SPECTRAL NORM OF PRODUCTS OF RANDOM AND DETERMINISTIC MATRICES

ROMAN VERSHYNIN

Abstract. We study the spectral norm of matrices M that can be factored as $M = BA$, where A is a random matrix with independent mean zero entries and B is a fixed matrix. Under the $(4+\varepsilon)$-th moment assumption on the entries of A, we show that the spectral norm of such an $m \times n$ matrix M is bounded by $\sqrt{m} + \sqrt{n}$, which is sharp. In other words, in regard to the spectral norm, products of random and deterministic matrices behave similarly to random matrices with independent entries. This result along with the previous work of M. Rudelson and the author implies that the smallest singular value of a random $m \times n$ matrix with i.i.d. mean zero entries and bounded $(4+\varepsilon)$-th moment is bounded below by $\sqrt{m} - \sqrt{n-1}$ with high probability.

1. Introduction

This paper grew out of an attempt to understand the class of random matrices with non-independent entries, but which can be *factorized* through random matrices with independent entries. Equivalently, we are interested in sample covariance matrices of a wide class of random vectors – the linear transformations of vectors with independent entries.

Here we study the spectral norm of such matrices. Recall that the spectral norm $\|M\|$ is defined as the largest singular value of a matrix M, which equals the largest eigenvalue of $\sqrt{M^*M}$. Equivalently, the spectral norm can be defined as the $\ell_2 \to \ell_2$ operator norm: $\|M\| = \sup_x \|Mx\|_2/\|x\|_2$ where $\|\cdot\|_2$ denotes the Euclidean norm. The spectral norm of random matrices plays a notable role in particular in geometric functional analysis, computer science, statistical physics, and signal processing.

1.1. Matrices with independent entries.

For random matrices with independent and identically distributed entries, the spectral norm is well studied. Let M be an $m \times n$ matrix whose entries are real independent and identically

Date: December 8, 2008.
Partially supported by NSF grant FRG DMS 0652617 and Alfred P. Sloan Research Fellowship.

distributed random variables with mean zero, variance 1 and finite fourth moment. Estimates of the type

$$\|M\| \sim \sqrt{n} + \sqrt{m} \tag{1.1}$$

are known to hold (and are sharp) in both asymptotic and non-asymptotic sense. The meaning of (1.1) in the asymptotic sense is that, for a family of matrices as above, whose dimensions m and n increase to infinity and whose aspect ratio m/n converges to a constant, the ratio $\|M\|/(\sqrt{n}+\sqrt{m})$ converges to 1 almost surely [18].

In the non-asymptotic regime, when the dimensions n and m are fixed, variants of (1.1) were proved by Y. Seginer [12] and R. Latala [9]. If M is an $m \times n$ matrix whose entries are real independent mean zero random variables, then denoting the rows of M by X_i and the columns by Y_j, the result of Y. Seginer [12] states that

$$\mathbb{E}\|M\| \le C\big(\mathbb{E} \max_i \|X_i\|_2 + \mathbb{E} \max_j \|Y_j\|_2\big)$$

where C is an absolute constant. This estimate is sharp because $\|M\|$ is obviously bounded below by the Euclidean norm of any row and any column of M. Furthermore, if the entries m_{ij} of the matrix M are not necessarily identically distributed, then R. Latala's result [9] states that

$$\mathbb{E}\|M\| \le C\Big(\max_i \mathbb{E}\|X_i\|_2 + \max_j \mathbb{E}\|Y_j\|_2 + \big(\sum_{i,j} \mathbb{E} m_{ij}^4\big)^{1/4}\Big).$$

In particular, if M is an $m \times n$ matrix whose entries are independent random variables with mean zero and fourth moments bounded by 1, then one can deduce from either Y. Seginer's or R. Latala's result that

$$\mathbb{E}\|M\| \le C(\sqrt{n} + \sqrt{m}). \tag{1.2}$$

This is a variant of (1.1) in the non-asymptotic sense.

The fourth moment hypothesis is known to be necessary. Consider again a family of matrices whose dimensions m and n increase to infinity, and whose aspect ratio m/n converges to a constant. If the entries are independent and identically distributed random variables with mean zero and infinite fourth moment, then the upper limit of the ratio $\|M\|/(\sqrt{n} + \sqrt{m})$ is infinite almost surely [18].

1.2. **The main result.** The main result of this paper is an extension of the optimal non-asymptotic bound (1.2) to the class of random matrices with non-independent entries, but which can be factored through a matrix with independent entries.

Theorem 1.1. *Consider a random $m \times n$ matrix $M = BA$, where A is an $N \times n$ random matrix whose entries are independent random variables with mean zero and $(4+\varepsilon)$-th moment bounded by 1, and B is an $m \times N$ non-random matrix such that $\|B\| \le 1$. Then*

$$\mathbb{E}\|M\| \le C(\varepsilon)(\sqrt{n} + \sqrt{m}) \tag{1.3}$$

where $C(\varepsilon)$ is a function that depends only on $\varepsilon > 0$.

Remarks. **1.** An important feature of this result is that its conclusion is independent of the dimension N.

2. As the proof of Theorem 1.1 shows, one can replace the $m \times N$ matrix B by an $M \times N$ matrix B with arbitrary M, and with controlled spectral and Hilbert-Schmidt norms:

$$\|B\| \le 1, \qquad \|B\|_{\mathrm{HS}} \le \sqrt{m}. \tag{1.4}$$

The bound (1.3) will still hold. This extension of Theorem 1.1 is thus independent of both dimensions N and M. Therefore the conclusion of Theorem 1.1 holds for an arbitrary linear operator B acting from the N-dimensional Euclidean space ℓ_2^N to an arbitrary Hilbert space, and which satisfies (1.4).

3. Theorem 1.1 can be interpreted in terms of *sample covariance matrices* of random vectors in $\mathbb{R}^m$ of the form BX, where X is a random vector in $\mathbb{R}^N$ with independent entries. Indeed, let A be the random matrix whose columns are n independent samples of the vector X. Then $M = BA$ is the matrix whose columns are n independent samples of the random vector BX. The sample covariance matrix of the random vector BX is defined as $\Sigma = \frac{1}{n}MM^*$. Theorem 1.1 states that the largest eigenvalue of Σ is bounded by $C_1(\varepsilon)(1+m/n)$, which is further bounded by $C_2(\varepsilon)$ for the number of samples $n \gtrsim m$ (and independently of the dimension N). This problem was previously studied in [2], [3] in the asymptotic regime for $m = N$, where the result must of course depend on N.

4. Under the stronger subgaussian moment assumption (1.6) on the entries, Theorem 1.1 is easy to prove using standard concentration and an ε-net argument. In contrast, if only some finite moment is assumed, we do not know any simple argument.

1.3. **The smallest singular value.** Our main motivation for Theorem 1.1 was to complete the analysis of the *smallest singular value* of random rectangular matrices carried out by M. Rudelson and the author in [17]. The smallest singular value $s_{\min}(M)$ of a matrix M can be equivalently described as $s_{\min}(M) = \inf_x \|Mx\|_2/\|x\|_2$.

Analyzing the smallest singular value is generally harder than analyzing the largest one (the spectral norm). The analogue of (1.1) for the smallest singular value of random $m \times n$ matrices M (for $m > n$) is

$$s_{\min}(M) \sim \sqrt{m} - \sqrt{n}. \tag{1.5}$$

The asymptotic version of this result proved in [5] holds under exactly the same hypotheses as (1.1) – for i.i.d. entries with mean zero, variance 1 and finite fourth moment.

An optimal non-asymptotic version of (1.5) was proved in [17] under somewhat stronger moment assumptions. Namely, we assumed that the entries m_{ij} of the matrix M are *subgaussian* random variables. This means that all their moments are bounded by the corresponding moments of the standard normal random variable, i.e.

$$(\mathbb{E}|m_{ij}|^p)^{1/p} \le B\sqrt{p} \qquad \text{for all } p \ge 1 \tag{1.6}$$

where B is called the subgaussian moment. It was proved in [17] that if the entries of M are i.i.d. mean zero subgaussian random variables with unit variance, then for every $t > 0$ one has

$$\mathbb{P}\big(s_{\min}(M) \le t\big(\sqrt{m} - \sqrt{n-1}\big)\big) \le (Ct)^{m-n+1} + e^{-cm}$$

where $C, c > 0$ depend only on the subgaussian moment B. In particular, for such matrices we have

$$s_{\min}(M) \ge c_1(\sqrt{m} - \sqrt{n-1}) \qquad \text{with high probability} \tag{1.7}$$

where $c_1 > 0$ depends only on the desired probability and the subgaussian moment.

Whether (1.7) holds under weaker moment assumptions was only known in the case of square matrices. It was proved in [16] using (1.2) that (1.7) holds under the fourth moment assumption for $m = n$. Whether the same is true for arbitrary rectangular matrices under the fourth moment assumption was left open in [17]. The bottleneck of the argument occurred in Proposition 7.3 on [17] where we needed a correct bound on the spectral norm of a product of a random matrix and a fixed orthogonal projection. Such a bound was easy to get only under the subgaussian hypothesis. Theorem 1.1 of the present paper extends the argument of [17] for random matrices with bounded $(4+\varepsilon)$-th moment. It follows directly from the argument of [17] and Theorem 1.1.

Corollary 1.2 (Smallest singular value)**.** *Let A be an $m \times n$ matrix, $m \ge n$, whose entries are i.i.d. random variables with mean zero, unit variance and $(4+\varepsilon)$-th moment bounded by B. Then, for every $\delta > 0$ there exist $t > 0$ and n_0 which depend only on ε, δ and B, and such that*

$$\mathbb{P}\big(s_{\min}(M) \le t\big(\sqrt{m} - \sqrt{n-1}\big)\big) \le \delta \qquad \textit{for all } n \ge n_0.$$

To prove this result, one makes in [17] the probability estimates conditional on the event that the norm of a random matrix (denoted W in Proposition 7.3 on [17]) is small – in a similar way as this was done in [16] for square matrices.

1.4. **Outline of the argument.** To sketch the proof of Theorem 1.1, let us assume for simplicity that $m = n$ and the entries of the matrix A are identically distributed. Since the columns of A are independent, the columns $X_1, \dots, X_n$ of the matrix M are independent random vectors in $\mathbb{R}^n$. Hence $M^*M = \sum_j X_j \otimes X_j$ is a sum of independent random operators. A sharp estimate for such sums has been first proved by M. Rudelson [14]. This approach, which we develop in Section 3, leads us to the bound

$$\mathbb{E}\|M\| \le C\sqrt{n \log n}. \tag{1.8}$$

This bound is already independent of the dimension N, but is off by $\sqrt{\log n}$ from being optimal. The logarithmic term is unfortunately a limitation of this method. This term comes from M. Rudelson's result, Theorem 3.1 below, where it is needed in full generality. It would be useful to understand the situations where the logarithmic term can be removed from M. Rudelson's theorem. So far, only one such situation is known from the work of G. Aubrun [1] – when the independent random vectors X_i are uniformly distributed in an unconditional convex body.

In absence of a suitable variant of M. Rudelson's theorem without the logarithmic term, the rest of our argument will proceed to remove this term from (1.8) using the rich independence structure, which is inherited by the vectors X_i from the random matrix A (but the independence structure is encoded nontrivially via the linear transformation B, which makes the entries of X_i dependent). A more delicate application of M. Rudelson's theorem allows one to transfer the logarithmic term from the conclusion to the assumption. Namely, Theorem 3.9 establishes the optimal bound $\mathbb{E}\|M\| \le C\sqrt{n}$ in the case when all columns of B are logarithmically small, i.e. their Euclidean norm is at most $\log^{-O(1)} n$. While some columns of a general matrix B may be large, the boundedness of B implies that most columns are always logarithmically small – all but all but $n \log^{O(1)} n$ of them. So, we can remove from B the already controlled small columns, which will make B an almost square matrix. In other words, we can assume hereafter that $N = n \log^{O(1)} n$.

