigl(\lambda a_{kk}\|X_k^2\|_{\psi_1}\bigr)^2 \left(\frac{1}{2\sqrt{\pi}} + \frac{1}{\sqrt{6\pi}}\right) \\
&\le 1 + \lambda a_{kk}\mathbb{E}X_k^2 + 16|\lambda a_{kk}|^2 K^4,
\end{aligned}$$

where

$$e|t_k| \le \frac{|a_{kk}|\|X_k^2\|_{\psi_1}}{4K^2 \max_k |a_{kk}|} \le \frac{1}{2}$$

and $2(e|t_k|)^3 \le (e|t_k|)^2$. The lemma is thus proved. □

Appendix C: Proof of (13)

The proof structure follows from the proof of Theorem 2.1 [14]. Recall $S := \sum_{i\neq j}^m a_{ij} X_i X_j \xi_i \xi_j$. We start with a decoupling step.

Step 1. Decoupling. Let $\delta = (\delta_1, \dots, \delta_m) \in \{0, 1\}^m$ be a random vector with independent Bernoulli random variables with $\mathbb{E}\delta_i = 1/2$, which is independent of X and ξ. Let X_{Λ_δ} denote $(X_i)_{i\in\Lambda_\delta}$ for a set $\Lambda_\delta := \{i \in [m] : \delta_i = 1\}$. Let $\mathbb{E}_X$, $\mathbb{E}_\xi$ and $\mathbb{E}_\delta$ denote the expectation with respect to random variables in X, ξ and δ respectively. Now consider random variable

$$S_\delta := \sum_{i,j} \delta_i(1-\delta_j) a_{ij} X_i X_j \xi_i \xi_j \quad \text{and hence} \quad S = 4\mathbb{E}_\delta S_\delta.$$

By Jensen's inequality, for all $\lambda \in \mathbf{R}$,

$$\mathbb{E}\exp(\lambda S) = \mathbb{E}_\xi \mathbb{E}_X \exp(\mathbb{E}_\delta 4\lambda S_\delta) \le \mathbb{E}_\xi \mathbb{E}_X \mathbb{E}_\delta \exp(4\lambda S_\delta), \tag{47}$$

where the last step holds because e^{ax} is convex on $\mathbf{R}$ for any $a \in \mathbf{R}$.

Consider $\Lambda_\delta := \{i \in [m] : \delta_i = 1\}$. Denote by $f(\xi, \delta, X_{\Lambda_\delta})$ the conditional moment generating function of random variable $4S_\delta$:

$$f(\xi, \delta, X_{\Lambda_\delta}) := \mathbb{E}\big(\exp(4\lambda S_\delta) | \xi, \delta, X_{\Lambda_\delta}\big).$$

Conditioned upon X_{Λ_δ} for a fixed realization of ξ and δ, we rewrite S_δ

$$S_\delta := \sum_{i \in \Lambda_\delta, j \in \Lambda_\delta^c} a_{ij} X_i X_j \xi_i \xi_j = \sum_{j \in \Lambda_\delta^c} X_j \left(\xi_j \sum_{i \in \Lambda_\delta} a_{ij} X_i \xi_i\right)$$

as a linear combination of mean-zero subgaussian random variables X_j, $j \in \Lambda_\delta^c$, with fixed coefficients. Thus the conditional distribution of S_δ is subgaussian with ψ_2 norm being upper bounded by the ℓ_2 norm of the coefficient vector $(\xi_j \sum_{i \in \Lambda_\delta} a_{ij} X_i \xi_i)_{j \in \Lambda_\delta^c}$ [17] (cf. Lemma 5.9).

