Bernoulli **26**(4), 2020, 3139–3162
https://doi.org/10.3150/20-BEJ1218

Concentration inequalities for random tensors

ROMAN VERSHYNIN

Department of Mathematics, University of California, Irvine, Irvine, CA 92617, U.S.A.
E-mail: rvershyn@uci.edu; url: https://www.math.uci.edu/~rvershyn/index.html

We show how to extend several basic concentration inequalities for simple random tensors $X = x_1 \otimes \cdots \otimes x_d$ where all x_k are independent random vectors in $\mathbb{R}^n$ with independent coefficients. The new results have optimal dependence on the dimension n and the degree d. As an application, we show that random tensors are well conditioned: $(1 - o(1))n^d$ independent copies of the simple random tensor $X \in \mathbb{R}^{n^d}$ are far from being linearly dependent with high probability. We prove this fact for any degree $d = o(\sqrt{n/\log n})$ and conjecture that it is true for any $d = O(n)$.

Keywords: concentration inequalities; condition numbers; polynomials; random tensors

1. Introduction

Concentration inequalities form a powerful toolset in probability theory and its many applications; see, for example, [12,18,19]. Perhaps the best known member of this large family of results is the *Gaussian concentration inequality*, which states that a standard normal random vector x in $\mathbb{R}^n$ satisfies

$$\mathbb{P}\big\{\big|f(x) - \mathbb{E}\, f(x)\big| > t\big\} \le 2\exp\bigg(-\frac{t^2}{2\|f\|_{\mathsf{Lip}}^2}\bigg) \tag{1.1}$$

for any Lipschitz function $f : (\mathbb{R}^n, \|\cdot\|_2) \to \mathbb{R}$, see e.g. [12], Theorem 5.6.

The Gaussian concentration inequality can be extended to some more general distributions on $\mathbb{R}^n$. A remarkable situation where this is possible is where x has a product distribution with *bounded* coordinates and f is *convex*. This result is due to by M. Talagrand [28]; see [18], Section 4.2, [12], Section 7.5:

Theorem 1.1 (Convex concentration). *Let $f : (\mathbb{R}^n, \|\cdot\|_2) \to \mathbb{R}$ be a convex and Lipschitz function. Let x be a random vector in $\mathbb{R}^n$ whose coordinates are independent random variables that are bounded a.s. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\big\{\big|f(x) - \mathbb{E}\, f(x)\big| > t\big\} \le 2\exp\bigg(-\frac{ct^2}{\|f\|_{\mathsf{Lip}}^2}\bigg). \tag{1.2}$$

Here $c > 0$ depends only on the bound on the coordinates.

The boundedness assumption in this result unfortunately excludes Gaussian distributions and many others. Significant efforts were made to extend Gaussian concentration to more general, not necessarily bounded, distributions, see, for example, [2,6,15] and the references therein. One

such result, which holds for a general random vector x with independent *subgaussian* coordinates, is for *Euclidean functions*, that is, the functions of the form $f(x) = \|Ax\|_H$ where A is a linear operator from $\mathbb{R}^n$ into a Hilbert space.

Theorem 1.2 (Euclidean concentration). *Let H be a Hilbert space and $A : (\mathbb{R}^n, \|\cdot\|_2) \to H$ be a linear operator. Let x be a random vector in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\big\{\big|\|Ax\|_H - \|A\|_{\mathsf{HS}}\big| \geq t\big\} \leq 2\exp\bigg(-\frac{ct^2}{\|A\|_{\mathsf{op}}^2}\bigg). \tag{1.3}$$

Here $c > 0$ depends only on the bound on the sub-Gaussian norms.

In this result, $\|A\|_{\mathsf{HS}}$ and $\|A\|_{\mathsf{op}}$ denote the Hilbert–Schmidt and operator norms of A, respectively. Theorem 1.2 can be derived from Hanson–Wright concentration inequality for quadratic forms [27], see [35], Section 6.3.

1.1. New results

The goal of this paper is to demonstrate how Theorems 1.1 and 1.2 can be extended for *simple random tensors*. These are tensors of the form

$$X := x_1 \otimes \cdots \otimes x_d$$

where x_k are independent random vectors in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance random variables that are either bounded a.s. (in Theorem 1.1) or sub-Gaussian (in Theorem 1.2).

Can we expect that a concentration inequality like (1.2) or (1.3) can hold for simple random tensors, that is, for $f(X)$ where $f : (\mathbb{R}^{n^d}, \|\cdot\|_2) \to \mathbb{R}$? Not really: if such inequality did hold, then it would imply that $\operatorname{Var}(f(X)) = O(1)$, but this is not the case. Indeed, consider the simplest case where f is given by the Euclidean norm, that is, $f(X) := \|X\|_2$, and let all x_k be standard normal random vectors in $\mathbb{R}^n$. Recall that a standard normal vector x in $\mathbb{R}^n$ satisfies

$$\mathbb{E}\,\|x\|_2^2 = n \quad \text{and} \quad \mathbb{E}\,\|x\|_2 \leq \sqrt{n - c}$$

for n large enough, where $c > 0$ is an absolute constant.[1] Then

$$\begin{aligned}\operatorname{Var}\big(f(X)\big) &= \mathbb{E}\,\|X\|_2^2 - \big(\mathbb{E}\,\|X\|_2\big)^2 = \big(\mathbb{E}\,\|x\|_2^2\big)^d - \big(\mathbb{E}\,\|x\|_2\big)^{2d} \\ &\geq n^d - (n-c)^d \geq cd(n-c)^{d-1} \quad \text{(by the binomial expansion)} \\ &\asymp dn^{d-1}.\end{aligned}$$

[1] To check the second bound, write $\|x\|_2^2 = n + \sum_{i=1}^n (x_i^2 - 1)$ and observe that, by the central limit theorem, the sum is approximately $\sqrt{2n}g$ where $g \sim N(0,1)$. Thus, $\|x\|_2 \approx \sqrt{n} + g/\sqrt{2}$, so $\operatorname{Var}(\|x\|_2) \gtrsim c$. On the other hand, $\operatorname{Var}(\|x\|_2) = \mathbb{E}\,\|x\|^2 - (\mathbb{E}\,\|x\|)^2 = n - (\mathbb{E}\,\|x\|)^2$. Thus, $\mathbb{E}\,\|x\| \leq \sqrt{n-c}$.

Thus, the strongest concentration inequality we can hope for must have the form

$$\mathbb{P}\big\{\big|f(X)-\mathbb{E}\,f(X)\big|>t\big\}\le 2\exp\bigg(-\frac{ct^2}{dn^{d-1}\|f\|_{\mathsf{Lip}}^2}\bigg). \tag{1.4}$$

Such inequality, however, can not be true for large t. The coefficients of X are d-fold products of sub-Gaussian random variables, and such products for $d\ge 2$ typically have tails that are heavier than sub-Gaussian. Nevertheless, we may still hope that the inequality (1.4) might hold for all t in some interesting range, for example for $0\le t\le |\mathbb{E}\,f(X)|$. This is what we prove in the current paper.

Theorem 1.3 (Convex concentration for random tensors). *Let n and d be positive integers and $f:(\mathbb{R}^{n^d},\|\cdot\|_2)\to\mathbb{R}$ be a convex and Lipschitz function. Consider a simple random tensor $X:=x_1\otimes\cdots\otimes x_d$ in $\mathbb{R}^{n^d}$, where all x_k are independent random vectors in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance random variables that are bounded a.s. Then, for every $0\le t\le 2(\mathbb{E}\,|f(X)|^2)^{1/2}$, we have*

$$\mathbb{P}\big\{\big|f(X)-\mathbb{E}\,f(X)\big|>t\big\}\le 2\exp\bigg(-\frac{ct^2}{dn^{d-1}\|f\|_{\mathsf{Lip}}^2}\bigg).$$

Here $c>0$ depends only on the bound on the coordinates.

Theorem 1.4 (Euclidean concentration for random tensors). *Let n and d be positive integers, H be a Hilbert space and $A:(\mathbb{R}^{n^d},\|\cdot\|_2)\to H$ be a linear operator. Consider a simple random tensor $X:=x_1\otimes\cdots\otimes x_d$ in $\mathbb{R}^{n^d}$, where all x_k are independent random vectors in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then, for every $0\le t\le 2\|A\|_{\mathsf{HS}}$, we have*

$$\mathbb{P}\big\{\big|\|AX\|_H-\|A\|_{\mathsf{HS}}\big|\ge t\big\}\le 2\exp\bigg(-\frac{ct^2}{dn^{d-1}\|A\|_{\mathsf{op}}^2}\bigg).$$

Here $c>0$ depends only on the bound on the sub-Gaussian norms.

Remark 1.5 (The range of concentration inequalities). Our argument shows that the main results actually hold in a somewhat wider range of t, namely for $0\le t\le 2n^{d/2}\|f\|_{\mathsf{Lip}}$ in Theorem 1.3 and $0\le t\le 2n^{d/2}\|A\|_{\mathsf{op}}$ in Theorem 1.4.