The advantage of almost square matrices is that the magnitude of their entries is easy to control. A simple consequence of the $(4 + \varepsilon)$-th moment hypothesis and Markov's inequality yields that the entries of $A = (a_{ij})$ satisfy $\max_{i,j} |a_{ij}| \le \sqrt{n}$ with high probability. Note that the same estimate holds for square matrices ($N = n$) under the fourth moment assumption. So, in regard to the magnitude of entries, almost square matrices are similar to exactly square matrices, for which the desired bound follows from R. Latala's result (1.2).

This prompts us to construct the proof of Theorem 1.1 for almost square matrices similarly to R. Latala's argument in [9], i.e. using fairly standard concentration of measure results in the Gauss space, coupled with delicate constructions of nets. We first decompose A into a sum of matrices which contain entries of similar magnitude. As the magnitude increases, these matrices become sparser. This quickly reduces the problem to random sparse matrices, whose entries are i.i.d. random variables valued in $\{-1, 0, 1\}$. The spectral norm of random sparse matrices was studied in [8] as a development of the work of Z. Furedi and J. Komlos [7]. However, we need to bound the spectral norm of the matrix $M = BA$ rather than A. Independence of entries is not available for M, which makes it difficult to use the known combinatorial methods based on the bounding trace of high powers of M.

To summarize, at this point we have an almost square random sparse matrix A, and we need to bound the spectral norm of $M = BA$, which is $\|M\| = \sup_x \|Mx\|_2$, where the supremum is over all unit vectors $x \in \mathbb{R}^n$. The well known method is to first fix x and bound $\|Mx\|_2$ with high probability; then take a union bound over all x in a sufficiently fine net of the unit sphere of $\mathbb{R}^n$. However, a probability bound for every fixed vector x, which follows from standard concentration inequalities, is not strong enough to make this method work. *Sparse vectors* – those which have few but large nonzero coordinates – produce worse concentration bounds than *spread vectors*, which have many but small nonzero coordinates. What helps us is that there are fewer sparse vectors on the sphere than there are spread vectors. This leads to a tradeoff between concentration and entropy, i.e. between the probability with which $\|Mx\|_2$ is nicely bounded, and the size of a net for the vectors x which achieve this probability bound. One then divides the unit Euclidean sphere in $\mathbb{R}^n$ into classes of vectors according to their "sparsity", and uses the entropy-concentration tradeoff for each class separately. This general line is already present in Latala's argument [9], and it was developed extensively in the recent work of M. Rudelson and the author on invertibility of large matrices [16], [17]. This argument is presented in Section 4, where it leads to a useful estimate for norms of sparse matrices, Corollary 4.9. With this in hand, one can quickly finish the proof of Theorem 1.1.

2. Preliminaries

2.1. **Notation.** Throughout the paper, the results are stated and proved over the field of real numbers. They are easy to generalize to complex numbers.

We denote absolute constants by $C, C_1, c, c_1 \dots$. Their values can change from line to line.

The standard inner product in $\mathbb{R}^n$ is denoted $\langle x, y\rangle$. Fir a vector $x \in \mathbb{R}^n$, we denote the cardinality of its support by $\|x\|_0 = |\{j : x_j \neq 0\}|$, the Euclidean norm by $\|x\|_2 = (\sum_j x_j^2)^{1/2}$, and the sup-norm by $\|x\|_\infty = \max_j |x_j|$. The unit Euclidean ball in $\mathbb{R}^n$ is denoted by $B_2^n = \{x : \|x\|_2 \leq 1\}$, and the unit Euclidean sphere in $\mathbb{R}^n$ is denoted by $S^{n-1} = \{x : \|x\|_2 = 1\}$.

The tensor product of vectors $x, y \in \mathbb{R}^n$ is a rank-one linear operator $x \otimes y$ on $\mathbb{R}^n$ defined as $(x \otimes y)(z) = \langle x, z\rangle y$ for $z \in \mathbb{R}^n$.

2.2. **Concentration of measure.** The method carried out in Section 4, uses concentration in the Gauss space in combination with constructions of ε-nets. Here we recall the basic facts we need.

The standard Gaussian random vector $g \in \mathbb{R}^m$ is a random vector whose coordinates are independent standard normal random variables. The following concentration inequality can be found e.g. in [11] inequality (1.5).

Theorem 2.1 (Gaussian concentration). *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz function. Let g be a standard Gaussian random vector in $\mathbb{R}^m$. Then for every $t > 0$ one has*

$$\mathbb{P}(f(g) - \mathbb{E}f(g) > t) \leq \exp(-c_0 t^2 / \|f\|_{\mathrm{Lip}}^2)$$

where $c_0 \in (0, 1)$ is an absolute constant.

As a very restrictive but useful example, Theorem 2.1 implies the following deviation inequality for sums of independent exponential random variables (which can also be derived by the more standard approach via moment generating functions).

Corollary 2.2 (Sums of exponential random variables). *Let $d = (d_1, \ldots, d_m)$ be a vector of real numbers, and let $g_1, \ldots, g_m$ be independent standard normal random variables. Then, for every $t > 0$ we have*

$$\mathbb{P}\Big\{\Big(\sum_{i=1}^m d_i^2 g_i^2\Big)^{1/2} > \|d\|_2 + t\Big\} \leq \exp(-c_0 t^2 / \|d\|_\infty^2).$$

Proof. The function $f(y) = (\sum_{i=1}^m d_i^2 y_i^2)^{1/2}$ is a Lipschitz function on $\mathbb{R}^m$, with $\|f\|_{\mathrm{Lip}} = \|d\|_\infty$. Moreover, Hölder's inequality implies that

$$\mathbb{E}f(g) = \mathbb{E}\Big(\sum_{i=1}^m d_i^2 g_i^2\Big)^{1/2} \leq \Big(\mathbb{E}\sum_{i=1}^m d_i^2 g_i^2\Big)^{1/2} = \|d\|_2.$$

Theorem 2.1 completes the proof. □

Another classical deviation inequality we will need is Bennett's inequality, see e.g. [6] Theorem 2:

Theorem 2.3 (Bennett's inequality)**.** *Let $X_1, \dots, X_N$ be independent mean zero random variables such that $|X_i| \le 1$ for all i. Consider the sum $S = X_1 + \dots + X_N$ and let $\sigma^2 := \mathrm{Var}(S)$. Then, for every $t > 0$ we have*

$$\mathbb{P}(S > t) \le \exp\big(-\sigma^2 h(t/\sigma^2)\big)$$

where $h(u) = (1+u)\log(1+u) - u$.

We will also need M. Talagrand's concentration inequality for convex Lipschitz funcitons from [13] Theorem 6.6; see also [10] Corollary 4.10 and the discussion below it.

Theorem 2.4. *Let $X_1, \dots, X_m$ be independent random variables such that $|X_i| \le K$ for all i. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a convex and 1-Lipschitz function. Then for every $t > 0$ one has*

$$\mathbb{P}\big(|f(X_1, \dots, X_m) - \mathbb{E} f(X_1, \dots, X_m)| > Kt\big) \le 4\exp(-t^2/4).$$

2.3. **Nets.** Consider a subset U of a normed space X, and let $\varepsilon > 0$. Recall that an *ε-net* of U is a subset $\mathcal{N}$ of U such that the distance from any point of U to $\mathcal{N}$ is at most ε. In other words, for every $x \in U$ there exists $y \in \mathcal{N}$ such that $\|x - y\|_X \le \varepsilon$.

The following estimate can be found e.g. in [11] Lemma 9.5.

Lemma 2.5 (Cardinality of an ε-net)**.** *Let $\varepsilon > 0$. The unit Euclidean ball B_2^n and the unit Euclidean sphere S^{n-1} in $\mathbb{R}^n$ both have ε-nets of cardinality at most $(1 + 2/\varepsilon)^n$.*

When computing norms of linear operators, ε-nets provide a convenient discretization of the problem, formalized in the next proposition.

Proposition 2.6 (Computing norm on nets)**.** *Let $A : X \to Y$ be a linear operator between normed spaces X and Y, and let $\mathcal{N}$ be an ε-net of the unit sphere $S(X)$ of X, where $\varepsilon \in (0,1)$. Then*

$$\|A\| \le \frac{1}{1-\varepsilon} \sup_{x \in \mathcal{N}} \|Ax\|_Y.$$

Proof. Every $z \in S(X)$ has the form $z = x + h$, where $x \in \mathcal{N}$ and $\|h\|_X \le \varepsilon$. Since $\|A\| = \sup_{z \in S(X)} \|Az\|_Y$, the triangle inequality yields

$$\|A\| \le \sup_{x \in \mathcal{N}} \|Ax\|_Y + \sup_{\|h\|_X \le \varepsilon} \|Ah\|_Y.$$

The last term in the right hand side is bounded by $\varepsilon\|A\|$. Therefore we have shown that

$$(1-\varepsilon)\|A\| \le \sup_{x \in \mathcal{N}} \|Ax\|_Y.$$

This completes the proof. $\square$

2.4. **Symmetrization.** We will use the standard symmetrization technique as was done in [9], see more general inequalities in e.g. [11] Section 6.1. To this end, let the matrices $A = (a_{ij})$ and B be as in Theorem 1.1. Let $A' = (a'_{ij})$ be an independent copy of A, and let ε_{ij} be independent symmetric Bernoulli random variables. Then, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}\|BA\| &= \mathbb{E}\|B(A - \mathbb{E}A')\| \le \mathbb{E}\|B(A - A')\| \\ &= \mathbb{E}\|B(\varepsilon_{ij}(a_{ij} - a'_{ij}))\| \le 2\mathbb{E}\|B(\varepsilon_{ij}a_{ij})\|.\end{aligned}$$

Therefore, we can assume without loss of generality in Theorem 1.1 that a_{ij} are symmetric random variables. Furthermore, let g_{ij} be independent standard normal random variables. Then, again by Jensen's inequality,

$$\begin{aligned}\mathbb{E}\|B(g_{ij}a_{ij})\| &= \mathbb{E}\|B(\varepsilon_{ij}|g_{ij}|a_{ij})\| \ge \mathbb{E}\|B(\varepsilon_{ij}\mathbb{E}(|g_{ij}|)a_{ij})\| \\ &= (2/\pi)^{1/2}\mathbb{E}\|B(\varepsilon_{ij}a_{ij})\|.\end{aligned}$$

Therefore,

$$\mathbb{E}\|BA\| \le (\pi/2)^{1/2}\mathbb{E}\|B(g_{ij}a_{ij})\|. \tag{2.1}$$

Conditioning on a_{ij}, we thus reduce the problem to random *gaussian* matrices.

2.5. **Truncation and conditioning.** We will need some elementary observations related to truncation and conditioning of random variables.