Thus, conditioned upon ξ, δ and X_{Λ_δ},

$$\|S_\delta\|_{\psi_2} \le C_0 \sigma_{\delta,\xi} \qquad \text{where } \sigma_{\delta,\xi}^2 = \sum_{j \in \Lambda_\delta^c} \xi_j \left(\sum_{i \in \Lambda_\delta} a_{ij} X_i \xi_i\right)^2. \tag{48}$$

Thus, we have for some large absolute $C > 0$

$$f(\xi, \delta, X_{\Lambda_\delta}) = \mathbb{E}\big(\exp(4\lambda S_\delta) | \xi, \delta, X_{\Lambda_\delta}\big) \le \exp\big(C\lambda^2 \|S_\delta\|_{\psi_2}^2\big) \le \exp\big(C'\lambda^2 \sigma_{\delta,\xi}^2\big). \tag{49}$$

Taking the expectations of both sides with respect to X_{Λ_δ} and ξ, we obtain

$$\begin{aligned} \mathbb{E}_\xi \mathbb{E}_{X_{\Lambda_\delta}} f(\xi, \delta, X_{\Lambda_\delta}) &= \mathbb{E}_\xi \mathbb{E}_{X_{\Lambda_\delta}} \mathbb{E}\big(\exp(4\lambda S_\delta) | \xi, \delta, X_{\Lambda_\delta}\big) \\ &\le \mathbb{E}_\xi \mathbb{E}_{X_{\Lambda_\delta}} \exp\big(C'\lambda^2 \sigma_{\delta,\xi}^2\big) =: \widetilde{f}_\delta. \end{aligned} \tag{50}$$

Step 2. Reduction to normal random variables. Let δ, ξ and X_{Λ_δ} be a fixed realization of the random vectors defined as above. Let $g = (g_1, \ldots, g_n)$, where g_i i.i.d. $\sim N(0, 1)$. Let $\mathbb{E}_g$ denote the expectation with respect to random variables in g. Consider random variable

$$Z := \sum_{j \in \Lambda_\delta^c} g_j \left(\xi_j \sum_{i \in \Lambda_\delta} a_{ij} X_i \xi_i\right).$$

By the rotation invariance of normal distribution, for a fixed realization of random vectors ξ, δ, X, the conditional distribution of Z follows $N(0, \sigma_{\delta,\xi}^2)$ for $\sigma_{\delta,\xi}^2$ as defined in (48). Thus we obtain the conditional moment generating function for Z denoted by

$$\mathbb{E}_g\big(\exp(tZ)\big) := \mathbb{E}\big(\exp(tZ) | \xi, \delta, X_{\Lambda_\delta}\big) = \exp\big(t^2 \sigma_{\delta,\xi}^2 / 2\big).$$

Choose $t = C_1 \lambda$ where $C_1 = \sqrt{2C'}$, we have

$$\mathbb{E}_g\big(\exp(C_1 \lambda Z)\big) = \exp\big(C'\lambda^2 \sigma_{\delta,\xi}^2\big) \qquad \text{which matches the RHS of (49).}$$

Hence for a fixed realization of δ, we can calculate $\widetilde{f_\delta}$ using Z as follows:

$$\widetilde{f_\delta} := \mathbb{E}_\xi \mathbb{E}_X \exp\big(C'\lambda^2 \sigma^2_{\delta,\xi}\big) = \mathbb{E}_\xi \mathbb{E}_X \mathbb{E}_g \big(\exp(C_1 \lambda Z)\big) = \mathbb{E}\big(\exp(C_1\lambda Z)|\delta\big). \tag{51}$$

Conditioned on δ, ξ and g, we can re-express Z:

$$Z = \sum_{i\in\Lambda_\delta} X_i \Bigg(\xi_i \sum_{j\in\Lambda_\delta^c} a_{ij} g_j \xi_j\Bigg)$$

as a linear combination of subgaussian random variables X_i, $i \in \Lambda_\delta$ with fixed coefficients, which immediately imply that

$$\mathbb{E}\big(\exp(C_1\lambda Z)|\delta, \xi, g\big) \le \exp\Bigg(C_3\lambda^2 \sum_{i\in\Lambda_\delta} \xi_i \Bigg(\sum_{j\in\Lambda_\delta^c} a_{ij} g_j \xi_j\Bigg)^2\Bigg).$$