1.2. Related results

Several existing techniques are already known to shed light on tensor concentration. Indeed, the quantity $\|AX\|_H^2$ in Theorem 1.4 can be expressed a polynomial of degree $2d$ in nd independent sub-Gaussian random variables, which are the coefficients of the random vectors x_i. A remarkable result of Latala [17] provides two-sided bounds on the moments of polynomials

of independent normal random variables, or *Gaussian chaoses*. These moment bounds can be translated into the following concentration inequality that is valid for any fixed tensor $B \in \mathbb{R}^{n^d}$ and a simple random tensor $X := x_1 \otimes \cdots \otimes x_d$ composed of independent, mean zero, standard normal random vectors x_k in $\mathbb{R}^n$:

$$\mathbb{P}\big\{\big|\langle B, X\rangle\big| \geq t\big\} \leq C_d \exp\bigg(-c_d \cdot \min_{1\leq k\leq d} \min_{I_1 \sqcup \cdots \sqcup I_k = [d]} \bigg(\frac{t}{\|B\|_{I_1,\ldots,I_k}}\bigg)^{2/k}\bigg). \tag{1.5}$$

The second minimum in the right-hand side is over all partitions of the index set $[d] = \{1, \ldots, d\}$ into k subsets, and $\|B\|_{I_1,\ldots,I_k}$ are certain norms of B, which interpolate between the Hilbert–Schmidt and the operator norms; see [17] for details.

Lehec [20] gave an alternative, shorter proof of (1.5) based on Talagrand's majorizing measure theorem. Adamczak and Latala [4] extended (1.5) for log-concave distributions. Adamczak and Wolff [7] considered distributions that satisfy a Sobolev inequality, and he gave an extension of (1.5) for general differentiable functions of X (not just polynomials). Götze, Sambale, and Sinulis extended (1.5) for α-subexponential distributions. Very recently, Adamczak, Latala, and Meller [5] explored extensions of (1.5) where the coefficients of B are not scalars but vectors in some Banach space. We refer the paper to the papers [5,7] for a review of the vast literature on concentration inequalities for multivariate polynomials, U-statistics, and more general functions of independent random variables.

As we noted, inequality (1.5) is well suited to study concentration properties of the quantity $\|AX\|_H^2$ that appears in Theorem 1.4. However, Latala's concentration inequality (1.5), as well as all its known extensions and ramifications, contain factors C_d and c_d that depend on the degree d of tensors in some unspecified way; the dependence seems to be at least exponential in d. In contrast, Theorems 1.3 and 1.4 feature the optimal dependence on the degree d of tensors: the constant c in both theorems does not depend on d at all. This can be critical in some applications, one of which we discuss next.

1.3. Application: Random tensors are well conditioned

Our work was primarily inspired by a question that arose recently in the theoretical computer science community [1,8,9,11]: *Are random tensors well conditioned?*

Suppose $X_1, \ldots, X_m$ are independent copies of a simple random tensor $X := x_1 \otimes \cdots \otimes x_d$. How large can m be so that these random tensors are linearly independent with high probability? Certainly m can not exceed the dimension n^d of the tensor space, but can it be arbitrarily close to the dimension, say $m = 0.99n^d$? Moreover, instead of linear independence we may ask for a stronger, more quantitative property of being well conditioned. We would like to have a uniform bound

$$\bigg\|\sum_{i=1}^m a_i X_i\bigg\|_2 \geq \sigma \|a\|_2 \quad \text{for all } a = (a_1, \ldots, a_m) \in \mathbb{R}^m \tag{1.6}$$

with σ as large as possible. Equivalently, we can understand (1.6) as a lower bound on the *smallest singular value* of the $n^d \times m$ matrix $\mathbf{X}^{\odot d}$ (called the *Khatri–Sidak product*) whose

columns are vectorized tensors $X_1, \dots, X_m$:

$$\sigma_{\min}\big(\mathbf{X}^{\odot d}\big) \geq \sigma.$$

Problems of this type were studied recently in the theoretical computer science community in the context of computing tensor decompositions [8,11], learning Gaussian mixtures [9] and estimating the capacity of error correcting codes [1].

For $d = 1$, this problem has been extensively studied in random matrix theory [3,10,13,16,22, 24–26,30–34]. Since $\mathbf{X}^{\odot 1}$ is an $n \times m$ random matrix whose entries are independent, mean zero, unit variance, sub-gaussian random variables, known optimal results [25] yield the bound

$$\sigma_{\min}\big(\mathbf{X}^{\odot 1}\big) \gtrsim \varepsilon\sqrt{n} \tag{1.7}$$

if $m = (1 - \varepsilon)n$ and $\varepsilon \in (0, 1]$.

For $d \geq 2$, optimal results on $\sigma = \sigma_{\min}(\mathbf{X}^{\odot d})$ are yet unknown. Various random models were studied: the factors x_k of the simple tensor X were assumed to be Gaussian (possibly with nonzero means) in [11], Bernoulli in [1] and are allowed to have very general general distributions with non-degenerate marginals in [8]. Symmetric random tensors are considered in [1]. Baskara et al. [11] obtained the lower bound (1.6) with $\sigma = (1/n)^{\exp(O(d))}$ for $d = o(\log\log n)$; Anari et al. [8] improves this to $\sigma = (1/O(n))^d$ for $d = o(\sqrt{n/\log n})$, and Abbe et al. [1] guarantees linear independence (i.e. $\sigma > 0$) for symmetric random tensors if $d = o(\sqrt{n/\log n})$. For a related notion of *row product* of random matrices, the problem was studied by Rudelson [23].

In this paper, we prove a bound on the smallest singular value $\sigma_{\min}(\mathbf{X}^{\odot d})$ is of constant order. We derive it as an application of Theorem 1.4. Let us give an informal statement here; Corollary 6.2 will provide a more rigorous version.

Corollary 1.6 (Random tensors are well conditioned). *If $d = o(\sqrt{n/\log n})$ and $\varepsilon \in (0, 1)$, then $m = (1 - \varepsilon)n^d$ independent simple sub-Gaussian random tensors $X_1, \dots, X_m$ in $\mathbb{R}^{n^d}$ are well conditioned with high probability*:

$$\left\| \sum_{i=1}^{m} a_i X_i \right\|_2 \geq \frac{\sqrt{\varepsilon}}{2} \|a\|_2 \quad \text{for all } a = (a_1, \dots, a_m) \in \mathbb{R}^m.$$

1.4. Our approach

Let us briefly explain our approach to tensor concentration. Suppose first that we do not care about the dependence on the degree d. Then Theorem 1.3 can be proved by expressing the deviation $f - \mathbb{E} f$ as a telescopic sum and controlling each increment by Talagrand's Theorem 1.1. For example, if $d = 3$, then for the function $f = f(x \otimes y \otimes z)$ we would write

$$f - \mathbb{E} f = (f - \mathbb{E}_x f) + (\mathbb{E}_x f - \mathbb{E}_{x,y} f) + (\mathbb{E}_{x,y} f - \mathbb{E}_{x,y,z} f) =: \Delta_1 + \Delta_2 + \Delta_3$$

where $\mathbb{E}_x$ denotes the conditional expectation with respect to x (conditional on y and z), and similarly for $\mathbb{E}_{x,y}$. Applying Theorem 1.1 for f as a function of x, we control Δ_1; applying the

same theorem for $\mathbb{E}_x f$ as a function of y, we control Δ_2, and applying it again for $\mathbb{E}_{x,y} f$ as a function of z, we control Δ_3. Then we combine all increments Δ_k by the triangle inequality.

This argument, however, would produce an exponential dependence of d in the concentration inequality. This is because the Lipschitz norm of f as a function of x is bounded by

$$\|y\|_2 \|z\|_2 \le K^{d/2}$$

if all coefficients of y and z are bounded by K a.s.

To get a better control of the Lipschitz norms of all functions that appear. in the telescopic sum, we prove a *maximal inequality* in Section 3, which provides us with a uniform bound on the products of norms of independent random vectors. This allows us to avoid losing any factors that are exponential in d.

However, combining the increments Δ_k by a simple union bound and triangle inequality is suboptimal and leads to an extra factor that is linear in d. One can avoid this by noting that Δ_k are martingale differences and using martingale concentration techniques (coupled with the maximal inequality). This is the approach we chose to prove Theorem 1.3.

One can try to prove Theorem 1.4 in a similar way, but a new difficulty arises here. We may not simply choose $f(x \otimes y \otimes z) = \|A(x \otimes y \otimes z)\|_2$ and write the telescopic sum for it. This is because $\mathbb{E}_x f$ is not a Euclidean function in y, it can not be expressed as $\|By\|$ for any linear operator B mapping $\mathbb{R}^n$ into a Hilbert space, so we may not use Theorem 1.2 to control the deviation of $\mathbb{E}_x f$.

This forces us to work with f^2 instead of f, since then $(\mathbb{E}_x f^2)^{1/2}$ is a Euclidean function. Thus, we write

$$f^2 - \mathbb{E} f^2 = \big(f^2 - \mathbb{E}_x f^2\big) + \big(\mathbb{E}_x f^2 - \mathbb{E}_{x,y} f^2\big) + \big(\mathbb{E}_{x,y} f^2 - \mathbb{E}_{x,y,z} f^2\big) =: \Delta_1 + \Delta_2 + \Delta_3.$$

Squaring f, however, produces tails of the increments that are heavier than sub-Gaussian. This prompts us to abandon the use of Theorem 1.2. Instead of controlling the tails of the increments, we control their moment generating function (MGF). In the end, we still combine the MGF's of the increments using a martingale-like argument coupled with the maximal inequality. This ultimately leads to Theorem 1.4.