Lemma 2.7 (Truncation). *Let X be a non-negative random variable, and let $M > 0$, $p \ge 1$. Then*

$$\mathbb{E}X\mathbf{1}_{\{X \ge M\}} \le \frac{\mathbb{E}X^p}{M^{p-1}}.$$

Proof. Applying Hölder's inequality with the number q satisfying $\frac{1}{p} + \frac{1}{q} = 1$, we obtain

$$\mathbb{E}X\mathbf{1}_{\{X \ge M\}} \le (\mathbb{E}X^p)^{1/p}\big(\mathbb{E}\mathbf{1}_{\{X \ge M\}}\big)^{1/q}.$$

By Markov's inequality,

$$\mathbb{E}\mathbf{1}_{\{X \ge M\}} = \mathbb{P}\{X \ge M\} = \mathbb{P}\{X^p \ge M^p\} \le \frac{\mathbb{E}X^p}{M^p}.$$

Therefore

$$\mathbb{E}X\mathbf{1}_{\{X \ge M\}} \le (\mathbb{E}X^p)^{1/p}\frac{(\mathbb{E}X^p)^{1/q}}{M^{p/q}} = \frac{\mathbb{E}X^p}{M^{p/q}} = \frac{\mathbb{E}X^p}{M^{p-1}}.$$

This completes the proof. □

We will also need two elementary conditioning lemmas. In Section 4, we will need to control the maximal magnitude of the entries $M_0 = \max_{ij}|a_{ij}|$ of the random matrix A. Conditioning on M_0 will unfortunately destroy the independence of the entries. So, we will instead condition on an event $\{M_0 \le t\}$ for fixed t, which will clearly preserve the independence. This conditional argument used in the proof of Corollary 4.11 relies on the following two elementary lemmas.

Lemma 2.8. *Let X be a random variable and K be a real number. Then*

$$\mathbb{E}(X \,|\, X \le K) \le \mathbb{E}X.$$

Proof. By the law of total probability,

$$\mathbb{E}X = \mathbb{E}(X \,|\, X \le K)\, \mathbb{P}(X \le K) + \mathbb{E}(X \,|\, X > K)\, \mathbb{P}(X > K).$$

Thus $\mathbb{E}X$ is a convex combination of the numbers $a = \mathbb{E}(X \,|\, X \le K)$ and $b = \mathbb{E}(X \,|\, X > K)$. Since clearly $a \le K \le b$, we must have $a \le \mathbb{E}X \le b$. □

Lemma 2.9. *Let X, Y be non-negative random variables. Assume there exists $K, L > 0$ such that one has for every $t \ge 1$:*

$$\mathbb{E}(X^2 \,|\, Y \le t) \le K^2 t, \qquad \mathbb{P}(Y > Lt) \le \frac{1}{t^2}.$$

Then $\mathbb{E}X \le CK\sqrt{L}$.

Proof. By scaling, we can assume without loss of generality that $K = L = 1$. Thus we have for every $t \ge 1$:

$$\mathbb{E}X^2 \mathbf{1}_{\{Y \le t\}} \le \mathbb{E}(X^2 \,|\, Y \le t) \le t, \qquad \mathbb{P}(Y > t) \le \frac{1}{t^2}. \tag{2.2}$$

We consider the decomposition

$$\mathbb{E}X = \mathbb{E}X\mathbf{1}_{\{Y \le 1\}} + \sum_{k=1}^{\infty} \mathbb{E}X\mathbf{1}_{\{2^{k-1} < Y \le 2^k\}}.$$

By (2.2) and Hölder's inequality, the first term is bounded as

$$\mathbb{E}X\mathbf{1}_{\{Y \le 1\}} \le \big(\mathbb{E}X^2\mathbf{1}_{\{Y \le 1\}}\big)^{1/2} \le 1.$$

Further, by Cauchy-Schwartz inequality and using (2.2), we have

$$\begin{aligned} \mathbb{E}X\mathbf{1}_{\{2^{k-1} < Y \le 2^k\}} = \mathbb{E}X\mathbf{1}_{\{Y \le 2^k\}}\mathbf{1}_{\{Y > 2^{k-1}\}} &\le \big(\mathbb{E}X^2\mathbf{1}_{\{Y \le 2^k\}}\big)^{1/2}\big(\mathbb{P}\{Y > 2^{k-1}\}\big)^{1/2} \\ &\le 2^{k/2} \cdot \frac{1}{2^{k-1}} = 2^{1-k/2}. \end{aligned}$$

Therefore

$$\mathbb{E}X \le 1 + \sum_{k=1}^{\infty} 2^{1-k/2} \le C.$$

This completes the proof. □

2.6. **Reductions and initial observations for Theorem 1.1.** We mention two useful reductions that will make our arguments less congested. In proving Theorem 1.1, we can assume that $N \geq n = m$ by adding an appropriate number of columns of rows to the matrices A and B.

Throughout the proof of Theorem 1.1, we denote the columns of B by $B_1, \ldots, B_N$. They are non-random vectors in $\mathbb{R}^n$, which satisfy

$$\max_i \|B_i\|_2 \leq \|B\| \leq 1; \qquad \sum_{i=1}^N \|B_i\|_2^2 = \|B\|_{\mathrm{HS}}^2 \leq n\|B\| \leq n \tag{2.3}$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm of a matrix. Throughout the proof, we will only have access to the matrix B through inequalities (2.3). This explains Remark 2 following Theorem 1.1, which states that the range space of B is irrelevant as long as we control the spectral and Hilbert-Schmidt norms of B.

3. Approach via M. Rudelson's theorem

3.1. **M. Rudelson's theorem.** Our first approach, which will yield Theorem 1.1 up to a logarithmic factor, rests on the following result. Here and thereafter, by $\varepsilon_1, \varepsilon_2, \ldots$ we denote independent symmetric Bernoulli random variables, i.e. independent random variables such that $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$.

Theorem 3.1 (M. Rudelson [14]). *Let $u_1, \ldots, u_M$ be vectors in $\mathbb{R}^m$. Then, for every $p \geq 1$, one has*

$$\Big(\mathbb{E}\Big\|\sum_{i=1}^M \varepsilon_i u_i \otimes u_i\Big\|^p\Big)^{1/p} \leq C(\sqrt{p} + \sqrt{\log m}) \cdot \max_i \|u_i\|_2 \cdot \Big\|\sum_{i=1}^M u_i \otimes u_i\Big\|^{1/2}.$$

In particular, for every $t > 0$, with probability at least $1 - me^{-ct^2}$ one has

$$\Big\|\sum_{i=1}^M \varepsilon_i u_i \otimes u_i\Big\| \leq t \cdot \max_i \|u_i\|_2 \cdot \Big\|\sum_{i=1}^M u_i \otimes u_i\Big\|^{1/2}.$$

The first estimate is taken from [14], inequality (3.4). The second estimate can be easily derived from it using the following elementary lemma:

Lemma 3.2. *Suppose a non-negative random variable X satisfies for some $m \geq 1$:*

$$(\mathbb{E}X^p)^{1/p} \leq \sqrt{p} + \sqrt{\log m} \quad \textit{for every } p \geq 1..$$

Then

$$\mathbb{P}(X \geq t) \leq me^{-ct^2} \quad \textit{for every } t > 0.$$

Proof. Suppose first that $t \ge \sqrt{\log m}$. Let $p := t^2$. Then $\sqrt{p} \ge \sqrt{\log m}$, and the hypothesis gives $(\mathbb{E}X^p)^{1/p} \le 2\sqrt{p}$. By Markov's inequality,

$$\mathbb{P}(X \ge 2et) = \mathbb{P}(X^p \ge 2et) \le \frac{(2\sqrt{p})^p}{(2et)^p} = e^{-t^2}.$$

Therefore, for every $t > 0$ one has

$$\mathbb{P}(X \ge 2et) \le m e^{-ct^2} \tag{3.1}$$

because if $t < \sqrt{\log m}$ then the right hand side of (3.1) is smaller than 1, which makes the inequality trivial. This completes the proof. □

The next lemma is a consequence of M. Rudelson's Theorem 3.1 and a standard symmetrization argument.

Lemma 3.3. *Let $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^m$ such that*

$$\|\mathbb{E}X_j \otimes X_j\| \le 1 \qquad \text{for every } j. \tag{3.2}$$

Then

$$\mathbb{E}\Big\| \sum_{j=1}^n X_j \otimes X_j \Big\| \le Cn + C\log(m)\,\mathbb{E}\max_j \|X_j\|_2^2.$$

Proof. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent symmetric Bernoulli random variables. By the triangle inequality, the standard symmetrization argument (see e.g. [11] Lemma 6.3), and the assumption, we have

$$E := \mathbb{E}\Big\| \sum_{j=1}^n X_j \otimes X_j \Big\| \le \mathbb{E}\Big\| \sum_{j=1}^n (X_j \otimes X_j - \mathbb{E}X_j \otimes X_j) \Big\| + \Big\| \sum_{j=1}^n \mathbb{E}X_j \otimes X_j \Big\|$$
$$\le 2\mathbb{E}\Big\| \sum_{j=1}^n \varepsilon_j X_j \otimes X_j \Big\| + n.$$

Condition on the random variables $X_1, \ldots, X_n$, and apply Theorem 3.1. Writing $\mathbb{E}_\varepsilon$ to denote the conditional expectation (i.e. the expectation with respect to the random variables $\varepsilon_1, \ldots, \varepsilon_n$), we have

$$\mathbb{E}_\varepsilon \Big\| \sum_{j=1}^n \varepsilon_j X_j \otimes X_j \Big\| \le C\sqrt{\log m} \cdot \max_j \|X_j\|_2 \cdot \Big\| \sum_{j=1}^n X_j \otimes X_j \Big\|^{1/2}.$$

Now we take expectation with respect to $X_1, \ldots, X_n$ and use Cauchy-Schwarz inequality to get

$$E \le C\sqrt{\log m} \cdot \big(\mathbb{E}\max_j \|X_j\|_2^2\big)^{1/2} \cdot E^{1/2} + n.$$

The conclusion of the lemma follows. □

3.2. **Theorem 1.1 up to a logarithmic term.** We now state a version of Theorem 1.1 with a logarithmic factor.

Proposition 3.4. *Let A be an $N \times n$ random matrix whose entries are independent random variables with mean zero and 4-th moment bounded by 1. Let B be an $n \times N$ matrix such that $\|B\| \le 1$. Then*

$$\mathbb{E}\|BA\| \le C\sqrt{n \log n}.$$

The proof will need two auxiliary lemmas. Recall that $B_1, \dots, B_N$ denote the columns of the matrix B.