Let P_δ denote the coordinate projection of $\mathbf{R}^m$ onto $\mathbf{R}^{\Lambda_\delta}$. Then conditioned on δ, we have by definition of $\widetilde{f_\delta}$ as in (51) and the bounds on the conditional moment generating function of Z immediately above,

$$\begin{aligned}
\widetilde{f_\delta} &= \mathbb{E}\big(\exp(C_1\lambda Z)|\delta\big) = \mathbb{E}_{\xi,g}\mathbb{E}\big(\exp(C_1\lambda Z)|\delta, \xi, g\big) \\
&\le \mathbb{E}\Bigg[\exp\Bigg(C_3\lambda^2 \sum_{i\in\Lambda_\delta} \xi_i \Bigg(\sum_{j\in\Lambda_\delta^c} a_{ij} g_j \xi_j\Bigg)^2\Bigg)|\delta\Bigg] \\
&= \mathbb{E}\big[\exp\big(C_3\lambda^2 \big\|D_\xi P_\delta A(I-P_\delta) D_\xi g\big\|_2^2\big)|\delta\big] \\
&= \mathbb{E}\big[\exp\big(C_3\lambda^2 \|A_{\delta,\xi} g\|_2^2\big)|\delta\big],
\end{aligned} \tag{52}$$

where we denote by $A_{\delta,\xi} := D_\xi P_\delta A(I-P_\delta)D_\xi$. We will integrate g out followed by ξ in the next two steps.

Step 3. Integrating out the normal random variables. Conditioned upon δ and ξ and by the rotation invariance of g, the random variables $\|A_{\delta,\xi} g\|_2^2$ follows the same distribution as $\sum_i s_i^2 g_i^2$ where s_i denote the singular values of $A_{\delta,\xi}$, with

$$\begin{aligned}
\max_i s_i &= \sqrt{\lambda_{\max}\big(A_{\delta,\xi}^T A_{\delta,\xi}\big)} =: \|A_{\delta,\xi}\|_2 \le \|A\|_2, \quad \text{and} \\
\sum_i s_i^2 &= \|A_{\delta,\xi}\|_F^2 = \operatorname{tr}\big(A_{\delta,\xi} A_{\delta,\xi}^T\big) \\
&= \operatorname{tr}\big(D_\xi P_\delta A(I-P_\delta) D_\xi A^T P_\delta D_\xi\big) = \operatorname{tr}\big(D_\xi P_\delta A(I-P_\delta)D_\xi A^T\big) \\
&= \sum_{i\in\Lambda_\delta} \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j.
\end{aligned} \tag{53}$$

First we note that $g_i^2, \forall i$ follow the χ^2 distribution with one degree of freedom, and $\mathbb{E}\exp(tg^2) = \frac{1}{\sqrt{1-2t}} \leq e^{2t}$ for $t < 1/4$. Thus, we have for a fixed realization of δ, ξ, and for all $|\lambda| \leq \frac{1}{2\sqrt{C_3}\|A\|_2}$,

$$\mathbb{E}\big[\exp\big(C_3\lambda^2 s_i^2 g_i^2\big)|\delta, \xi\big] = \frac{1}{\sqrt{1 - 2C_3\lambda^2 s_i^2}} \leq \exp\big(2C_3\lambda^2 s_i^2\big).$$

Hence for any fixed δ and ξ, and for $C_4 = 2C_3$ and $|\lambda| \leq \frac{1}{2\sqrt{C_3}\|A\|_2}$, we have by independence of $g_1, g_2, \ldots,$

$$\begin{aligned}
\mathbb{E}\big[\exp\big(C_3\lambda^2 \|A_{\delta,\xi} g\|_2^2\big)|\delta, \xi\big] &= \mathbb{E}\left[\exp\left(C_3\lambda^2 \sum_i s_i^2 g_i^2\right)\Big|\delta, \xi\right] \\
&= \prod_i \mathbb{E}\big[\exp\big(C_3\lambda^2 s_i^2 g_i^2\big)|\delta, \xi\big] \\
&\leq \prod_i \exp\big(2C_3\lambda^2 s_i^2\big).
\end{aligned} \tag{54}$$