Remark 1.7 (An alternative approach to convex concentration for random tensors). There is an alternative and somewhat simpler way to prove Theorem 1.3, where one won't have to use a maximal inequality. Instead, one can deduce this result (with some work) from a version of convex concentration (Theorem 1.1) that holds for a weaker notion of convexity, namely for separately convex functions, that is, functions that are convex in each coordinate [21,28], see [12], Theorems 6.10, 6.9. However, there seem to be no simpler way to prove Theorem 1.4, a result we care most about in view of applications. So, for pedagogical reasons we choose to prove Theorem 1.3 using maximal inequality, so we can use it later as a stepping stone for the proof of Theorem 1.4.

Remark 1.8 (A broader view). The method we develop here is flexible and might be used to "tensorize" some other concentration inequalities. For example, if all x_k have the standard normal distribution, then the convexity requirement is not needed in Theorem 1.3, and we get a tensor

version of the Gaussian concentration inequality (1.1). Furthermore, one should be able to relax the sub-Gaussian assumption in Theorem 1.4 by using in our argument a version Theorem 1.2 for heavier tails obtained recently by Götze, Sambale, and Sinulis [14].

1.5. Open problems

We do not know optimal concentration inequalities for *symmetric* tensors $X = x^{\otimes d} = x \otimes \cdots \otimes x$. One could possibly use decoupling to reduce the problem to concentration to tensors with independent factors, and then apply Theorem 1.3 or 1.4. However, decoupling will likely cause a loss of factors that are exponential in d, which will defeat our purpose.

There are many directions in which Corollary 1.6 should be strengthened and generalized. Can the bound be improved to $\varepsilon n^{d/2}$, matching the inequality (1.7) for $d = 1$? Does it hold for degrees higher than $\sqrt{n}$, for example for $d \asymp n$? Even linear independence is unknown for higher degrees. Can Corollary 1.6 be extended to other models of randomness considered in the theoretical computer science community [1,8,9,11]? For example, does it hold for symmetric tensors, and can the mean zero and sub-Gaussian assumptions be significantly weakened?

1.6. The rest of the paper

In Section 2, we collect some basic facts from high-dimensional probability that will be needed later. Most importantly, in Proposition 2.2 we show how to control the MGF of a random chaos of order 2, which is the quadratic form $x^{\mathsf{T}} M x$ where x is a random vector and M is a fixed matrix. In Proposition 2.2, we derive a version of Hanson–Wright inequality in the MGF form. These results, although possibly known as a folklore, are hard to find in the literature and could be useful for future applications.

In Section 3, we prove a sharp *maximal inequality* for products of norms of independent random vectors. We use it to establish our main results: in Section 4 we prove Theorem 1.3 and in Section 5 we prove Theorem 1.4.

In Section 6, we give two applications to the geometry of random tensors. We prove a concentration inequality for the distance between a random tensor and a given subspace in Corollary 6.1, and then use it to show that random tensors are well conditioned, proving a formal version of Corollary 1.6.

2. Preliminaries

Throughout this paper, we use basic facts about subgaussian and subexponential random variables that can be found for example, in [35], Chapters 2–3, and [36], Chapter 2. Positive constants are denoted by $C, c, C_1, c_1, \ldots$, and their specific values can be different in different parts of this paper. We allow these constants to depend only on the a.s. bound of the coefficients (for Theorem 1.1) or the sub-Gaussian norms of the coefficients (for Theorem 1.2), but not on any other parameters.

2.1. Concentration of the norm

One such fact that follows immediately from Theorem 1.2 for the identity matrix A is a concentration inequality for the norm of a random vector.

Corollary 2.1 (Concentration of norm). *Let x be a random vector in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\big\{\big|\|x\|_2 - \sqrt{n}\big| > t\big\} \leq 2\exp\big(-ct^2\big).$$

Here $c > 0$ depends only on the bound on the sub-Gaussian norms.

2.2. MGF of a quadratic form

Another useful fact is the following bound on the moment generating function (MGF) of a chaos of order 2. It might be known but is hard to find in the literature in this generality.

Proposition 2.2 (MGF of a sub-Gaussian chaos). *Let M be an $n \times n$ matrix. Let x be a random vector in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then*

$$\mathbb{E}\exp\big(\lambda\big(x^{\mathsf{T}}Mx - \operatorname{tr} M\big)\big) \leq \exp\big(C\lambda^2\|M\|_{\mathsf{HS}}^2\big)$$

for every $\lambda \in \mathbb{R}$ such that $|\lambda| \leq c/\|M\|_{\mathsf{op}}$.

Proof. *Step 1. Separating the diagonal and off-diagonal parts.* We can break the quadratic form as follows:

$$S := x^{\mathsf{T}}Mx - \operatorname{tr} M = \sum_{i=1}^n M_{ii}\big(x_i^2 - 1\big) + \sum_{i,j:i\neq j} M_{ij}x_ix_j =: S_{\mathrm{diag}} + S_{\mathrm{offdiag}}.$$

By Cauchy–Schwarz inequality, we have

$$\mathbb{E}\exp(\lambda S) \leq \big[\mathbb{E}\exp(2\lambda S_{\mathrm{diag}})\big]^{1/2}\big[\mathbb{E}\exp(2\lambda S_{\mathrm{offdiag}})\big]^{1/2}.$$

Let us consider the diagonal and off-diagonal parts separately.

Step 2. Diagonal part. Since x_i are sub-Gaussian random variables with unit variance, $x_i^2 - 1$ are mean-zero, subexponential random variables, and

$$\big\|x_i^2 - 1\big\|_{\psi_1} \lesssim \big\|x_i^2\big\|_{\psi_1} = \|x_i\|_{\psi_2}^2 \lesssim 1.$$

(This is a combination of some basic facts about sub-Gaussian and subexponential distributions, see [35], Exercise 2.7.10 and Lemma 2.7.6.) Then a standard bound on the MGF of a mean-zero, subexponential distribution ([35], Property 5 in Proposition 2.7.1) gives

$$\mathbb{E}\exp\big(\lambda_i\big(x_i^2 - 1\big)\big) \leq \exp\big(C\lambda_i^2\big) \quad \text{if } |\lambda_i| \leq c. \tag{2.1}$$

Therefore

$$\begin{aligned}\mathbb{E}\exp(2\lambda S_{\mathrm{diag}}) &= \prod_{i=1}^{n}\mathbb{E}\exp\bigl(2\lambda M_{ii}\bigl(x_i^2-1\bigr)\bigr) \quad \text{(by independence)}\\ &\le \exp\left(C\lambda^2\sum_{i=1}^{n}M_{ii}^2\right) \quad \text{if } |\lambda|\le\frac{c}{|M_{ii}|} \quad \text{(by (2.1))}\\ &\le \exp\bigl(C\lambda^2\|M\|_{\mathsf{HS}}^2\bigr) \quad \text{if } |\lambda|\le\frac{c}{\|M\|_{\mathsf{op}}}.\end{aligned}$$

Step 2. Off-diagonal part. Let $x_1',\dots,x_n'$ be independent copies of $x_1,\dots,x_n$. We have

$$\begin{aligned}\mathbb{E}\exp(2\lambda S_{\mathrm{offdiag}}) &= \mathbb{E}\exp\left(2\lambda\sum_{i,j:i\ne j}M_{ij}x_ix_j\right)\\ &\le \mathbb{E}\exp\left(8\lambda\sum_{i,j=1}^{n}M_{ij}x_ix_j'\right) \quad \text{(by decoupling, see [35], Remark 6.1.3)}\\ &\le \exp\bigl(C\lambda^2\|M\|_{\mathsf{HS}}^2\bigr) \quad \text{if } |\lambda|\le\frac{c}{\|M\|_{\mathsf{op}}},\end{aligned}$$

where the last bound follows from [35], Lemmas 6.2.3 and 6.2.2.

Combining the diagonal and off-diagonal contributions, we complete the proof. □

Corollary 2.3. *Let H be a Hilbert space and $A:(\mathbb{R}^n,\|\cdot\|_2)\to H$ be a linear operator. Let x be a random vector in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then*

$$\mathbb{E}\exp\bigl(\lambda\bigl(\|Ax\|_H^2-\|A\|_{\mathsf{HS}}^2\bigr)\bigr)\le\exp\bigl(C\lambda^2\|A\|_{\mathsf{op}}^2\|A\|_{\mathsf{HS}}^2\bigr)$$

for every $\lambda\in\mathbb{R}$ such that $|\lambda|\le c/\|A\|_{\mathsf{op}}^2$.