Lemma 3.5. *Let $a_1, \dots, a_N$ be independent random variables with mean zero and 4-th moment bounded by 1. Consider the random vector X in $\mathbb{R}^n$ defined as*

$$X = \sum_{i=1}^N a_i B_i.$$

Then

$$\mathbb{E}\|X\|_2^2 \le n, \qquad \mathrm{Var}(\|X\|_2^2) \le 3n.$$

Proof. The estimate on the expectation follows easily from (2.3):

$$\mathbb{E}\|X\|_2^2 = \sum_{i=1}^N \mathbb{E}(a_i^2)\|B_i\|_2^2 \le \sum_{i=1}^N \|B_i\|_2^2 \le n. \tag{3.3}$$

To estimate the variance, we need to compute

$$\mathbb{E}\|X\|_2^4 = \mathbb{E}\langle X, X\rangle^2 = \sum_{i,j,k,l=1}^N \mathbb{E}(a_i a_j a_k a_l)\langle B_i, B_j\rangle\langle B_k, B_l\rangle.$$

By independence and the mean zero assumption, the only nonzero terms in this sum are those for which $i = j; k = l$ or $i = k; j = l$ or $i = l; j = k$. Therefore

$$\begin{aligned}\mathbb{E}\|X\|_2^4 &= \sum_{i,j=1}^N \mathbb{E}(a_i^2 a_j^2)\|B_i\|^2\|B_j\|^2 + 2\sum_{i,j=1}^N \mathbb{E}(a_i^2 a_j^2)\langle B_i, B_j\rangle^2 \\ &= \sum_{i=1}^N \mathbb{E}(a_i^4)\|B_i\|^4 + \sum_{\substack{i,j=1 \\ i\neq j}}^N \mathbb{E}(a_i^2)\mathbb{E}(a_j^2)\|B_i\|_2^2\|B_j\|_2^2 + 2\sum_{i,j=1}^N \mathbb{E}(a_i^2 a_j^2)\langle B_i, B_j\rangle^2 \\ &=: I_1 + I_2 + I_3.\end{aligned}$$

By the fourth moment assumption and using (2.3) we have

$$I_1 \le \sum_{i=1}^N \|B_i\|_2^4 \le \max_i(\|B_i\|_2^2)\sum_{i=1}^N \|B_i\|_2^2 \le n$$

Squaring the sum in (3.3), we see that

$$I_2 \le (\mathbb{E}\|X\|_2^2)^2.$$

Finally, since by Cauchy-Schwarz inequality $\mathbb{E}(a_i^2 a_j^2) \le \sqrt{\mathbb{E}(a_i^4)\mathbb{E}(a_j^4)} \le 1$, and using (2.3) again, we obtain

$$I_3 \le 2 \sum_{i,j=1}^{N} \langle B_i, B_j \rangle^2 = 2\|B^*B\|_{\mathrm{HS}}^2 \le 2\|B^*\|^2 \|B\|_{\mathrm{HS}}^2 = 2\|B\|^2 \|B\|_{\mathrm{HS}}^2 \le 2n.$$

Putting all this together, we obtain

$$\mathrm{Var}(\|X\|_2^2) = \mathbb{E}\|X\|_2^4 - (\mathbb{E}\|X\|_2^2)^2 \le I_1 + I_3 \le 3n.$$

This completes the proof. □

Lemma 3.6. *Let A and B be matrices as in Proposition 3.4. Let $X_1, \dots, X_n \in \mathbb{R}^n$ be the columns of the matrix BA. Then*

$$\mathbb{E} \max_{j=1,\dots,n} \|X_j\|_2^2 \le Cn.$$

Remark. This result says that all columns of the matrix BA have norm $O(\sqrt{n})$ with high probability. Since the spectral norm of a matrix is bounded below by the norm of any column, this result is a necessary step in proving our desired estimate $\|BA\| = O(\sqrt{n})$.

Proof. Let, as usual, $B_1, \dots, B_N \in \mathbb{R}^n$ denote the columns of the matrix B, and let $X_1, \dots, X_n \in \mathbb{R}^n$ denote the columns of the matrix BA. Then

$$X_j = \sum_{i=1}^{N} a_{ij} B_i, \qquad j = 1, \dots, n. \tag{3.4}$$

Let us fix $j \in \{1, \dots, n\}$ and use Lemma 3.5. This gives

$$\mathbb{E}\|X_j\|_2^2 \le n, \qquad \mathrm{Var}(\|X_j\|_2^2) \le 3n. \tag{3.5}$$

Now we use Chebychev's inequality, which states that for a random variable Z with $\sigma^2 = \mathrm{Var}(Z)$ and for an arbitrary $k > 0$, one has

$$\mathbb{P}(|Z - \mathbb{E}Z| > k\sigma) \le \frac{1}{k^2}.$$

Let $t > 0$ be arbitrary. Using Chebychev's inequality along with (3.5) for $Z = \|X_j\|_2^2$, $k = t\sqrt{n}$, we obtain

$$\mathbb{P}\big(\|X_j\|_2^2 > (1 + \sqrt{3}\, t)n\big) \le \frac{1}{t^2 n}.$$

Taking the union bound over all $j = 1, \dots, n$, we conclude that

$$\mathbb{P}\big(\max_{j=1,\dots,n} \|X_j\|_2^2 > (1 + \sqrt{3}\, t)n\big) \le n \cdot \frac{1}{t^2 n} = \frac{1}{t^2}.$$

Integration completes the proof. □

Proof of Proposition 3.4. Let $X_1, \dots, X_n \in \mathbb{R}^n$ denote the columns of the matrix BA. We are going to apply Lemma 3.3. In order to check that condition (3.2) holds, we consider an arbitrary vector $x \in S^{n-1}$ and use representation (3.4) to compute

$$\begin{aligned}\mathbb{E}\langle X_j, x\rangle^2 &= \mathbb{E}\Big(\sum_{i=1}^N a_{ij}\langle B_i, x\rangle\Big)^2 = \sum_{i=1}^N \mathbb{E}(a_{ij}^2)\langle B_i, x\rangle^2 \le \sum_{i=1}^N \langle B_i, x\rangle^2 \\ &= \|B^* x\|_2^2 \le \|B^*\| = \|B\| \le 1.\end{aligned}$$

This shows that condition (3.2) holds. Lemma 3.3 then gives

$$\mathbb{E}\|BA\|^2 = \mathbb{E}\Big\|\sum_{j=1}^n X_j \otimes X_j\Big\| \le Cn + C\log(n)\,\mathbb{E}\max_{j=1,\dots,n}\|X_j\|_2^2.$$

Estimating the maximum in the right hand side using Lemma 3.6, we conclude that

$$\mathbb{E}\|BA\|^2 \le C_1 n \log n.$$

This completes the proof. □

3.3. **Tradeoff between the matrix norm and the magnitude of entries.** We would like now to gain more control over the logarithmic factor than we have in Proposition 3.4. Our next result establishes a tradeoff between the logarithmic factor and the magnitude of the matrices A, B. It will be used in the proof of Theorem 3.9.

Proposition 3.7. *Let A be an $N \times n$ matrix whose entries are random independent variables a_{ij} with mean zero and such that*

$$\mathbb{E}a_{ij}^2 \le 1, \qquad |a_{ij}| \le a \qquad \textit{for every } i, j.$$

Let B be an $n \times N$ matrix such that $\|B\| \le 1$, and whose columns satisfy

$$\|B_i\|_2 \le b \qquad \textit{for every } i.$$

Then

$$\mathbb{E}\|BA\| \le C(1 + ab^{1/2}\log^{1/4} n)\sqrt{n}.$$

The proof will again be based on M. Rudelson's Theorem 3.1, although this time we use Rudelson's theorem in a more delicate way:

Lemma 3.8. *Under the assumptions of Proposition 3.7, we have*

$$\mathbb{E}\max_{j=1,\dots,n}\Big\|\sum_{i=1}^N a_{ij}^2 B_i \otimes B_i\Big\| \le C(1 + a^2 b\sqrt{\log n}).$$

Proof. Fix $j \in \{1, \dots, n\}$. Let $\mu_{ij}^2 := \mathbb{E} a_{ij}^2$. By the triangle inequality,

$$\Big\| \sum_{i=1}^N a_{ij}^2 B_i \otimes B_i \Big\| \le \Big\| \sum_{i=1}^N (a_{ij}^2 - \mu_{ij}^2) B_i \otimes B_i \Big\| + \Big\| \sum_{i=1}^N \mu_{ij}^2 B_i \otimes B_i \Big\|. \tag{3.6}$$

Since $0 \le \mu_{ij}^2 \le 1$ and

$$\Big\| \sum_{i=1}^N B_i \otimes B_i \Big\| = \|B^* B\| \le \|B\|^2 \le 1, \tag{3.7}$$

we have

$$\Big\| \sum_{i=1}^N \mu_{ij}^2 B_i \otimes B_i \Big\| \le \Big\| \sum_{i=1}^N B_i \otimes B_i \Big\| \le 1. \tag{3.8}$$

Next, clearly $\mu_{ij}^2 \le a^2$, so

$$\mathbb{E}(a_{ij}^2 - \mu_{ij}^2) = 0, \qquad |a_{ij}^2 - \mu_{ij}^2| \le 2a^2.$$

Let ε_{ij} be independent symmetric Bernoulli random variables. Using the standard symmetrization ([11] Lemma 6.3) and the contraction principle ([11] Theorem 4.4), we obtain

$$\mathbb{E} \max_{j=1,\dots,n} \Big\| \sum_{i=1}^N (a_{ij}^2 - \mu_{ij}^2) B_i \otimes B_i \Big\| \le 2a^2 \, \mathbb{E} \max_{j=1,\dots,n} \Big\| \sum_{i=1}^N \varepsilon_{ij} B_i \otimes B_i \Big\|. \tag{3.9}$$

Let $t > 0$. By the second part of M. Rudelson's Theorem 3.1 and taking the union bound over n random variables, we conclude that, with probability at least $1 - n^2 e^{-ct^2}$, we have

$$\max_{j=1,\dots,n} \Big\| \sum_{i=1}^N \varepsilon_{ij} B_i \otimes B_i \Big\| \le t \cdot \max_{i=1,\dots,N} \|B_i\|_2 \cdot \Big\| \sum_{i=1}^N B_i \otimes B_i \Big\| \le tb$$

The second estimate follows from (3.7) and since $\max_i \|B_i\|_2 \le b$ by the hypothesis.

Let $s > 0$ be arbitrary. We apply the above estimate for $t = C(\sqrt{\log n} + s)$, where $C = 1/\sqrt{c}$. This shows that, with probability at least $1 - e^{-s^2}$, one has

$$\max_{j=1,\dots,n} \Big\| \sum_{i=1}^N \varepsilon_{ij} B_i \otimes B_i \Big\| \le Cb(\sqrt{\log n} + s).$$

Integration implies that

$$\mathbb{E} \max_{j=1,\dots,n} \Big\| \sum_{i=1}^N \varepsilon_{ij} B_i \otimes B_i \Big\| \le Cb\sqrt{\log n}.$$

Putting this into (3.9) and, together with (3.8), back into (3.6), we complete the proof. □

Proof of Proposition 3.7. By the symmetrization argument (see (2.1)), we can assume that the entries of the matrix A are $g_{ij}a_{ij}$, where a_{ij} are random variables satisfying the assumptions of the proposition, and g_{ij} are independent standard normal random variables. We will write $\mathbb{E}_g$, $\mathbb{P}_g$ when we take expectations and probability estimates with respect to (g_{ij}) (i.e. conditioned on (a_{ij})), and we write $\mathbb{E}_a$ to denote the expectation with respect to (a_{ij}).

By Lemma 3.8, the random variable

$$K^2 := \max_{j=1,\dots,n} \Big\| \sum_{i=1}^N a_{ij}^2 B_i \otimes B_i \Big\|$$

which does not depend on the random variables (g_{ij}), has expectation

$$\mathbb{E}_a(K^2) \le C(1 + a^2 b\sqrt{\log n}). \tag{3.10}$$

We condition on the random variables (a_{ij}); this fixes a value of K.