Thus we have by (52), (53) and (54)

$$\begin{aligned}
\widetilde{f_\delta} &\leq \mathbb{E}_\xi \mathbb{E}\big[\exp\big(C_3\lambda^2 \|A_{\delta,\xi} g\|_2^2\big)|\delta, \xi\big] \\
&\leq \mathbb{E}\left[\exp\left(2C_3\lambda^2 \sum_i s_i^2\right)\Big|\delta\right] \\
&= \mathbb{E}\left[\exp\left(C_4\lambda^2 \sum_{i\in\Lambda_\delta} \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\right)\Big|\delta\right].
\end{aligned} \tag{55}$$

The key observation here is we are dealing with a quadratic form on the RHS of (55) which is already decoupled thanks to the decoupling Step 1.

Step 4. Integrating out the Bernoulli random variables. For any given realization of δ, we now need to bound the moment generating function for the decoupled quadratic form on the RHS of (55), which is the content of Lemma C.1 where we take $t = C_4\lambda^2$ and conclude that for all δ and for all $|\lambda| \leq \frac{1}{2\sqrt{C_4}\|A\|_2}$,

$$\widetilde{f_\delta} \leq \exp\left(1.44 C_4\lambda^2 \sum_{i\neq j} a_{ij}^2 p_i p_j\right).$$

Lemma C.1. *Let $0 < \tau \leq \frac{1}{4\|A\|_2^2}$. For any fixed realization of δ, we have*

$$\mathbb{E}\left[\exp\left(\tau \sum_{i\in\Lambda_\delta} \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\right)\Big|\delta\right] \leq \exp\left(1.44\tau \sum_{i\neq j} a_{ij}^2 p_i p_j\right).$$

Proof. As mentioned, we are dealing with a quadratic form which is already decoupled. Thus, we integrate out ξ_i for all $i \in \Lambda_\delta$ followed by those in Λ_δ^c. Recall for any realization of δ and $0 < \tau \le \frac{1}{4\|A\|_2^2}$ we have by independence of $\xi_1, \xi_2, \dots,$

$$\begin{aligned} f_\delta &:= \mathbb{E}\biggl[\exp\biggl(\tau \sum_{i\in\Lambda_\delta} \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr)\Bigm|\delta\biggr] \\ &= \mathbb{E}_{\xi_{\Lambda_\delta^c}} \mathbb{E}\biggl[\exp\biggl(\tau \sum_{i\in\Lambda_\delta} \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr)\Bigm|\xi_{\Lambda_\delta^c}, \delta\biggr] \\ &= \mathbb{E}_{\xi_{\Lambda_\delta^c}} \prod_{i\in\Lambda_\delta} \mathbb{E}\biggl[\exp\biggl(\tau \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr)\Bigm|\xi_{\Lambda_\delta^c}, \delta\biggr]. \end{aligned} \tag{56}$$

We will use the following approximation twice in our proof:

$$e^x - 1 \le 1.2x \qquad \text{which holds for } 0 \le x \le 0.35. \tag{57}$$

First notice that for all realizations of δ and ξ, we have for $0 < \tau \le \frac{1}{4\|A\|_2^2}$

$$0 \le \tau \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j \le \tau \sum_j a_{ij}^2 \le \tau \|A\|_2^2 \le 1/4$$

given that the maximum row ℓ_2 norm of A is bounded by the operator norm of matrix A^T: $\|A^T\|_2 = \|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$. Hence, we have for $|\lambda| \le \frac{1}{2\sqrt{C_4}\|A\|_2}$, (57) and the fact that $1 + x \le e^x$,

$$\begin{aligned} \mathbb{E}\biggl[\exp\biggl(\tau \xi_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr)\Bigm|\xi_{\Lambda_\delta^c}, \delta\biggr] &= p_i \exp\biggl(\tau \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr) + (1 - p_i) \\ &\le p_i\biggl(1.2\tau \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr) + 1 \le \exp\biggl(1.2\tau p_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr). \end{aligned} \tag{58}$$