Proof. Apply Proposition 2.2 for $M:=A^*A$ and note that

$$x^{\mathsf{T}}Mx=\|Ax\|_H^2,\quad \operatorname{tr}M=\|A\|_{\mathsf{HS}}^2,\quad \|M\|_{\mathsf{op}}=\|A\|_{\mathsf{op}}^2,\quad \|M\|_{\mathsf{HS}}\le\|A\|_{\mathsf{op}}\|A\|_{\mathsf{HS}}.$$ □

Note in passing that Corollary 2.3 implies Euclidean concentration Theorem 1.2. All one needs to do is use exponential Markov's inequality and optimize the resulting bound in λ. We leave this as an exercise.

2.3. Euclidean functions

Theorems 1.2 and 1.4 can be conveniently stated as results about concentration of Euclidean functions.

Definition 2.4 (Euclidean functions). A function $f : \mathbb{R}^n \to [0, \infty)$ is called a *Euclidean* function on $\mathbb{R}^n$ if it can be expressed as

$$f(x) = \|Ax\|_H$$

where H is a Hilbert space and $A : \mathbb{R}^n \to H$ is a linear operator. Equivalently, f is Euclidean if f^2 is a positive-semidefinite quadratic form, that is, if

$$f(x)^2 = x^{\mathsf{T}} M x$$

for some $n \times n$ positive-semidefinite matrix M.

Let us note a few obvious facts about Euclidean functions.

Lemma 2.5 (Properties of Euclidean functions).

(i) *If f is a Euclidean function on $\mathbb{R}^n$, then af is, for any $a \geq 0$.*
(ii) *If f and g are Euclidean functions on $\mathbb{R}^n$ then $\sqrt{f^2 + g^2}$ is.*
(iii) *If f is a random Euclidean function on $\mathbb{R}^n$, then $(\mathbb{E} f^2)^{1/2}$ is.*
(iv) *The Lipschitz norm of a Euclidean function can be computed as follows*:

$$\|f\|_{\mathsf{Lip}} = \max_{x \in \mathbb{R}^n : \|x\|_2 = 1} f(x).$$

In particular, if $f(x) = \|Ax\|_H$ then $\|f\|_{\mathsf{Lip}} = \|A\|_{\mathsf{op}}$.

For future convenience, we restate Corollary 2.3 in terms of Euclidean functions.

Corollary 2.6 (MGF of a Euclidean function). *Let $f : (\mathbb{R}^n, \|\cdot\|_2) \to [0, \infty)$ be a Euclidean function. Let x be a random vector in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then*

$$\mathbb{E} \exp\bigl(\lambda\bigl(f(x)^2 - \mathbb{E} f(x)^2\bigr)\bigr) \leq \exp\bigl(C\lambda^2 \|f\|_{\mathsf{Lip}}^2 \, \mathbb{E} f(x)^2\bigr)$$

for every $\lambda \in \mathbb{R}$ such that $|\lambda| \leq c/\|f\|_{\mathsf{Lip}}^2$.

3. A maximal inequality

The proof of both of our main results, Theorems 1.3 and 1.4, relies on a tight control of the norm of the simple random tensor

$$\|X\|_2 = \|x_1 \otimes \cdots \otimes x_d\|_2 = \prod_{i=1}^{d} \|x_i\|_2. \tag{3.1}$$

Lemma 3.1 (The norm of a random tensor). *Let $x_1, \dots, x_d \in \mathbb{R}^n$ be independent random vectors with independent, mean zero, unit variance, sub-Gaussian coordinates. Then, for every $0 \le t \le 2n^{d/2}$, we have*

$$\mathbb{P}\left\{\prod_{i=1}^{d} \|x_i\|_2 > n^{d/2} + t\right\} \le 2\exp\left(-\frac{ct^2}{dn^{d-1}}\right).$$

Note in passing that this result is a partial case of Theorem 1.3 for the function $f(X) = \|X\|_2$ and of Theorem 1.4 for the identity map A.

Proof. Let $s \ge 0$. Then

$$\mathbb{P}\left\{\prod_{i=1}^{d} \|x_i\|_2 > (\sqrt{n} + s)^d\right\} \le \mathbb{P}\left\{\frac{1}{d}\sum_{i=1}^{d}\big(\|x_i\|_2 - \sqrt{n}\big) > s\right\}. \tag{3.2}$$

To check this, take d-th root on both sides of the inequality $\prod_{i=1}^{d} \|x_i\|_2 > (\sqrt{n} + s)^d$, apply the inequality of arithmetic and geometric means, and subtract $\sqrt{n}$ from both parts. Furthermore, we can replace all terms $\sqrt{n}$ in (3.2) by the means $\mu_i := \mathbb{E}\,\|x_i\|_2$ since they satisfy

$$\mu_i \le \big(\mathbb{E}\,\|x_i\|_2^2\big)^{1/2} = \sqrt{n}.$$

Thus the probability in (3.2) is bounded by

$$\mathbb{P}\left\{\frac{1}{d}\sum_{i=1}^{d}\big(\|x_i\|_2 - \mu_i\big) > s\right\}, \tag{3.3}$$

which is a tail probability for a sum of independent, mean zero random variables.

By the concentration of the norm (Corollary 2.1) and a standard centering argument ([35], Lemma 2.6.8), we have[2]

$$\big\|\,\|x_i\|_2 - \mu_i\big\|_{\psi_2} \le C, \quad i = 1, \dots, d,$$

and this implies that

$$\left\|\frac{1}{d}\sum_{i=1}^{d}\big(\|x_i\|_2 - \mu_i\big)\right\|_{\psi_2} \le \frac{C}{\sqrt{d}},$$

see [35], Proposition 2.6.1. Thus the probability in (3.3) is bounded by

$$2\exp\big(-cs^2 d\big).$$

[2]Here $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$ denote the sub-Gaussian and subexponential norms, respectively; see [35], Sections 2.5, 2.7, for definition and basic properties.

Let $0 \le u \le 2$ and apply this bound for $s := u\sqrt{n}/(2d)$. With this choice,

$$(\sqrt{n}+s)^d = n^{d/2}\left(1+\frac{u}{2d}\right)^d \le n^{d/2}(1+u).$$

Thus we have shown that

$$\mathbb{P}\left\{\prod_{i=1}^d \|x_i\|_2 > n^{d/2}(1+u)\right\} \le 2\exp\left(-\frac{cnu^2}{4d}\right). \tag{3.4}$$

Using this inequality for $u := t/n^{d/2}$, we complete the proof. □

A stronger statement will be needed in the proof of our main results: we will require a tight control of the products (3.4) *for all d simultaneously*. The following maximal inequality will be used for the following.

Lemma 3.2 (A maximal inequality). *Let $x_1, \dots, x_d \in \mathbb{R}^n$ be independent random vectors with independent, mean zero, unit variance, sub-Gaussian coordinates. Then, for every $0 \le u \le 2$, we have*

$$\mathbb{P}\left\{\max_{1\le k\le d} n^{-k/2}\prod_{i=1}^k \|x_i\|_2 > 1+u\right\} \le 2\exp\left(-\frac{cnu^2}{d}\right).$$

Proof. *Step 1. A binary partition.* By increasing d if necessary, we can assume that

$$d = 2^L \quad \text{for some } L \in \mathbb{N}.$$

For each level $\ell \in \{0, 1, \dots, L\}$, consider the partition $\mathcal{I}_\ell$ of the integer interval $[1, d] = \{1, \dots, d\}$ into 2^ℓ successive intervals of length

$$d_\ell := \frac{d}{2^\ell}.$$

We call each of these intervals a *binary interval*. For example, the family $\mathcal{I}_0$ consists of just one binary interval $[1, d]$, and the family $\mathcal{I}_1$ consists of two binary intervals $[1, d/2]$ and $[d/2+1, d]$.

For every integer $k \in [1, d]$, the interval $[1, k]$ can be partitioned into binary intervals of different lengths. (The binary representation of the number k/d determines which intervals participate in this partition.) As a consequence, such partition of $[1, k]$ must include no more than one interval from each family $\mathcal{I}_\ell$.

Step 2. Controlling the product over binary sets. Fix a level $\ell \in \{0, 1, \dots, L\}$ and a binary interval $I \in \mathcal{I}_\ell$. Apply Lemma 3.1 with d replaced by $|I| = d_\ell = d/2^\ell$ and for $t := 2^{-\ell/4} n^{d_\ell/2} u$. It gives

$$\mathbb{P}\left\{\prod_{i\in I} \|x_i\|_2 > \left(1+2^{-\ell/4}u\right)n^{d_\ell/2}\right\} \le 2\exp\left(-\frac{c_0 nu^2}{2^{\ell/2}d_\ell}\right) = 2\exp\left(-2^{\ell/2}\cdot\frac{c_0 nu^2}{d}\right).$$

Taking a union bound over all levels ℓ and all 2^ℓ binary intervals I in the family $\mathcal{I}_\ell$, we get

$$\mathbb{P}\left\{\exists \ell \in \{0,\dots,L\}, \exists I \in \mathcal{I}_\ell : \prod_{i\in I} \|x_i\|_2 > \big(1+2^{-\ell/4}u\big)n^{d_\ell/2}\right\}$$
$$\le \sum_{\ell=0}^{L} 2^\ell \cdot 2\exp\left(-2^{\ell/2}\cdot\frac{c_0 n u^2}{d}\right).$$

To simplify this bound, we can assume that $c_0 n u^2/d \ge 1$, otherwise the probability bound in the conclusion of the lemma becomes trivial if $c < c_0/2$. Also, $2^{\ell/2} \ge 1$, and thus

$$2^{\ell/2}\cdot\frac{c_0 n u^2}{d} \ge \frac{1}{2}\left(2^{\ell/2} + \frac{c_0 n u^2}{d}\right).$$

Substituting this into our probability bound, we can continue it as

$$2\exp\left(-\frac{c_0 n u^2}{2d}\right)\cdot\sum_{\ell=0}^{L} 2^\ell \cdot 2\exp\big(-2^{\ell/2-1}\big) \le C\exp\left(-\frac{c_0 n u^2}{2d}\right).$$

By reducing the absolute constant c_0, we can make $C = 2$.