Let $X_1, \dots, X_n \in \mathbb{R}^n$ denote the columns of the matrix BA; then

$$X_j = \sum_{i=1}^N g_{ij} a_{ij} B_i, \qquad j = 1, \dots, n.$$

Consider a $(1/2)$-net $\mathcal{N}$ of the unit Euclidean sphere S^{n-1} of cardinality $|\mathcal{N}| \le 5^n$, which exists by Lemma 2.5. Using Proposition 2.6, we have

$$\|BA\|^2 = \|(BA)^*\|^2 \le 4 \max_{x\in\mathcal{N}} \|(BA)^* x\|_2^2 = 4 \max_{x\in\mathcal{N}} \sum_{j=1}^n \langle X_j, x\rangle^2. \tag{3.11}$$

Fix $x \in \mathcal{N}$. For every $j = 1, \dots, n$, the random variable

$$\langle X_j, x\rangle = \sum_{i=1}^N g_{ij} \langle a_{ij} B_i, x\rangle$$

is a Gaussian random variable with mean zero and variance

$$\sum_{i=1}^N \langle a_{ij} B_i, x\rangle^2 \le \Big\| \sum_{i=1}^N a_{ij}^2 B_i \otimes B_i \Big\| \le K^2.$$

(To obtain the first inequality, take the supremum over $x \in S^{n-1}$). Therefore, by Corollary 2.2 with $d_i = K$, we have for every $t > 0$:

$$\mathbb{P}_g \Big\{ \Big(\sum_{j=1}^n \langle X_j, x\rangle^2 \Big)^{1/2} > K\sqrt{n} + t \Big\} \le e^{-t^2/K^2}.$$

Let $s > 0$ be arbitrary. The previous estimate for $t = sK\sqrt{n}$ gives

$$\mathbb{P}_g\Big\{\Big(\sum_{j=1}^n \langle X_j, x\rangle^2\Big)^{1/2} > (1+s)K\sqrt{n}\Big\} \le e^{-s^2 n}.$$

Taking the union bound over $x \in \mathcal{N}$ and using (3.11), we obtain

$$\mathbb{P}_g\big\{\|BA\| > 2(1+s)K\sqrt{n}\big\} \le |\mathcal{N}|e^{-s^2 n} = 5^n e^{-s^2 n} \le e^{(2-s^2)n}.$$

Integration yields

$$\mathbb{E}_g\|BA\| \le CK\sqrt{n}.$$

Finally, we take expectation with respect to the random variables (a_{ij}) and use (3.10) to conclude that

$$\mathbb{E}\|BA\| \le C\mathbb{E}_a(K)\sqrt{n} \le C_1\big(1 + a^2 b\sqrt{\log n}\big)^{1/2}\sqrt{n}.$$

This completes the proof. □

3.4. **Theorem 1.1 for logarithmically small columns.** Our next step is to combine Propositions 3.4 and 3.7 and obtain a weaker version of the main Theorem 1.1 – this time with the correct bound $O(\sqrt{n})$ on the norm, but under the additional assumption that the columns of the matrix B are logarithmically small.

Theorem 3.9. *Let A be an $N \times n$ random matrix whose entries are independent random variables with mean zero and $(4+\varepsilon)$-th moment bounded by 1. Let B be an $n \times N$ matrix such that $\|B\| \le 1$, and whose columns satisfy*

$$\|B_i\|_2 \le \log^{-\frac{1}{2}-\frac{1}{\varepsilon}} n \qquad \text{for every } i.$$

Then

$$\mathbb{E}\|BA\| \le C\sqrt{n}.$$

Proof. By the symmetrization argument described in Section 2, we can assume without loss of generality that all entries a_{ij} of the matrix $A = (a_{ij})$ are symmetric random variables. Let

$$a := \log^{\frac{1}{2\varepsilon}} n.$$

We decompose every entry of the matrix A according to its absolute value as

$$\bar{a}_{ij} := a_{ij}\mathbf{1}_{\{|a_{ij}| \le a\}}, \qquad \tilde{a}_{ij} := a_{ij}\mathbf{1}_{\{|a_{ij}| > a\}}.$$

Then all random variables $\bar{a}_{ij}$ and $\tilde{a}_{ij}$ have mean zero, and we have the following decomposition of matrices:

$$BA = B\bar{A} + B\tilde{A}, \qquad \text{where } \bar{A} = (\bar{a}_{ij}),\ \tilde{A} = (\tilde{a}_{ij}).$$

The norm of $B\tilde{A}$ can be bounded using Proposition 3.4. Indeed, by the Truncation Lemma 2.7 with $p = 1 + \varepsilon/4$, we have

$$\mathbb{E}\tilde{a}_{ij}^4 = \mathbb{E}a_{ij}^4 \mathbf{1}_{\{a_{ij}^4 > a^4\}} \leq \frac{\mathbb{E}a_{ij}^{4+\varepsilon}}{a^\varepsilon} \leq a^{-\varepsilon},$$

where the last inequality follows from the fourth moment hypothesis. Therefore, the matrix $a^\varepsilon A$ satisfies the hypothesis of Proposition 3.4, which then yields

$$\mathbb{E}\|B\tilde{A}\| \leq Ca^{-\varepsilon}\sqrt{n \log n} \leq C\sqrt{n},$$

where the last inequality follows by our choice of a.

The norm of $B\bar{A}$ can be bounded using Proposition 3.7, which we can apply with a as above and $b = \log^{-\frac{1}{2}-\frac{1}{\varepsilon}} n$. This gives

$$\mathbb{E}\|B\bar{A}\| \leq C(1 + ab^{1/2}\log^{1/4} n)\sqrt{n} \leq 2C\sqrt{n},$$

where the last inequality follows by our choice of a and b.

Putting the two estimates together, we conclude by the triangle inequality that

$$\mathbb{E}\|BA\| \leq \mathbb{E}\|B\bar{A}\| + \mathbb{E}\|B\tilde{A}\| \leq C'\sqrt{n}.$$

This completes the proof. □

4. Approach via concentration

In this section, we develop an alternative way to bound the norm of BA, which rests on Gaussian concentration inequalities and elaborate choice of ε-nets. The main technical result of this section is the following theorem, which, like Theorem 3.9, gives the correct bound $O(\sqrt{n})$ under some boundedness assumptions on the entries of A.

Theorem 4.1. *Let $\varepsilon > 0$ and $M \geq 1$. Let A be an $N \times n$ random matrix with $\log N \leq Mn$, and whose entries are independent random variables a_{ij} with mean zero and such that*

$$\mathbb{E}|a_{ij}|^{2+\varepsilon} \leq 1, \qquad |a_{ij}| \leq \Big(\frac{Mn}{\log N}\Big)^{\frac{1}{2+\varepsilon}} \qquad \textit{for every } i, j.$$

Let B be an $n \times N$ matrix such that $\|B\| \leq 1$. Then

$$\mathbb{E}\|BA\| \leq C(\varepsilon)\sqrt{Mn}$$

where $C(\varepsilon)$ depends only on ε.

Remarks. **1.** If the entries a_{ij} have bounded $(4+\varepsilon)$-th moment, it is easy to check that $\max_{ij} a_{ij} \sim (nN)^{\frac{1}{4+\varepsilon}}$ holds with high probability. Therefore, under the $(4+\varepsilon)$-th moment assumption, the hypotheses of Theorem 4.1 are satisfied for almost square matrices, i.e. those for which $N \leq n^{1+c\varepsilon}$. This will quickly yield the main Theorem 1.1 for almost square matrices, see Corollary 4.11 below.

2. The hypotheses of Theorem 4.1 are almost sharp when $N \sim n$. Indeed, let us assume for simplicity that the random variables a_{ij} are identically distributed and B is the identity matrix. The $(2+\varepsilon)$-th moment hypothesis is almost sharp: if $\mathbb{E}a_{ij}^2 \gg 1$ then $(\mathbb{E}\|A\|^2)^{1/2} \geq \big(\frac{1}{n}\|A\|_{\mathrm{HS}}^2\big)^{1/2} \gg \sqrt{n}$. Also, the boundedness hypothesis is almost sharp, since $\|A\| \geq \max_{i,j}|a_{ij}|$.

3. Using M. Talagrand's concentration result, Theorem 2.4, one can also obtains tail bounds for the norm $\|BA\|$:

Corollary 4.2. *Under the assumptions of Theorem 4.1, one has for every $t > 0$:*

$$\mathbb{P}\big(\|BA\| > (C(\varepsilon) + t)\sqrt{Mn}\big) \leq e^{-ct^2}.$$

In particular, one has for every $q \geq 1$:

$$(\mathbb{E}\|BA\|^q)^{1/q} \leq C_0(\varepsilon)\sqrt{qMn}.$$

Proof. We can consider the $N \times n$ matrix A as a vector in $\mathbb{R}^{Nn}$. The Euclidean norm of such a vector equals the Hilbert-Schmidt norm $\|A\|_{\mathrm{HS}}$. Since $\|BA\| \leq \|B\|\|A\| \leq 1 \cdot \|A\|_{\mathrm{HS}}$, the function $f : \mathbb{R}^{Nn} \to \mathbb{R}$ defined by $f(A) = \|BA\|$ is 1-Lipschitz and convex. Since we have $|a_{ij}| \leq \sqrt{Mn}$ for all i, j by the assumptions, M. Talagrand's Theorem 2.4 gives

$$\mathbb{P}\big(\|BA\| - \mathbb{E}\|BA\| > t\sqrt{Mn}\big) \leq 4e^{-t^2/4}, \qquad t > 0.$$

The estimate for $\mathbb{E}\|BA\|$ in Theorem 4.1 completes the proof. □

4.1. **Sparse matrices: rows and columns.** Theorem 4.1 will follow from our analysis of sparse matrices. Namely, we will decompose the entries a_{ij} according to their magnitude; as the magnitude increases, the moment assumptions will ensure that there will be fewer such entries, i.e. the resulting matrix becomes sparser.

We start with an elementary lemma, which will help us analyze the magnitude of the rows and columns of the matrix BA when A is a sparse matrix.

Lemma 4.3. *Let a_{ij} be independent random variables, $i = 1, \ldots, N$, $j = 1, \ldots, n$. Let $p \in (0, 1]$, and suppose that*

$$\mathbb{E}a_{ij}^2 \leq p, \qquad |a_{ij}| \leq 1 \qquad \text{for every } i, j.$$

Let B be an $n \times N$ matrix such that $\|B\| \le 1$, whose columns are denoted B_i. Then

$$\mathbb{E} \max_{i=1,\dots,N} \sum_{j=1}^{n} a_{ij}^2 \le C(np + \log N), \tag{4.1}$$

$$\mathbb{E} \max_{j=1,\dots,n} \sum_{i=1}^{N} a_{ij}^2 \|B_i\|_2^2 \le C(np + \log n). \tag{4.2}$$

Remark. The test case for this lemma, as well as for most of the results that follow, is the random variables a_{ij} with values in $\{-1, 0, 1\}$ and such that $\mathbb{P}(a_{ij} \neq 0) = p$. The $N \times n$ random matrix $A = (a_{ij})$ will then become sparser as we decrease p; it will have on average np nonzero entries per row. Estimate (4.1) gives a bound on the Euclidean norm of all rows of A.