Thus we have by independence of $\xi_1, \xi_2, \dots,$ (56) and (58)

$$\begin{aligned} f_\delta &\le \mathbb{E}_{\xi_{\Lambda_\delta^c}} \prod_{i\in\Lambda_\delta} \exp\biggl(1.2\tau p_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr) = \mathbb{E}_{\xi_{\Lambda_\delta^c}} \exp\biggl(\sum_{i\in\Lambda_\delta} 1.2\tau p_i \sum_{j\in\Lambda_\delta^c} a_{ij}^2 \xi_j\biggr) \\ &= \mathbb{E}_{\xi_{\Lambda_\delta^c}} \exp\biggl(1.2\tau \sum_{j\in\Lambda_\delta^c} \xi_j \sum_{i\in\Lambda_\delta} a_{ij}^2 p_i\biggr) = \prod_{j\in\Lambda_\delta^c} \mathbb{E}_{\xi_j} \exp\biggl(1.2\tau \xi_j \sum_{i\in\Lambda_\delta} a_{ij}^2 p_i\biggr) \\ &= \prod_{j\in\Lambda_\delta^c} p_j \exp\biggl(1.2\tau \sum_{i\in\Lambda_\delta} a_{ij}^2 p_i\biggr) + (1 - p_j), \end{aligned} \tag{59}$$

where for all δ and $0 < \tau \le \frac{1}{4\|A\|_2^2}$, we have by the approximation in (57)

$$\exp\left(1.2\tau \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i\right) - 1 \le 1.44\tau \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i \tag{60}$$

given that the column ℓ_2 norm of A is bounded by the operator norm of A, and thus

$$1.2\tau \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i \le 1.2\tau \sum_{i=1}^{m} a_{ij}^2 p_i \le 1.2 \sum_{i=1}^{m} a_{ij}^2 / \left(4\|A\|_2^2\right) \le 0.3.$$

Now by (59), (60) and the fact that $x + 1 \le e^x$

$$\begin{aligned} f_\delta &\le \prod_{j \in \Lambda_\delta^c} p_j \left[\exp\left(1.2\tau \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i\right) - 1\right] + 1 \\ &\le \prod_{j \in \Lambda_\delta^c} p_j \left(1.44\tau \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i\right) + 1 \le \prod_{j \in \Lambda_\delta^c} \exp\left(1.44\tau p_j \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i\right) \\ &= \exp\left(\sum_{j \in \Lambda_\delta^c} 1.44\tau p_j \sum_{i \in \Lambda_\delta} a_{ij}^2 p_i\right) \le \exp\left(1.44\tau \sum_{i \ne j} a_{ij}^2 p_i p_j\right). \end{aligned}$$

The lemma thus holds. □

Step 5. Putting things together. By Jensen's inequality (47), definition of $f(\xi, \delta, X_{\Lambda_\delta})$ in (49) and (50), we have for all $|\lambda| \le \frac{1}{2\sqrt{C_4}\|A\|_2}$

$$\begin{aligned} \mathbb{E}\exp(\lambda S) &\le \mathbb{E}_\delta \mathbb{E}_\xi \mathbb{E}_X \exp(4\lambda S_\delta) \\ &= \mathbb{E}_\delta \mathbb{E}_\xi \mathbb{E}_{X_{\Lambda_\delta}} \mathbb{E}\left(\exp(4\lambda S_\delta) | \xi, \delta, X_{\Lambda_\delta}\right) \\ &= \mathbb{E}_\delta \mathbb{E}_\xi \mathbb{E}_{X_{\Lambda_\delta}} f(\xi, \delta, X_{\Lambda_\delta}) \\ &\le \mathbb{E}_\delta \widetilde{f_\delta} \le \exp\left(1.44 C_4 \lambda^2 \sum_{i \ne j} a_{ij}^2 p_i p_j\right). \end{aligned}$$

Thus, (13) holds. □

Acknowledgements

Shuheng Zhou is supported in part by NSF Grant DMS 1316731 and Elizabeth Caroline Crosby Research Award from the Advance Program at the University of Michigan.