Step 3. Controlling the product over any interval. Let us fix a realization of random vectors for which the good event considered above occurs, that is,

$$\prod_{i\in I}\|x_i\|_2 \le \big(1+2^{-\ell/4}u\big)n^{d_\ell/2} \quad \text{for every } \ell \in \{0,\dots,L\} \text{ and } I \in \mathcal{I}_\ell. \tag{3.5}$$

Let $1 \le k \le d$. As we noted in Step 1, we can partition the interval $[1,k]$ into binary intervals $I \in \mathcal{I}_\ell$ so that at most one binary interval is taken from each family $\mathcal{I}_\ell$. Let us multiply the inequalities (3.5) for all binary intervals I that participate in this partition. Note that sum of exponents d_ℓ is the sum of the length of these intervals I, which equals k. Thus, we obtain

$$\prod_{i=1}^{k}\|x_i\|_2 \le n^{k/2}\prod_{\ell=0}^{L}\big(1+2^{-\ell/4}u\big) \le n^{k/2}\exp\left(u\sum_{\ell=0}^{L}2^{-\ell/4}\right) \quad \big(\text{using } 1+x\le e^x\big)$$
$$\le n^{k/2}\exp(Cu) \le n^{k/2}\big(1+e^{2C}u\big) \quad (\text{since } 0\le u\le 2).$$

This yields the conclusion of the lemma with Cu instead of u in the bound. One can get rid of C by reducing the constant c in the probability bound. The proof is complete. □

4. Proof of Theorem 1.3

Proof of Theorem 1.3. *Step 1. Applying the maximal inequality.* We can assume without loss of generality that $\|f\|_{\mathsf{Lip}} = 1$. Consider the events

$$\mathcal{E}_k := \left\{ \prod_{i=k}^{d} \|x_i\|_2 \le 2n^{(d-k+1)/2} \right\}, \quad k = 1, \dots, d,$$

and let $\mathcal{E}_{d+1}$ be the entire probability space for convenience. Applying the maximal inequality of Lemma 3.2 for $u = 1$ and for the reverse ordering of the vectors, we see that the event

$$\mathcal{E} := \mathcal{E}_2 \cap \cdots \cap \mathcal{E}_d$$

is likely:

$$\mathbb{P}(\mathcal{E}) \ge 1 - 2\exp\left(-\frac{cn}{d}\right). \tag{4.1}$$

Step 2. Applying the convex concentration inequality. Fix any realization of the random vectors $x_2, \dots, x_d$ that satisfy $\mathcal{E}_2$ and apply the Convex Concentration Theorem 1.1 for f as a function of x_1. It is a convex and Lipschitz function. To get a quantitative bound on its Lipschitz norm, consider any $x, y \in \mathbb{R}^n$ and note that

$$\begin{aligned} &\big|f(x \otimes x_2 \otimes \cdots \otimes x_d) - f(y \otimes x_2 \otimes \cdots \otimes x_d)\big| \\ &\quad \le \big\|(x-y) \otimes x_2 \otimes \cdots \otimes x_d\big\|_2 \quad \big(\text{since } \|f\|_{\mathsf{Lip}} = 1\big) \\ &\quad = \|x-y\|_2 \cdot \prod_{i=2}^{d} \|x_i\|_2 \le \|x-y\|_2 \cdot 2n^{(d-1)/2} \quad (\text{since } \mathcal{E}_2 \text{ holds}). \end{aligned}$$

This shows that f as a function of x_1 has Lipschitz norm bounded by

$$L := 2n^{(d-1)/2}. \tag{4.2}$$

The Convex Concentration theorem 1.1 then yields

$$\|f - \mathbb{E}_{x_1} f\|_{\psi_2(x_1)} \le CL \quad \text{for any } x_2, \dots, x_d \text{ that satisfy } \mathcal{E}_2.$$

In this inequality, x_1 indicates that the expectation and the ψ_2 norm is taken with respect to the random vector x_1, that is, conditioned on all other random vectors.

Fix any realization of the random vectors $x_3, \dots, x_d$ that satisfy $\mathcal{E}_3$ and apply the Convex Concentration theorem 1.1 for $\mathbb{E}_{x_1} f$ as a function of x_2. It is a convex and Lipschitz function. To get a quantitative bound on its Lipschitz norm, consider any $x, y \in \mathbb{R}^n$ and note that

$$\begin{aligned} &\big|\mathbb{E}_{x_1} f(x_1 \otimes x \otimes x_3 \otimes \cdots \otimes x_d) - \mathbb{E}_{x_1} f(x_1 \otimes y \otimes x_3 \otimes \cdots \otimes x_d)\big| \\ &\quad \le \mathbb{E}_{x_1} \big\|x_1 \otimes (x-y) \otimes x_3 \otimes \cdots \otimes x_d\big\|_2 \quad \big(\text{by Jensen's inequality and since } \|f\|_{\mathsf{Lip}} = 1\big) \end{aligned}$$

$$\le \big(\mathbb{E}_{x_1} \|x_1\|_2^2\big)^{1/2} \cdot \|x-y\|_2 \cdot \prod_{i=3}^{d} \|x_i\|_2$$
$$\le \sqrt{n} \cdot \|x-y\|_2 \cdot 2n^{(d-2)/2} \quad \text{(since } \mathcal{E}_3 \text{ holds)}$$
$$= \|x-y\|_2 \cdot 2n^{(d-1)/2}.$$

This shows that $\mathbb{E}_{x_1} f$ as a function of x_2 has Lipschitz norm bounded by $L = 2n^{(d-1)/2}$. The Convex Concentration theorem 1.1 then yields

$$\|\mathbb{E}_{x_1} f - \mathbb{E}_{x_1,x_2} f\|_{\psi_2(x_2)} \le CL \quad \text{for any } x_3,\dots,x_d \text{ that satisfy } \mathcal{E}_3.$$

Continuing in a similar way, we can show that for every $k = 1,\dots,d$:

$$\|\mathbb{E}_{x_1,\dots,x_{k-1}} f - \mathbb{E}_{x_1,\dots,x_k} f\|_{\psi_2(x_k)} \le CL \quad \text{for any } x_{k+1},\dots,x_d \text{ that satisfy } \mathcal{E}_{k+1}. \tag{4.3}$$

Step 3. Combining the increments using a martingale-like argument. Let us look at the differences

$$\Delta_k = \Delta_k(x_k,\dots,x_d) := \mathbb{E}_{x_1,\dots,x_{k-1}} f - \mathbb{E}_{x_1,\dots,x_k} f.$$

The estimate (4.3) on the sub-Gaussian norm yields the following bound on the moment generating function [35], Proposition 2.5.2:

$$\mathbb{E}_{x_k} \exp(\lambda\Delta_k) \le \exp\big(CL^2\lambda^2\big) \quad \text{for any } x_{k+1},\dots,x_d \text{ that satisfy } \mathcal{E}_{k+1}$$

and for any $\lambda \in \mathbb{R}$. We can combine these pieces using a martingale-like argument, which we defer to Lemma 4.1. It gives that for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\exp\big(\lambda(f-\mathbb{E}f)\big)\mathbf{1}_{\mathcal{E}} = \mathbb{E}\exp\big(\lambda(\Delta_1+\dots+\Delta_d)\big)\mathbf{1}_{\mathcal{E}} \le \exp\big(CdL^2\lambda^2\big) \tag{4.4}$$

where $\mathcal{E} = \mathcal{E}_2 \cap \dots \cap \mathcal{E}_d$ is the event whose probability we estimated in (4.1).