Proof. We will only prove inequality (4.2); the proof of inequality (4.1) is similar. By the assumptions, we have

$$\mathrm{Var}(a_{ij}^2) \le \mathbb{E}a_{ij}^4 \le \mathbb{E}a_{ij}^2 \le p \qquad \text{for every } i, j.$$

Also, recall that (2.3) give

$$\sum_{i=1}^{N} \|B_i\|_2^2 \le n, \qquad \sum_{i=1}^{N} \|B_i\|_2^4 \le \max_i \|B_i\|_2^2 \cdot \sum_{i=1}^{N} \|B_i\|_2^2 \le n.$$

Consider the sums of independent random variables

$$S_j := \sum_{i=1}^{N} a_{ij}^2 \|B_i\|_2^2, \qquad j = 1, \dots, n.$$

The above estimates show that for every j we have

$$\mathbb{E}S_j = \sum_{i=1}^{N} \mathbb{E}(a_{ij}^2) \|B_i\|_2^2 \le np, \qquad \mathrm{Var}(S_j) = \sum_{i=1}^{N} \mathrm{Var}(a_{ij}^2) \|B_i\|_2^4 \le np.$$

We apply Bennett's inequality, Theorem 2.3, for $X_i = \frac{1}{2}\big(a_{ij}^2 - \mathbb{E}a_{ij}^2\big)\|B_i\|_2$, which clearly satisfy $|X_j| \le 1$ because $|a_{ij}| \le 1$ and $\|B_i\|_2 \le 1$ by (2.3). We obtain

$$\mathbb{P}\Big\{\frac{1}{2}(S_j - \mathbb{E}S_j) > t\Big\} \le \exp\big(-\sigma^2 h(t/\sigma^2)\big) \tag{4.3}$$

where $\mathbb{E}S_j \le np$ and $\sigma^2 = \mathrm{Var}(S_j) \le np$. Note that $h(x) \ge cx$ for $x \ge 1$, where c is some positive absolute constant. Therefore, if $t \ge np$, then $\sigma^2 h(t/\sigma^2) \ge ct$, so (4.3) yields

$$\mathbb{P}\{S_j > 3t\} \le e^{-ct} \qquad \text{for } t \ge np.$$

Taking the union bound over all j, we conclude that

$$\mathbb{P}\Big\{\max_{j=1,\dots,n} S_j > 3t\Big\} \le ne^{-ct} \qquad \text{for } t \ge np.$$

Now let $s \ge 1$ be arbitrary, and use the last inequality for $t = (np + \log n)s$. We obtain

$$\mathbb{P}\Big\{\max_{j=1,\dots,n} S_j > 3(np + \log n)s\Big\} \le n^{1-3cs} e^{-3cnps} \le n^{1-3cs}.$$

Integration yields

$$\mathbb{E} \max_{j=1,\dots,n} S_j \le C(np + \log n).$$

This completes the proof of (4.2). □

The estimates in Lemma 4.3 motivate us to consider a class of $N \times n$ matrices $A = (a_{ij})$ whose entries satisfy the following inequalities for some parameters $p \in (0, 1]$ and $K > 0$:

$$\begin{aligned}
&\max_{i,j} |a_{ij}| \le 1; \\
&\max_{i=1,\dots,N} \Big(\sum_{j=1}^n a_{ij}^2\Big)^{1/2} \le K\sqrt{np + \log N}; \qquad (4.4)\\
&\max_{j=1,\dots,n} \Big(\sum_{i=1}^N a_{ij}^2 \|B_i\|_2^2\Big)^{1/2} \le K\sqrt{np + \log n}.
\end{aligned}$$

We have proved that for random matrices whose entries satisfy $|a_{ij}| \le 1$ and $Ea_{ij}^2 \le p$, conditions (4.4) hold with a random parameter K that satisfies $\mathbb{E}K \le C$.

4.2. **Concentration for a fixed vector.** Our goal will be to estimate the magnitude of $\|BA\|$ where $A = (g_{ij} a_{ij})$, where g_{ij} are independent standard normal random variables, and a_{ij} are fixed numbers that satisfy conditions (4.4). Such an estimate will be established in Proposition 4.8 below. By the standard symmetrization, the same estimate will hold true if $A = (a_{ij})$ is a random matrix with entries as in Lemma 4.3. This will be done in Corollary 4.9. Finally, Theorem 4.1 will be deduced from this by decomposing the entries of a random matrix according to their magnitude.

Our first step toward this goal is to check the magnitude of $\|BAx\|_2$ for a fixed vector x.

Lemma 4.4. *Let $A = (g_{ij} a_{ij})$ be an $N \times n$ matrix, where g_{ij} are independent standard normal random variables, and a_{ij} are fixed numbers that satisfy conditions (4.4). Let B be an $n \times N$ matrix such that $\|B\| \le 1$. Then, for every vector $x \in B_2^n$ we have*

$$\mathbb{E}\|BAx\|_2 \le K\sqrt{np + \log n}.$$

Proof. Denoting as usual the columns of B by B_i, we have

$$BAx = \sum_{i=1}^{N} \Big(\sum_{j=1}^{n} g_{ij} a_{ij} x_j \Big) B_i.$$

Since $\|x\|_2 \le 1$ and using the last condition in (4.4), we have

$$\begin{aligned}\mathbb{E}\|BAx\|_2^2 &= \sum_{i=1}^{N} \sum_{j=1}^{n} a_{ij}^2 x_j^2 \|B_i\|_2^2 \\ &= \sum_{j=1}^{n} \Big(\sum_{i=1}^{N} a_{ij}^2 \|B_i\|_2^2 \Big) x_j^2 \\ &\le \max_{j=1,\dots,n} \sum_{i=1}^{N} a_{ij}^2 \|B_i\|_2^2 \le K^2(np + \log n).\end{aligned}$$

This completes the proof. □

We will now strengthen Lemma 4.4 into a deviation inequality for $\|BAx\|_2$. This is a simple consequence of the Gaussian concentration, Theorem 2.1. This deviation inequality is universal in that it holds for any vector x; in the sequel we will need more delicate inequalities that depend on the distribution of the coordinates in x.

Lemma 4.5 (Universal deviation)**.** *Let A and B be matrices as in Lemma 4.4. Then, for every vector $x \in B_2^n$ we have*

$$\mathbb{P}\big\{ \|BAx\|_2 > K\sqrt{np + \log n} + t \big\} \le e^{-c_0 t^2}. \tag{4.5}$$

Proof. As in the proof of Lemma 4.4, we write

$$BAx = \sum_{i=1}^{N} \Big(\sum_{j=1}^{n} g_{ij} a_{ij} x_j \Big) B_i$$

where B_i are the columns of the matrix B. Therefore, the random vector BAx is distributed identically with the random vector

$$\sum_{i=1}^{N} g_i h_i B_i, \qquad \text{where } h_i = \Big(\sum_{j=1}^{n} a_{ij}^2 x_j^2 \Big)^{1/2}$$

and where g_i are independent standard normal random variables. Since all $|a_{ij}| \le 1$ by conditions (4.4), and $\|x\|_2 \le 1$ by the assumptions, we have

$$0 \le h_i \le 1, \qquad i = 1, \dots, N.$$

Consider the map $f : \mathbb{R}^N \to \mathbb{R}$ given by

$$f(y) = \Big\| \sum_{i=1}^N y_i h_i B_i \Big\|_2.$$

Its Lipschitz norm equals

$$\|f\|_{\mathrm{Lip}} = \Big\| \sum_{i=1}^N h_i^2 B_i \otimes B_i \Big\|^{1/2} \le \max_i |h_i| \cdot \Big\| \sum_{i=1}^N B_i \otimes B_i \Big\|^{1/2} \le 1 \cdot \|B\| \le 1.$$

Then the Gaussian concentration, Theorem 2.1, gives for every $t > 0$:

$$\mathbb{P}(f(g) - \mathbb{E} f(g) > t) \le \exp(-c_0 t^2),$$

where $g = (g_1, \dots, g_N)$. Since as we noted above, $f(g)$ is distributed identically with $\|BAx\|_2$, Lemma 4.4 completes the proof. □

4.3. **Control of sparse vectors.** Since the spectral norm of BA is the supremum of $\|BAx\|_2$ over all $x \in S^{n-1}$, the result of Lemma 4.5 suggests that $\mathbb{E}\|BA\| \lesssim \sqrt{np + \log N}$ should be true. However, the deviation inequality in Lemma 4.5 is not strong enough to prove this bound. This is because the metric entropy of the sphere, measured e.g. as the cardinality of its $\frac{1}{2}$-net, is e^{cn}. If we are to make the bound on $\|BAx\|_2$ uniform over the net, we would need the probability estimate in (4.5) at most e^{-cn} (to allow a room for the union bound over e^{cn} points x in the net). This however would force us to make $t \sim \sqrt{n}$ or larger, so the best bound we can get this way is $\mathbb{E}\|BA\|_2 \lesssim \sqrt{n}$. This bound is too weak as it ignores the last two assumptions in (4.4).

Nevertheless, the bound in Lemma 4.5 can be made uniform over a set of sparse vectors, whose metric entropy is smaller than that of the whole sphere:

Proposition 4.6 (Sparse vectors)**.** *Let A and B be matrices as in Lemma 4.4. There exists an absolute constant $c > 0$ such that the following holds. Consider the set of vectors*

$$B_{2,0} := \Big\{ x \in \mathbb{R}^n,\ \|x\|_2 \le 1,\ \|x\|_0 \le cnp/\log(e/p) \Big\}.$$

Then

$$\mathbb{E} \sup_{x \in B_{2,0}} \|BAx\|_2 \le 3K\sqrt{np + \log n}.$$

Proof. Let $c > 0$ be a constant to be determined later, and let $\lambda := cnp/\log(1/p)$. Then

$$B_{2,0} = \bigcup_{|J| \le \lambda n} B_2^J,$$

where the union is over all subsets $J \subset \{1, \dots, n\}$ of cardinality at most λn, and where $B_2^J = \{x \in \mathbb{R}^J : \ \|x\|_2 \le 1\}$ is the unit Euclidean ball in $\mathbb{R}^J$. By

Lemma 2.5, B_2^J has a $\frac{1}{2}$-net $\mathcal{N}_J$ of cardinality at most $e^{2\lambda n}$. Let $t \geq 1$. For a fixed $x \in \mathcal{N}_J$, Lemma 4.5 gives

$$\mathbb{P}\big\{\|BAx\|_2 > (K+1)\sqrt{np+\log n} + t\big\} \leq \exp\big(-c_0(np+t^2)\big).$$

Using Lemma 2.6 and taking the union bound over all $x \in \mathcal{N}_J$, we obtain

$$\begin{aligned}
&\mathbb{P}\big\{\sup_{x \in B_2^J} \|BAx\|_2 > (K+1)\sqrt{np+\log n} + t\big\} \\
&\leq \mathbb{P}\big\{\sup_{x \in \mathcal{N}} \|BAx\|_2 > 2(K+1)\sqrt{np+\log n} + t\big\} \\
&\leq |\mathcal{N}| \exp\big(-c_0(np+t^2)\big) \leq \exp\big(2\lambda n - c_0(np+t^2)\big).
\end{aligned}$$

Since there are $\binom{n}{\lceil \lambda n \rceil} \leq (e/\lambda)^{\lambda n}$ ways to choose the subset J, by taking the union bound over all J we conclude that

$$\begin{aligned}
(4.6)\quad &\mathbb{P}\big\{\sup_{x \in B_{2,0}} \|BAx\|_2 > 2(K+1)\sqrt{np+\log n} + t\big\} \\
&\qquad\qquad \leq \exp\big(\lambda \log(e/\lambda) + 2\lambda n - c_0(np+t^2)\big).
\end{aligned}$$

Finally, if the absolute constant $c > 0$ in the definition of λ is chosen sufficiently small, we have $\lambda \log(e/\lambda) + 2\lambda n \leq c_0 np$. Thus the right hand side of (4.6) is at most

$$\exp(-c_0 t^2).$$

Integration completes the proof. □

4.4. **Control of spread vectors.** Although we now have a good control of sparse vectors, they unfortunately comprise a small part of the unit ball B_2^n. More common but harder to deal with are "spread vectors" – those having many coordinates that are not close to zero. The next result gains control of the spread vectors.

Proposition 4.7 (Spread vectors)**.** *Let A and B be matrices as in Lemma 4.4. Let $M \geq 2$. Consider the set of vectors*

$$B_{2,\infty} := \Big\{x \in \mathbb{R}^n,\ \|x\|_2 \leq 1,\ \|x\|_\infty \leq \frac{M}{\sqrt{n}}\Big\}.$$

Then

$$\mathbb{E} \sup_{x \in B_{2,\infty}} \|BAx\|_2 \leq C \log^{3/2}(M) \cdot K\sqrt{np+\log N}.$$

Proof. This time we will need to work with multiple nets to account for different possible distributions of the magnitude of the coordinates of vectors $x \in B_{2,\infty}$. Since $\|x\|_\infty \leq \|x\|_2$, without loss of generality we can assume that $M \leq \sqrt{n}$.