Mark Rudelson encouraged me to apply the method from [14] to prove the first result in the current paper. The author is also grateful for discussions with Tailen Hsing, which helped improving the presentation of this paper tremendously. The proof presented here was filed in part as Technical Report, 539, October, 2015, Department of Statistics, University of Michigan.

References

[1] Adamczak, R. and Wolff, P. (2015). Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probab. Theory Related Fields* **162** 531–586. MR3383337

[2] Barthe, F. and Milman, E. (2012). Transference principles for log-Sobolev and spectral-gap with applications to conservative spin systems.

[3] Dawid, A.P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

[4] de la Peña, V.H. and Giné, E. (1999). *Decoupling*: *From Dependence to Independence*: *Randomly Stopped Processes*. *U-Statistics and Processes*. *Martingales and Beyond*. *Probability and Its Applications* (*New York*). New York: Springer. MR1666908

[5] de la Peña, V.H. and Montgomery-Smith, S.J. (1995). Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Ann*. *Probab*. **23** 806–816. MR1334173

[6] Diakonikolas, I., Kane, D.M. and Nelson, J. (2010). Bounded independence fools degree-2 threshold functions. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 11–20. Los Alamitos, CA: IEEE Computer Soc. MR3024771

[7] Foucart, S. and Rauhut, H. (2013). *A Mathematical Introduction to Compressive Sensing*. *Applied and Numerical Harmonic Analysis*. New York: Birkhäuser/Springer. MR3100033

[8] Gupta, A.K. and Varga, T. (1992). Characterization of matrix variate normal distributions. *J*. *Multivariate Anal*. **41** 80–88. MR1156682

[9] Hanson, D.L. and Wright, F.T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann*. *Math*. *Stat*. **42** 1079–1083. MR0279864

[10] Horn, R.A. and Johnson, C.R. (1991). *Topics in Matrix Analysis*. Cambridge: Cambridge Univ. Press. MR1091716

[11] Hsu, D., Kakade, S.M. and Zhang, T. (2012). Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron*. *Commun*. *Probab*. **17** Art. ID 14. MR2900355

[12] Latała, R. (2006). Estimates of moments and tails of Gaussian chaoses. *Ann*. *Probab*. **34** 2315–2331. MR2294983

[13] Rudelson, M. (2016). On the complexity of the set of unconditional convex bodies. *Discrete Comput*. *Geom*. **55** 185–202. MR3439264

[14] Rudelson, M. and Vershynin, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron*. *Commun*. *Probab*. **18** Art. ID 82. MR3125258

[15] Talagrand, M. (1995). Sections of smooth convex bodies via majorizing measures. *Acta Math*. **175** 273–300. MR1368249

[16] Vershynin, R. (2011). A simple decoupling inequality in probability theory. Available at http://www-personal.umich.edu/~romanv/papers/papers.html.

[17] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170

[18] Wright, F.T. (1973). A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *Ann*. *Probab*. **1** 1068–1070. MR0353419

[19] Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *Ann*. *Statist*. **42** 532–562.

[20] Zhou, S. (2014). Supplement to "Gemini: Graph estimation with matrix variate normal instances". *Ann*. *Statist*. DOI:10.1214/13-AOS1187SUPP.

[21] Zhou, S. (2019). The concentration of measure phenomenon on sparse matrix variate random matrices. Working paper.

Received February 2017 and revised July 2017