Step 4. Deriving the concentration via exponential Markov's inequality. To derive a probability bound, we can use a standard argument based on exponential Markov's inequality. Namely, we have for every $\lambda > 0$:

$$\begin{aligned}
\mathbb{P}\{f-\mathbb{E}f > t\} &\le \mathbb{P}\{f-\mathbb{E}f > t \text{ and } \mathcal{E}\} + \mathbb{P}\big(\mathcal{E}^c\big) \\
&= \mathbb{P}\big\{\exp\big(\lambda(f-\mathbb{E}f)\big)\mathbf{1}_{\mathcal{E}} > \exp(\lambda t)\big\} + \mathbb{P}\big(\mathcal{E}^c\big) \\
&\le \exp(-\lambda t)\,\mathbb{E}\exp\big(\lambda(f-\mathbb{E}f)\big)\mathbf{1}_{\mathcal{E}} + \mathbb{P}\big(\mathcal{E}^c\big) \quad \big(\text{by Markov's inequality}\big) \\
&\le \exp\big(-\lambda t + CdL^2\lambda^2\big) + 2\exp\Big(-\frac{cn}{d}\Big) \quad \text{(by (4.4) and (4.1)).}
\end{aligned}$$

This bound is minimized for $\lambda := t/(2CdL^2)$. With this choice of λ, and with the choice of L made in (4.2), our bound becomes

$$\mathbb{P}\{f-\mathbb{E}f > t\} \le \exp\Big(-\frac{t^2}{16Cdn^{d-1}}\Big) + 2\exp\Big(-\frac{cn}{d}\Big). \tag{4.5}$$

Since by assumption of the theorem,

$$t^2 \le 4\,\mathbb{E}\big|f(x_1\otimes\cdots\otimes x_d)\big|^2 \le 4\,\mathbb{E}\,\|x_1\otimes\cdots\otimes x_d\|_2^2 \quad \big(\text{since } \|f\|_{\mathsf{Lip}}=1\big)$$
$$= 4\,\mathbb{E}\,\|x_1\|_2^2\cdots\|x_d\|_2^2 = 4n^d,$$

we have

$$\frac{t^2}{dn^{d-1}} \le \frac{4n}{d}.$$

This implies that the first term in the bound (4.5) dominates over the second, if the constant C is sufficiently large compared to $1/c$. This gives

$$\mathbb{P}\{f-\mathbb{E}\,f>t\} \le 3\exp\bigg(-\frac{t^2}{16Cdn^{d-1}}\bigg).$$

Finally, repeating the argument for $-f$ instead of f we obtain the same probability bound for $\mathbb{P}\{-f+\mathbb{E}\,f>t\}$. Combining the two bounds, we get

$$\mathbb{P}\big\{|f-\mathbb{E}\,f|>t\big\} \le 6\exp\bigg(-\frac{t^2}{16Cdn^{d-1}}\bigg).$$

We can replace the factor 6 by 2 by making C larger if necessary. Theorem 1.3 is proved. □

Our argument above used on the following martingale-like inequality, which we will carefully state and prove now.

Lemma 4.1 (A martingale-type inequality). *Let $x_1,\dots,x_d$ be independent random vectors. For each $k=1,\dots,d$, let $f_k=f_k(x_k,\dots,x_d)$ be an integrable real-valued function and $\mathcal{E}_k$ be an event that is uniquely determined by the vectors $x_{k+1},\dots,x_d$. Let $\mathcal{E}_{d+1}$ be the entire probability space for convenience. Suppose that, for every $k=1,\dots,d$:*

$$\mathbb{E}_{x_k}\exp(f_k)\le\pi_k \quad \textit{for every choice of } x_{k+1},\dots,x_d \textit{ satisfying } \mathcal{E}_{k+1}.$$

Then, for $\mathcal{E}:=\mathcal{E}_2\cap\cdots\cap\mathcal{E}_d$, we have

$$\mathbb{E}\exp(f_1+\cdots+f_d)\mathbf{1}_{\mathcal{E}}\le\pi_1\cdots\pi_d.$$

Proof. We have

$$\mathbb{E}\exp(f_1+\cdots+f_d)\mathbf{1}_{\mathcal{E}_2\cap\cdots\cap\mathcal{E}_d} = \mathbb{E}\exp(f_2+\cdots+f_d)\mathbf{1}_{\mathcal{E}_3\cap\cdots\cap\mathcal{E}_d}\,\mathbb{E}_{x_1}\exp(f_1)\mathbf{1}_{\mathcal{E}_2}$$
$$\le \pi_1\,\mathbb{E}\exp(f_2+\cdots+f_d)\mathbf{1}_{\mathcal{E}_3\cap\cdots\cap\mathcal{E}_d}$$

since $\mathbb{E}_{x_1}\exp(f_1)\mathbf{1}_{\mathcal{E}_2}\le\pi_1$ a.s. by assumption. Iterating this argument, we complete the proof. □

5. Proof of Theorem 1.4

Let us restate Theorem 1.4 in terms of Euclidean functions which were introduced in Section 2.3.

Theorem 5.1 (Euclidean concentration for random tensors). *Let n and d be positive integers and $f : (\mathbb{R}^{n^d}, \|\cdot\|_2) \to [0,\infty)$ be a Euclidean function. Consider a simple random tensor $X := x_1 \otimes \cdots \otimes x_d$ in $\mathbb{R}^{n^d}$, where all x_k are independent random vectors in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then, for every $0 \le t \le 2(\mathbb{E}\, f(X)^2)^{1/2}$, we have*

$$\mathbb{P}\big\{\big|f(X) - \big(\mathbb{E}\, f(X)^2\big)^{1/2}\big| \ge t\big\} \le 2\exp\bigg(-\frac{ct^2}{dn^{d-1}\|f\|_{\mathsf{Lip}}^2}\bigg).$$

Here $c > 0$ depends only on the bound on the sub-Gaussian norms.

Proof. *Step 1. Applying the maximal inequality.* The proof starts as in the proof of Theorem 1.3 in Section 4. We define the norm-controlling events $\mathcal{E}_k$ and estimate the probability of $\mathcal{E} = \mathcal{E}_2 \cap \cdots \mathcal{E}_d$ by a maximal inequality in the same way as before.

Step 2. Applying a sub-Gaussian concentration inequality. Fix any realization of the random vectors $x_2, \dots, x_d$ that satisfy $\mathcal{E}_2$ and apply Corollary 2.6 for f as a function of x_1. It is a Euclidean function, and one can check as before that its Lipschitz norm is bounded by

$$L := 2n^{(d-1)/2}. \tag{5.1}$$

Corollary 2.6 then yields

$$\mathbb{E}_{x_1} \exp\big(\lambda\big(f^2 - \mathbb{E}_{x_1} f^2\big)\big) \le \exp\big(C\lambda^2 L^2\, \mathbb{E}_{x_1} f^2\big)$$

provided that $|\lambda| \le c/L^2$. For future convenience, let us restate this bound as follows. Choose $\lambda \in \mathbb{R}$ and denote

$$\lambda_0 := \lambda; \qquad \lambda_1 := \lambda_0 + C\lambda_0^2 L^2.$$

Then we have

$$\mathbb{E}_{x_1} \exp\big(\lambda_0 f^2 - \lambda_1\, \mathbb{E}_{x_1} f^2\big) \le 1 \quad \text{for any } x_2, \dots, x_d \text{ that satisfy } \mathcal{E}_2,$$

provided that $|\lambda_0| \le c/L^2$.

Fix any realization of the random vectors $x_3, \dots, x_d$ that satisfy $\mathcal{E}_3$ and apply Corollary 2.6 for $(\mathbb{E}_{x_1} f^2)^{1/2}$ as a function of x_2. It is a Euclidean function whose Lipschitz norm is bounded by L as before. Corollary 2.6 then yields

$$\mathbb{E}_{x_2} \exp\big(\lambda_1\big(\mathbb{E}_{x_1} f^2 - \mathbb{E}_{x_1,x_2} f^2\big)\big) \le \exp\big(C\lambda_1^2 L^2\, \mathbb{E}_{x_1,x_2} f^2\big)$$

provided that $|\lambda_1| \le c/L^2$. We can restate this bound as follows. Denote

$$\lambda_2 := \lambda_1 + C\lambda_1^2 L^2.$$

Then we have

$$\mathbb{E}_{x_2} \exp\bigl(\lambda_1 \mathbb{E}_{x_1} f^2 - \lambda_2 \mathbb{E}_{x_1,x_2} f^2\bigr) \le 1 \quad \text{for any } x_3, \dots, x_d \text{ that satisfy } \mathcal{E}_3,$$

provided that $|\lambda_1| \le c/L^2$.

Continuing in a similar way, we can show the following for every $k = 1, \dots, d$. Denote

$$\lambda_k := \lambda_{k-1} + C\lambda_{k-1}^2 L^2.$$

Then we have

$$\mathbb{E}_{x_2} \exp\bigl(\lambda_{k-1} \mathbb{E}_{x_1,\dots,x_{k-1}} f^2 - \lambda_k \mathbb{E}_{x_1,\dots,x_k} f^2\bigr) \le 1 \quad \forall x_{k+1}, \dots, x_d \text{ that satisfy } \mathcal{E}_{k+1},$$

provided that $|\lambda_{k-1}| \le c/L^2$.