Step 1: construction of nets. Let

$$h_k := \frac{2^k}{\sqrt{n}}, \qquad k = -2, -1, 0, 1, 2, \dots, \log_2 M$$

and let

$$\mathcal{N} := \{x \in B_{2,\infty} : \ \forall j \ \exists k \text{ such that } |x_j| = h_k\}.$$

A standard calculation shows that $\mathcal{N}$ is an $\frac{1}{2}$-net of $B_{2,\infty}$ in the $B_{2,\infty}$-norm, i.e. for every $x \in B_{2,\infty}$ there exists $y \in \mathcal{N}$ such that $x - y \in \frac{1}{2} B_{2,\infty}$. Therefore, by Lemma 2.6,

$$\sup_{x \in B_{2,\infty}} \|BAx\|_2 \le 2 \sup_{x \in \mathcal{N}} \|BAx\|_2.$$

Fix $x \in \mathcal{N}$. Since $\|x\|_2 \le 1$, the number of coordinates of x that satisfy $|x_j| = h_k$ is at most $\lfloor h_k^{-2} \rfloor$, for every k. Decomposing x according to the coordinates whose absolute value is h_k, we have by the triangle inequality that

$$\sup_{x \in B_{2,\infty}} \|BAx\|_2 \le 2 \sum_{k=-2}^{\log_2 M} \sup_{x \in \mathcal{N}_k} \|BAy\|_2, \tag{4.7}$$

where

$$\mathcal{N}_k = \big\{ y \in B_2^n : \ \|x\|_0 \le \lfloor h_k^{-2} \rfloor; \text{ all nonzero coordinates of } x \text{ satisfy } |x_j| = h_k \big\}.$$

Fix k. Since $h_k \le M/\sqrt{n}$, we have

$$m := \lfloor h_k^{-2} \rfloor \ge \lfloor n/M^2 \rfloor \ge 1. \tag{4.8}$$

To estimate the cardinality of $\mathcal{N}_k$, note that there are at most $\min(m, n)$ ways to choose $\|x\|_0 := l$; there are $\binom{n}{l}$ ways to choose the support of x; and there are 2^l ways to choose the (signs of) nonzero coordinates of x. Hence by Stirling's approximation and using (4.8), we have

$$|\mathcal{N}_k| \le \sum_{l=1}^{\min(m,n)} \binom{n}{l} 2^l \le \Big(\frac{2en}{m} \Big)^m \le (4eM^2)^m \le \exp(Cm \log M) \tag{4.9}$$

where $C \ge 1$ is an absolute constant.

Step 2: control of a fixed vector. Fix m and fix $x \in \mathcal{N}_k$. As we saw in the proof of Lemma 4.5,

$$\|BAx\|_2 \text{ is distributed identically with } \Big\| \sum_{i=1}^N g_i h_i B_i \Big\|_2$$

where

$$h_i = \Big(\sum_{j=1}^n a_{ij}^2 x_j^2 \Big)^{1/2}$$

and where g_i are independent standard normal random variables. Since $x \in \mathcal{N}_k$, we have $\|x\|_\infty = h_k \le \frac{1}{\sqrt{m}}$. This and the second condition in (4.4) yield

$$h_i \le \Big(\frac{1}{m}\sum_{j=1}^n a_{ij}^2\Big)^{1/2} \le K\sqrt{\frac{np+\log N}{m}}.$$

We consider the map $f : \mathbb{R}^N \to \mathbb{R}$ given by

$$f(y) = \Big\|\sum_{i=1}^N y_i h_i B_i\Big\|_2.$$

Repeating the estimate in the proof of Lemma 4.5, we bound the Lipschitz norm as

$$\|f\|_{\mathrm{Lip}} \le \max_i |h_i| \le K\sqrt{\frac{np+\log N}{m}}.$$

Then the Gaussian concentration, Theorem 2.1, gives for every $t > 0$:

$$\mathbb{P}(f(g) - \mathbb{E}f(g) > t) \le \exp\Big(-\frac{c_0 t^2 m}{K^2(np+\log N)}\Big),$$

where $g = (g_1, \dots, g_N)$. Since as we noted above, $f(g)$ is distributed identically with $\|BAx\|_2$, we can Lemma 4.4 yields that

$$\mathbb{P}(\|BAx\|_2 > K\sqrt{np+\log N} + t) \le \exp\Big(-\frac{c_0 t^2 m}{K^2(np+\log N)}\Big),$$

Let $u > 0$ be arbitrary. Applying the above estimate for $t = uK\sqrt{np+\log N}$, we conclude that

$$\mathbb{P}\big(\|BAx\|_2 > (1+u)K\sqrt{np+\log N}\big) \le \exp(-c_0 u^2 m). \tag{4.10}$$

Step 3: union bound. Taking the union bound in (4.10) over all $x \in \mathcal{N}_k$ and using estimate (4.9) on the cardinality of $\mathcal{N}_k$, we have for all $u > 0$:

$$\begin{aligned}\mathbb{P}\big(\sup_{x\in\mathcal{N}_k}\|BAx\|_2 > (1+u)K\sqrt{np+\log N}\big) &\le |\mathcal{N}_k|\exp(-c_0u^2m)\\ &\le \exp(Cm\log M - c_0u^2m).\end{aligned}$$

Let $s \ge 1$. We choose $u = C_1 s\sqrt{\log M}$, where $C_1 := \sqrt{C/c_0}$. Since $u \ge 1$ and $m \ge 1$, $M \ge 2$, we obtain from the above estimate that

$$\begin{aligned}\mathbb{P}\big(\sup_{x\in\mathcal{N}_k}\|BAx\|_2 > 2C_1sK\sqrt{\log(M)(np+\log N)}\big) &\le \exp(C(1-s^2)m\log M)\\ &\le \exp(c(1-s^2)).\end{aligned}$$

Integrating yields that

$$\mathbb{E}\sup_{x\in\mathcal{N}_k}\|BAx\|_2 \le C_2K\sqrt{\log(M)(np+\log N)}.$$

Putting this back in (4.7), we conclude that

$$\mathbb{E} \sup_{x \in B_{2,\infty}} \|BAx\|_2 \le 2(3 + \log M) \cdot C_2 K \sqrt{\log(M)(np + \log N)}.$$

This completes the proof. $\square$

4.5. **Norms of sparse matrices, and proof of Theorem 4.1.** Propositions 4.6 and 4.7 together handle all vectors in the unit ball, and yield the following norm estimate:

Proposition 4.8. *Let A and B be matrices as in Lemma 4.4. Then*

$$\mathbb{E}\|BA\| \le C \log^{2/3}\Big(\frac{1}{p}\Big) \cdot K\sqrt{np + \log N}.$$

Proof. Let c be the absolute constant as in Proposition 4.6. We can clearly assume that $c \le 1/4$. We define

$$M = \sqrt{\frac{1}{cp} \log \frac{e}{p}}.$$

Note that $M \ge 2$ as required in Proposition 4.6.

Fix a vector $x \in B_2^n$. We decompose it according to the magnitude of the coordinates, as follows:

$$x = y + z, \qquad y := x\,\mathbf{1}_{\{j:\, |x_j| > M/\sqrt{n}\}}, \qquad z := x\,\mathbf{1}_{\{j:\, |x_j| \le M/\sqrt{n}\}}.$$

Clearly, $\|y\|_2 \le \|x\| \le 1$, $\|z\|_2 \le \|x\| \le 1$. By Markov's inequality, we have

$$\|y\|_0 = \big|\{j:\ |x_j| > M/\sqrt{n}\}\big| \le \frac{n}{M^2} = \frac{cnp}{\log(e/p)}.$$

Then $y \in B_{2,0}$ as in Proposition 4.6. On the other hand, $\|z\|_\infty \le \frac{M}{\sqrt{n}}$ by definition, so $z \in B_{2,\infty}$ as in Proposition 4.7. Therefore, by Propositions 4.6 and 4.7we have

$$\begin{aligned}\mathbb{E}\|BA\| = \mathbb{E} \sup_{x \in B_2^n} \|BAx\|_2 &\le \mathbb{E} \sup_{y \in B_{2,0}} \|BAy\|_2 + \mathbb{E} \sup_{z \in B_{2,\infty}} \|BAz\|_2 \\ &\le 3K\sqrt{np + \log n} + C\log^{3/2}(M) \cdot K\sqrt{np + \log N}.\end{aligned}$$

Our choice of M completes the proof. $\square$

Finally, a standard symmetrization argument yields the following norm estimate, which we shall use for sparse random matrices.

Corollary 4.9. *Let $p \in (0,1]$. Let A be an $N \times n$ random matrix whose entries are independent random variables a_{ij} with mean zero and such that*

$$\mathbb{E}|a_{ij}|^2 \le p, \qquad |a_{ij}| \le 1 \qquad \text{for every } i, j.$$

Let B be an $n \times N$ matrix such that $\|B\| \le 1$. Then

$$\mathbb{E}\|BA\| \le C \log^{3/2}\left(\frac{e}{p}\right)\sqrt{np + \log N}.$$

Remark. It would be interesting to remove the logarithmic term from this estimate.

Proof. Let g_{ij} be independent standard normal random variables. Consider the random matrix $\tilde{A} = (g_{ij}a_{ij})$. By (2.1), we have

$$\mathbb{E}\|BA\| \le (\pi/2)^{1/2}\,\mathbb{E}\|B\tilde{A}\|. \tag{4.11}$$

By Lemma 4.3, conditions (4.4) hold with a random parameter K which only depends on the random variables (a_{ij}) and not on (g_{ij}), and which satisfies

$$\mathbb{E}_a K \le C_1 \tag{4.12}$$

where C_1 is an absolute constant. Here and below we write $\mathbb{E}_a$ when the expectation is with respect to (a_{ij}), and $\mathbb{E}_g$ if the expectation is with respect to (g_{ij}).

Condition on random variables (a_{ij}). Proposition 4.8 then yields

$$\mathbb{E}_g\|B\tilde{A}\| \le C \log^{2/3}\left(\frac{1}{p}\right)\cdot K\sqrt{np + \log N}.$$

Therefore, when we remove the conditioning, we obtain by (4.12) that

$$\mathbb{E}\|B\tilde{A}\| = \mathbb{E}_a\mathbb{E}_g\|B\tilde{A}\| \le C \log^{2/3}\left(\frac{1}{p}\right)\cdot C_1\sqrt{np + \log N}.$$

This and (4.11) complete the proof. □

Proof of Theorem 4.1. By the standard symmetrization technique described in Section 2, we can assume without loss of generality that all a_{ij} are symmetric random variables. We decompose the matrix A according to the magnitude of its entries as follows. Given a subset $I \subset \mathbb{R}$, we define the truncated matrix

$$\operatorname{trunc}(A, I) = (a_{ij}\mathbf{1}_{\{|a_{ij}| \in I\}}).$$

Consider

$$A^{(0)} = \operatorname{trunc}(A, [0,1]);$$
$$A^{(k)} = 2^{-k}\operatorname{trunc}(A, (2^{k-1}, 2^k]), \quad k = 1, 2, \ldots$$

Then we have a decomposition $A = \sum_{k=0}^{\infty} 2^k A^{(k)}$. This sum is actually finite because of the assumptions on a_{ij}. Indeed, we have

$$A = A^{(0)} + \sum_{k=1}^{k_0} 2^k A^{(k)} \tag{4.13}$$

where k_0 is the maximal integer such that

$$2^{k_0} \le \Big(\frac{Mn}{\log N} \Big)^{\frac{1}{2+\varepsilon}}. \tag{4.14}$$

Because a_{ij} are symmetric random variables, all entries $a_{ij}^{(k)}$ of the matrices $A^{(k)}$ satisfy $\mathbb{E} a_{ij}^{(k)} = 0$ and $|a_{ij}^{(k)}| \le 1$.