Step 3. Combining the increments using a martingale-like argument. Combining the pieces into a telescoping sum using Lemma 4.1, we obtain

$$\mathbb{E} \exp\bigl(\lambda_0 f^2 - \lambda_d \mathbb{E} f^2\bigr) \mathbf{1}_{\mathcal{E}} \le 1 \tag{5.2}$$

provided that

$$|\lambda_k| \le \frac{c}{L^2} \quad \text{for all } k = 0, \dots, d-1. \tag{5.3}$$

If we choose $\lambda = \lambda_0 \in \mathbb{R}$ so that $|\lambda| \le c_0/(dL^2)$ with a sufficiently small absolute constant c_0, then we can show by induction that (5.3) holds and, moreover,

$$\lambda_d \le \lambda + 2CdL^2\lambda^2.$$

We defer the verification of both of these bounds to Lemma 5.2 below. Substituting them into (5.2) and rearranging the terms, we conclude that

$$\mathbb{E} \exp\bigl(\lambda\bigl(f^2 - \mathbb{E} f^2\bigr)\bigr) \mathbf{1}_{\mathcal{E}} \le \exp\bigl(2CdL^2\lambda^2 \mathbb{E} f^2\bigr) \quad \text{if } |\lambda| \le \frac{c_0}{dL^2}.$$

Replacing λ with $-\lambda$, we see that the same bound holds for $\mathbb{E} \exp(-\lambda f^2 + \lambda \mathbb{E} f^2)$. Since the inequality $e^{|z|} \le e^z + e^{-z}$ holds for all $z \in \mathbb{R}$, we obtain

$$\mathbb{E} \exp\bigl(\lambda\bigl|f^2 - \mathbb{E} f^2\bigr|\bigr) \mathbf{1}_{\mathcal{E}} \le 2 \exp\bigl(2CdL^2\lambda^2 \mathbb{E} f^2\bigr) \quad \text{if } |\lambda| \le \frac{c_0}{dL^2}.$$

Step 4. Deriving a concentration inequality for f^2. Using the exponential Markov's inequality just like we did in the proof of Theorem 1.3 in Section 4, we get

$$\mathbb{P}\bigl\{\bigl|f^2 - \mathbb{E} f^2\bigr| > u\bigr\} \le 2 \exp\bigl(-\lambda u + 2CdL^2\lambda^2 \mathbb{E} f^2\bigr) + 2 \exp\biggl(-\frac{cn}{d}\biggr)$$

for any $u > 0$ and any $0 \le \lambda \le c_0/(dL^2)$.

Let us optimize the right-hand side in λ. A good choice is

$$\lambda := \frac{c_1}{dL^2} \min\left(\frac{u}{\mathbb{E} f^2}, 1\right)$$

for a sufficiently small constant $c > 0$. Indeed, if $c_1 \le c_0$ then λ lies in the required range $0 \le \lambda \le c_0/(dL^2)$, and if $c_1 \le 1/(4C)$, then substituting this choice of λ into our probability bound gives

$$\mathbb{P}\{|f^2 - \mathbb{E} f^2| > u\} \le 2\exp\left(-\frac{c_1}{2dL^2} \min\left(\frac{u^2}{\mathbb{E} f^2}, u\right)\right) + 2\exp\left(-\frac{cn}{d}\right).$$

Step 5. Deriving a concentration inequality for f. Choose any $\varepsilon \ge 0$ and substitute $u := \varepsilon \mathbb{E} f^2$ into our probability bound. We get

$$\mathbb{P}\{|f^2 - \mathbb{E} f^2| > \varepsilon \mathbb{E} f^2\} \le 2\exp\left(-\frac{c_1}{2dL^2} \min(\varepsilon^2, \varepsilon) \mathbb{E} f^2\right) + 2\exp\left(-\frac{cn}{d}\right).$$

Now choose any $\delta \ge 0$ and apply this bound for $\varepsilon := \max(\delta, \delta^2)$. Then $\min(\varepsilon^2, \varepsilon) = \delta^2$, and one can easily check the following implication

$$|f - (\mathbb{E} f^2)^{1/2}| > \delta (\mathbb{E} f^2)^{1/2} \quad \Longrightarrow \quad |f^2 - \mathbb{E} f^2| > \varepsilon \mathbb{E} f^2.$$

(This follows from the implication $|z - 1| \ge \delta \Rightarrow |z^2 - 1| \ge \max(\delta, \delta^2)$ that is valid for all $z \ge 0$.) Hence, we obtain

$$\mathbb{P}\{|f - (\mathbb{E} f^2)^{1/2}| > \delta (\mathbb{E} f^2)^{1/2}\} \le 2\exp\left(-\frac{c_1 \delta^2 \mathbb{E} f^2}{2dL^2}\right) + 2\exp\left(-\frac{cn}{d}\right).$$

Now choose any $t \ge 0$ and apply this bound for $\delta := t/(\mathbb{E} f^2)^{1/2}$. Recalling the value of L from (5.1), we get

$$\mathbb{P}\{|f - (\mathbb{E} f^2)^{1/2}| > t\} \le 2\exp\left(-\frac{c_1 t^2}{8dn^{d-1}}\right) + 2\exp\left(-\frac{cn}{d}\right).$$

Finally, we can use the theorem's assumption on t to get rid of the second exponential term just like we did in in the proof of Theorem 1.3 in Section 4. The proof is complete. □

In Step 3 of the argument above, we used the following bound on the multipliers λ_k, which we promised to prove later. Let us do it now.

Lemma 5.2 (Multipliers). *Let $d, M \ge 0$ and consider a number $\lambda_0 \in \mathbb{R}$ such that*

$$|\lambda_0| \le \frac{1}{8dM}. \tag{5.4}$$

Define $\lambda_1, \dots, \lambda_d \in \mathbb{R}$ *inductively by the formula*

$$\lambda_k := \lambda_{k-1} + M\lambda_{k-1}^2, \quad k = 1, \dots, d.$$

Then, for every $k = 1, \dots, d$, *we have*:

$$|\lambda_k| \le \frac{1}{6dM} \quad \text{and} \quad \lambda_k \le \lambda_0 + 2kM\lambda_0^2.$$

Proof. We can prove the second inequality in the conclusion by induction. Assume that it holds for some k, that is,

$$\lambda_k \le \lambda_0 + 2kM\lambda_0^2. \tag{5.5}$$

By construction, the sequence (λ_k) is increasing, so the triangle inequality gives

$$|\lambda_k| \le |\lambda_0| + |\lambda_k - \lambda_0| \le |\lambda_0| + \lambda_k - \lambda_0 \le |\lambda_0| + 2kM\lambda_0^2. \tag{5.6}$$

Furthermore, the assumption (5.4) implies that

$$2kM\lambda_0^2 \le 2dM\lambda_0^2 \le \frac{|\lambda_0|}{4}.$$

Substituting this into (5.6), we get

$$|\lambda_k| \le \frac{5}{4}|\lambda_0|. \tag{5.7}$$

Then we have

$$\begin{aligned}
\lambda_{k+1} &= \lambda_k + M\lambda_k^2 \quad \text{(by construction)} \\
&\le \lambda_0 + 2kM\lambda_0^2 + M\left(\frac{5}{4}|\lambda_0|\right)^2 \quad \text{(by (5.5) and (5.7))} \\
&\le \lambda_0 + 2(k+1)M\lambda_0^2.
\end{aligned}$$

Thus we proved (5.5) for $k+1$, so the second inequality in the conclusion is verified.

The first bound in the conclusion follows from the first. Indeed, using (5.7) and (5.6), we get

$$|\lambda_k| \le \frac{5}{4}|\lambda_0| \le \frac{5}{4} \cdot \frac{1}{8dM} \le \frac{1}{6dM}$$

as claimed. The proof is complete. □

6. Applications

In this section, we state and prove a full version of Theorem 1.6 that states that random tensors are well conditioned. But before we do so, let us prove a result that may have an independent interest, namely a concentration inequality for the distance between a random tensor X and a given subspace L.

Corollary 6.1 (Distance to a subspace). *Let n and d be positive integers and $L \subset \mathbb{R}^{n^d}$ be a linear subspace with $k := \operatorname{codim}(L)$. Consider a simple random tensor $X := x_1 \otimes \cdots \otimes x_d$ in $\mathbb{R}^{n^d}$, where all x_k are independent random vectors in $\mathbb{R}^n$ whose coordinates are independent, mean zero, unit variance, sub-Gaussian random variables. Then, for every $0 \le t \le 2\sqrt{k}$, we have*

$$\mathbb{P}\big\{\big|\operatorname{dist}(X, L) - \sqrt{k}\big| \ge t\big\} \le 2\exp\left(-\frac{ct^2}{dn^{d-1}}\right).$$

Here $c > 0$ depends only on the bound on the sub-Gaussian norms.

Proof. Apply Theorem 1.4 for the orthogonal projection P in $\mathbb{R}^{n^d}$ onto $L^\perp$ and note that $\|P\|_{\mathsf{op}} = 1$ and $\|P\|_{\mathsf{HS}} = \sqrt{\dim(L^\perp)} = \sqrt{k}$. □

For $d = 1$, Corollary 6.1 recovers the known optimal concentration inequalities for the distance between a random vector and a fixed subspace are known (see, e.g., [29], Corollary 2.1.19, [27, 35], Exercise 6.3.4), which are frequently used in random matrix theory. Some previously known extensions for tensors of degrees $d \ge 2$ were given in [1,8,9,11].

Now, we are ready to state and prove a rigorous version of Theorem 1.6.