Using Corollary 4.9 for the matrix $A^{(0)}$ and $p = 1$, we obtain

$$\mathbb{E}\|BA^{(0)}\| \le C_1 \sqrt{np + \log N} \le 2C_1 \sqrt{Mn}, \tag{4.15}$$

where the last line follows because $\log N \le Mn$ and $M \ge 1$ by the hypothesis.

Now we fix $1 \le k \le k_0$. Using the $(2+\varepsilon)$-th moment assumption, we have by Markov's inequality that

$$\mathbb{P}(a_{ij}^{(k)} \ne 0) \le \mathbb{P}(a_{ij} > 2^{k-1}) \le 2^{-(2+\varepsilon)(k-1)} =: p_k.$$

This and the bound $|a_{ij}^{(k)}| \le 1$ yield $\mathbb{E}(a_{ij}^{(k)})^2 \le p_k$. With this, we apply Corollary 4.9 for the matrix $A^{(k)}$ and obtain

$$\mathbb{E}\|BA^{(k)}\| \le C \log^{3/2} \Big(\frac{e}{p_k} \Big) \sqrt{np_k + \log N}.$$

By the definition of p_k and by (4.14), we have

$$p_k \ge p_{k_0} \ge \frac{\log N}{Mn}.$$

Therefore, $np_k + \log N \le (1+M)np_k \le 2Mnp_k$, so

$$\begin{aligned} \mathbb{E}\|BA^{(k)}\| &\le C \log^{3/2} \Big(\frac{e}{p_k} \Big) \sqrt{2Mnp_k} \\ &\le C_2 \big[1 + (2+\varepsilon)(k-1) \big]^{3/2} 2^{-(1+\varepsilon/2)(k-1)} \cdot \sqrt{Mn}. \end{aligned} \tag{4.16}$$

Using (4.13) and the triangle inequality, then using (4.15) and (4.16), we conclude that

$$\begin{aligned} \mathbb{E}\|BA\| &\le \mathbb{E}\|BA^{(0)}\| + \sum_{k=1}^{k_0} 2^k \, \mathbb{E}\|BA^{(k)}\| \\ &\le 2C_1\sqrt{Mn} + \sum_{k=1}^{k_0} C_2 \big[1 + (2+\varepsilon)(k-1) \big]^{3/2} 2^{k-(1+\varepsilon/2)(k-1)} \cdot \sqrt{Mn} \\ &\le C_2 \sqrt{Mn} \cdot \sum_{k=1}^{\infty} k^{3/2} 2^{-(\varepsilon/2)k} \\ &= C(\varepsilon)\sqrt{Mn}. \end{aligned}$$

This completes the proof of Theorem 4.1. ∎

4.6. **Almost square matrices.** The main application of Theorem 4.1 is for almost square matrices – those for which $N = n^{1+o(1)}$. The next lemma verifies the hypotheses of Theorem 4.1 for such matrices.

Lemma 4.10. *Let $\varepsilon \in (0,1)$, and assume that $N \le n^{1+\varepsilon/10}$. Let A be an $N \times n$ random matrix whose entries are inedpendent random variables with $(4+\varepsilon)$-th moment bounded by 1. Define the random variable M by the equation*

$$\max_{i,j} |a_{ij}| = \Big(\frac{Mn}{\log N}\Big)^{\frac{1}{2+\varepsilon/4}}. \tag{4.17}$$

Then, for every $t \ge 1$, one has

$$\mathbb{P}\big(M > C(\varepsilon)t\big) \le \frac{1}{t^2}.$$

In particular, one has $\mathbb{E}M \le C_1(\varepsilon)$.

Proof. By Markov inequality, we have for every i,j that

$$\mathbb{P}\big(|a_{ij}| > s\big) \le \frac{1}{s^{4+\varepsilon}}, \qquad s > 0.$$

Let $t \ge 1$. We then have

$$\mathbb{P}\big(|a_{ij}| > (t^2 nN)^{\frac{1}{4+\varepsilon}}\big) \le \frac{1}{t^2 nN}.$$

Taking the union bound over all nN random variables a_{ij}, we obtain

$$\mathbb{P}\big(\max_{i,j} |a_{ij}| > (t^2 nN)^{\frac{1}{4+\varepsilon}}\big) \le \frac{1}{t^2}. \tag{4.18}$$

The assumption $N \le n^{1+\varepsilon/10}$ yields that

$$nN \le \Big(\frac{C(\varepsilon)n}{\log N}\Big)^{2+\varepsilon/8}.$$

Therefore, since $\frac{2+\varepsilon/8}{4+\varepsilon} \le \frac{1}{2+\varepsilon/4}$ and $t \ge 1$, we have

$$(t^2 nN)^{\frac{1}{4+\varepsilon}} \le \Big(\frac{C(\varepsilon)tn}{\log N}\Big)^{\frac{1}{2+\varepsilon/4}}.$$

Using this in (4.18), we obtain

$$\mathbb{P}\big(M > C(\varepsilon)t\big) \le \mathbb{P}\Big(\max_{i,j} |a_{ij}| > \Big(\frac{C(\varepsilon)tn}{\log N}\Big)^{\frac{1}{2+\varepsilon/4}}\Big) \le \frac{1}{t^2}.$$

Integration completes the proof. □

We are now ready to state and prove a partial case of Theorem 1.1 for almost square matrices.

Corollary 4.11. *Let $\varepsilon \in (0,1)$, and assume that $N \le n^{1+\varepsilon/10}$. Let A be an $N \times n$ random matrix whose entries are independent random variables with mean zero and $(4+\varepsilon)$-th moment bounded by 1. Let B be an $n \times N$ matrix such that $\|B\| \le 1$. Then*

$$\mathbb{E}\|BA\| \le C(\varepsilon)\sqrt{n}.$$

Proof. By the standard symmetrization, we can assume that the random variables a_{ij} are symmetric. Let M be the random variable as in Lemma 4.10, and let $t \ge 1$. By the definition, $\{M \le t\}$ is the product event. Therefore, conditioning on this event (i) preserves the independence of the entries of A; (ii) makes all these entries bounded as in (4.17); (iii) can only reduce their moments by Lemma 2.8, thus for all i, j we have

$$\mathbb{E}\big(|a_{ij}|^{2+\varepsilon/4} \,|\, M \le t\big) \le \mathbb{E}|a_{ij}|^{2+\varepsilon/4} \le 1.$$

Therefore, we can apply Corollary 4.2 conditionally and with $\varepsilon/4$, which gives

$$\big[\mathbb{E}\big(\|BA\|^2 \,|\, M \le t\big)\big]^{1/2} \le C_0(\varepsilon/4)\sqrt{tn}.$$

Additionally, by Lemma 4.10 we have

$$\mathbb{P}\big(M > C(\varepsilon)t\big) \le \frac{1}{t^2}.$$

By Lemma 2.9, this yields

$$\mathbb{E}\|BA\| \le C_1(\varepsilon)\sqrt{n}$$

as claimed. □

5. Completion of the proof of Theorem 1.1

Proof of Theorem 1.1. As we mentioned in Section 2.6, we can assume that $N \ge n = m$. Fix a value of ε, and let

$$K = K(\varepsilon) = \frac{1}{2} + \frac{1}{\varepsilon}.$$

As usual, let $B_1, \ldots, B_N$ be the columns of the matrix B. Consider the subset $I \subset \{1, \ldots, \}$ of large columns defined as

$$I := \Big\{i : \ \|B_i\|_2 > \frac{1}{\log^K n}\Big\}.$$

By (2.3), we have

$$|I| < n \log^{2K} n =: N_0.$$

Denote by A_I the $N_0 \times n$ submatrix of A whose rows are in I, by B_I the $n \times N_0$ submatrix of B whose columns are in I (and similarly for I^c). The decomposition $BA = B_I A_I + B_{I^c} A_{I^c}$ implies by the triangle inequality that

$$\|BA\| \le \|B_I A_I\| + \|B_{I^c} A_{I^c}\|. \tag{5.1}$$

This splits our problem into two subproblems, one for I and one for I^c.

The matrices A_I, B_I are almost square, so Corollary 4.11 applies for them, which yields

$$\mathbb{E}\|B_I A_I\| \le C(\varepsilon)\sqrt{n}. \tag{5.2}$$

On the other hand, the columns of the matrix B_{I^c} are small by the definition of I:

$$\|B_i\|_2 \le \frac{1}{\log^K n} \qquad \text{for every} i \in I^c.$$

Therefore, Theorem 3.9 applies to the matrices A_{I^c}, B_{I^c}, which gives

$$\mathbb{E}\|B_{I^c} A_{I^c}\| \le C\sqrt{n}. \tag{5.3}$$

Putting estimates (5.2) and (5.3) into (5.1), we conclude that

$$\|BA\| \le C_1(\varepsilon)\sqrt{n}.$$

Theorem 1.1 is proved. □

References

[1] G. Aubrun, *Sampling convex bodies: a random matrix approach*, Proc. Amer. Math. Soc. 135 (2007), 1293–1303

[2] Z. D. Bai, J. W. Silverstein, *No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices*, Ann. Probab. 26 (1998), 316–345

[3] Z. D. Bai, J. W. Silverstein, *Exact separation of eigenvalues of large-dimensional sample covariance matrices*, Ann. Probab. 27 (1999), 1536–1555

[4] Z. D. Bai, J. W. Silverstein, Y. Q. Yin, *A note on the largest eigenvalue of a large-dimensional sample covariance matrix*, J. Multivariate Anal. 26 (1988), 166–168

[5] Z. D. Bai, Y. Q. Yin, *Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix*, Ann. Probab. 21 (1993), 1275–1294

[6] S. Boucheron, O. Bousquet, and G. Lugosi, *Concentration inequalities.* In: O. Bousquet, U.V. Luxburg, and G. Rätsch (editors), Advanced Lectures in Machine Learning, Springer, pp. 208–240, 2004

[7] Z. Füredi, J. Komlós, *The eigenvalues of random symmetric matrices*, Combinatorica 1 (1981), 233–241

[8] A. Khorunzhy, *Sparse random matrices: spectral edge and statistics of rooted trees*, Adv. in Appl. Probab. 33 (2001), 124–140

[9] R. Latala, *Some estimates of norms of random matrices*, Proc. Amer. Math. Soc. 133 (2005), 1273–1282

[10] M. Ledoux, *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001

[11] M. Ledoux and M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23 Springer-Verlag, Berlin, 1991.

[12] Y. Seginer, *The expected norm of random matrices*, Combin. Probab. Comput. 9 (2000), 149–166

[13] M. Talagrand, *A new look at independence*, Ann. Probab. 24 (1996), 1–34

[14] M. Rudelson, *Random vectors in the isotropic position*, J. Funct. Anal. 164 (1999), 60–72

[15] M. Rudelson, R. Vershynin, *Sampling from large matrices: an approach through geometric functional analysis*, Journal of the ACM (2007), Art. 21, 19 pp.

[16] M. Rudelson, R. Vershynin, *The Littlewood-Offord Problem and invertibility of random matrices*, Advances in Mathematics 218 (2008), 600–633

[17] M. Rudelson, R. Vershynin, *The smallest singular value of a random rectangular matrix*, submitted

[18] Y. Q. Yin, Z. D. Bai, P. R. Krishnaiah, *On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix*, Probab. Theory Related Fields 78 (1988), 509–521

Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

E-mail address: romanv@umich.edu