Corollary 6.2 (Random tensors are well conditioned). *Consider independent simple sub-Gaussian random tensors $X_1, \dots, X_m$ (defined like a tensor X in Corollary 6.1). Let ε be such that $Cd^2 \log(n)/n \le \varepsilon \le 1/2$. If $m \le (1-\varepsilon)n^d$ then, with probability at least $1 - 2\exp(-c\varepsilon n/d)$, we have*

$$\left\| \sum_{i=1}^m a_i X_i \right\|_2 \ge \frac{\sqrt{\varepsilon}}{2} \|a\|_2 \quad \text{for all } a = (a_1, \dots, a_m) \in \mathbb{R}^m.$$

Note that this result is nontrivial for $d = O(\sqrt{n/\log n})$, because only in this range is the range of ε nonempty.

Proof. We can assume that $\|a\|_2 = 1$ without loss of generality. A simple "leave-one-out" bound gives

$$\left\| \sum_{i=1}^m a_i X_i \right\|_2 \ge \frac{1}{\sqrt{m}} \min_{j=1,\dots,m} \operatorname{dist}(X_j, L_j) \tag{6.1}$$

where L_j is the linear span of the $m-1$ vectors $(X_i)_{i \ne j}$. Since $\dim(L_j) \le m - 1 \le (1-\varepsilon)n^d$, we have $\operatorname{codim}(L_j) \ge \varepsilon n^d$.

Fix j and apply Corollary 6.1 with $t = \sqrt{\varepsilon n^d}/2$ conditionally on $(X_i)_{i \ne j}$. It gives

$$d(X_j, L_j) \ge \frac{\varepsilon n^d}{2} \tag{6.2}$$

with probability at least $1 - 2\exp(-c\varepsilon n/2d)$. Taking the union bound over $j = 1, \dots, m$, we conclude that all events (6.2) hold simultaneously with probability at least

$$1 - 2m \exp\left(-\frac{c\varepsilon n}{2d}\right) \geq 1 - 2\exp\left(-\frac{c\varepsilon n}{4d}\right),$$

where we used that $m \leq n^d$ and the assumption on ε. Substitute this into the leave-one-out bound (6.1) to complete the proof. □

Acknowledgement

The author is grateful to Mark Rudelson for useful comments on the preliminary version of this manuscript, in particular for pointing out an alternative approach to Theorem 1.1 mentioned in Remark 1.7.

References

[1] Abbe, E., Shpilka, A. and Wigderson, A. (2015). Reed–Muller codes for random erasures and errors. *IEEE Trans. Inf. Theory* **61** 5229–5252. MR3400278 https://doi.org/10.1109/TIT.2015.2462817

[2] Adamczak, R. (2005). Logarithmic Sobolev inequalities and concentration of measure for convex functions and polynomial chaoses. *Bull. Pol. Acad. Sci. Math.* **53** 221–238. MR2163396 https://doi.org/10.4064/ba53-2-10

[3] Adamczak, R., Guédon, O., Litvak, A., Pajor, A. and Tomczak-Jaegermann, N. (2008). Smallest singular value of random matrices with independent columns. *C. R. Math. Acad. Sci. Paris* **346** 853–856. MR2441920 https://doi.org/10.1016/j.crma.2008.07.011

[4] Adamczak, R. and Latała, R. (2012). Tail and moment estimates for chaoses generated by symmetric random variables with logarithmically concave tails. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1103–1136. MR3052405 https://doi.org/10.1214/11-AIHP441

[5] Adamczak, R., Latala, R. and Meller, R. (2020). Moments of Gaussian chaoses in Banach spaces. Preprint.

[6] Adamczak, R. and Strzelecki, M. (2015). Modified log-Sobolev inequalities for convex functions on the real line. Sufficient conditions. *Studia Math.* **230** 59–93. MR3456588 https://doi.org/10.4064/sm8319-12-2015

[7] 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 https://doi.org/10.1007/s00440-014-0579-3

[8] Anari, N., Daskalakis, C., Maass, W., Papadimitriou, C., Saberi, A. and Vempala, S. (2018). Smoothed analysis of discrete tensor decomposition and assemblies of neurons. In *Advances in Neural Information Processing Systems* 10857–10867.

[9] Anderson, J., Belkin, M., Goyal, N., Rademacher, L. and Voss, J. (2014). The more, the merrier: The blessing of dimensionality for learning large Gaussian mixtures. *J. Mach. Learn. Res. Workshop Conf. Proc.* **35** 1–30.

[10] Basak, A. and Rudelson, M. (2017). Invertibility of sparse non-Hermitian matrices. *Adv. Math.* **310** 426–483. MR3620692 https://doi.org/10.1016/j.aim.2017.02.009

[11] Bhaskara, A., Charikar, M., Moitra, A. and Vijayaraghavan, A. (2014). Smoothed analysis of tensor decompositions. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* 594–603. ACM.

[12] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. With a foreword by Michel Ledoux. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[13] Götze, F., Naumov, A. and Tikhomirov, A. (2015). On minimal singular values of random matrices with correlated entries. *Random Matrices Theory Appl.* **4** 1550006. MR3356884 https://doi.org/10.1142/S2010326315500069

[14] Götze, F., Sambale, H. and Sinulis, A. (2019). Concentration inequalities for polynomials in α-subexponential random variables. Preprint.

[15] Gozlan, N., Roberto, C., Samson, P.-M. and Tetali, P. (2017). Kantorovich duality for general transport costs and applications. *J. Funct. Anal.* **273** 3327–3405. MR3706606 https://doi.org/10.1016/j.jfa.2017.08.015

[16] Kahn, J., Komlós, J. and Szemerédi, E. (1995). On the probability that a random ± 1-matrix is singular. *J. Amer. Math. Soc.* **8** 223–240. MR1260107 https://doi.org/10.2307/2152887

[17] Latała, R. (2006). Estimates of moments and tails of Gaussian chaoses. *Ann. Probab.* **34** 2315–2331. MR2294983 https://doi.org/10.1214/009117906000000421

[18] Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Providence, RI: Amer. Math. Soc. MR1849347

[19] Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces*: *Isoperimetry and Processes*. Berlin: Springer.

[20] Lehec, J. (2011). Moments of the Gaussian chaos. In *Séminaire de Probabilités XLIII*. *Lecture Notes in Math.* **2006** 327–340. Berlin: Springer. MR2790379 https://doi.org/10.1007/978-3-642-15217-7_13

[21] Maurer, A. (2006). Concentration inequalities for functions of independent variables. *Random Structures Algorithms* **29** 121–138. MR2245497 https://doi.org/10.1002/rsa.20105

[22] Rudelson, M. (2008). Invertibility of random matrices: Norm of the inverse. *Ann. of Math.* (2) **168** 575–600. MR2434885 https://doi.org/10.4007/annals.2008.168.575

[23] Rudelson, M. (2012). Row products of random matrices. *Adv. Math.* **231** 3199–3231. MR2980497 https://doi.org/10.1016/j.aim.2012.08.010

[24] Rudelson, M. and Vershynin, R. (2008). The Littlewood–Offord problem and invertibility of random matrices. *Adv. Math.* **218** 600–633. MR2407948 https://doi.org/10.1016/j.aim.2008.01.010

[25] Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.* **62** 1707–1739. MR2569075 https://doi.org/10.1002/cpa.20294

[26] Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians*. *Volume III* 1576–1602. New Delhi: Hindustan Book Agency. MR2827856

[27] Rudelson, M. and Vershynin, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82. MR3125258 https://doi.org/10.1214/ECP.v18-2865

[28] Talagrand, M. (1996). A new look at independence. *Ann. Probab.* **24** 1–34. MR1387624 https://doi.org/10.1214/aop/1042644705

[29] Tao, T. (2012). *Topics in Random Matrix Theory*. *Graduate Studies in Mathematics* **132**. Providence, RI: Amer. Math. Soc. MR2906465 https://doi.org/10.1090/gsm/132

[30] Tao, T. and Vu, V. (2006). On random ± 1 matrices: Singularity and determinant. *Random Structures Algorithms* **28** 1–23. MR2187480 https://doi.org/10.1002/rsa.20109

[31] Tao, T. and Vu, V. (2010). Random matrices: The distribution of the smallest singular values. *Geom. Funct. Anal.* **20** 260–297. MR2647142 https://doi.org/10.1007/s00039-010-0057-8

[32] Tao, T. and Vu, V.H. (2009). Inverse Littlewood–Offord theorems and the condition number of random discrete matrices. *Ann. of Math.* (2) **169** 595–632. MR2480613 https://doi.org/10.4007/annals.2009.169.595

[33] Tikhomirov, K. (2015). The limit of the smallest singular value of random matrices with i.i.d. entries. *Adv. Math.* **284** 1–20. MR3391069 https://doi.org/10.1016/j.aim.2015.07.020

[34] Tikhomirov, K.E. (2016). The smallest singular value of random rectangular matrices with no moment assumptions on entries. *Israel J. Math.* **212** 289–314. MR3504328 https://doi.org/10.1007/s11856-016-1287-8

[35] Vershynin, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge: Cambridge Univ. Press. With a foreword by Sara van de Geer. MR3837109 https://doi.org/10.1017/9781108231596

[36] Wainwright, M.J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge: Cambridge Univ. Press. MR3967104 https://doi.org/10.1017/9781108627771

Received August 2019 and revised March 2020