Noname manuscript No.
(will be inserted by the editor)

Matrix Concentration for Products

De Huang · Jonathan Niles-Weed · Joel A. Tropp · Rachel Ward

2 March 2021

Abstract This paper develops nonasymptotic growth and concentration bounds for a product of independent random matrices. These results sharpen and generalize recent work of Henriksen–Ward, and they are similar in spirit to the results of Ahlswede–Winter and of Tropp for a sum of independent random matrices. The argument relies on the uniform smoothness properties of the Schatten trace classes.

Keywords Random matrices · large deviation

Mathematics Subject Classification (2010) 60B20 · 60F10 · 47B10

1 Motivation

Products of random matrices arise in many contemporary applications in the mathematics of data science. For instance, they describe the evolution of stochastic linear dynamical systems, which include popular stochastic algorithms for optimization such as Oja's algorithm for streaming principal component analysis [31] and the

Communicated by Alan Edelman

The authors gratefully acknowledge the funding for this work. DH was supported under NSF grant DMS-1613861. JNW and RW were supported in part by the Institute for Advanced Study, where some of this research was conducted. JAT was supported under ONR Awards N00014-17-1-2146 and N00014-18-1-2363. RW also received support from AFOSR MURI Award N00014-17-S-F006.

D. Huang
Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125, USA
E-mail: dhuang@caltech.edu

J. Niles-Weed
Courant Institute of Mathematical Sciences and Center for Data Science
New York University
New York, NY 10012, USA
E-mail: jnw@cims.nyu.edu

J.A. Tropp
Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125, USA
E-mail: jtropp@cms.caltech.edu

R. Ward
Oden Institute for Computational Engineering and Sciences
University of Texas
Austin, TX 78712, USA
E-mail: rward@math.utexas.edu

randomized Kaczmarz method for solving linear systems [39]. To understand the detailed behavior of these algorithms, such as the rate of convergence, we may seek out methods for studying a product of random matrices.

Unfortunately, the tools currently available in the literature are poorly adapted to these circumstances. Indeed, an instantiation of a stochastic optimization algorithm involves a finite product of finite-dimensional matrices, often with a particular structure (e.g., low-rank perturbations of the identity). But most existing theoretical results are limit laws that require the number of factors in the product or the dimension of the factors to tend to infinity. Furthermore, strong assumptions on the random matrices (e.g., independent and identically distributed entries) are usually required.

This paper offers some new tools for studying random matrix products that arise from stochastic optimization algorithms and related problems. The research is inspired by the recent paper [21] of Henriksen and Ward. Our hope is to replicate the successful program for studying sums of random matrices, implemented in the works [1, 32, 41, 42, 43, 44, 45]. In particular, we seek to develop methods that are flexible, easy to use, and powerful. We also aspire to use transparent theoretical arguments that can be adapted to new situations.

2 Contributions

To motivate our work, we start with an elementary concentration inequality for a product of independent random numbers. We will generalize this bound, and others, to the matrix setting.

2.1 Context: A Product of Random Numbers Near 1

Consider an independent family $\{X_1, X_2, \dots\} \subset \mathbb{R}$ of bounded random variables that satisfy

$$\mathbb{E}\, X_i = \mu \quad \text{and} \quad |X_i - \mu|^2 \le b^2 \quad \text{almost surely.}$$

Form a product of random perturbations of 1, and compute its mean:

$$Z_n := \prod_{i=1}^{n} \left(1 + \frac{X_i}{n}\right) \quad \text{and} \quad \mathbb{E}\, Z_n = \left(1 + \frac{\mu}{n}\right)^n = \mathrm{e}^{\mu} \cdot (1 - O(n^{-1})).$$

We anticipate that the random product Z_n concentrates around its expectation $\mathbb{E}\, Z_n \approx \mathrm{e}^{\mu}$.

To check this surmise, we can use standard methods from scalar concentration theory. For $s > 0$,

$$\begin{aligned} \mathbb{P}\{Z_n \ge (1+s)\,\mathrm{e}^{\mu}\} &= \mathbb{P}\left\{\prod_{i=1}^{n}\left(1 + \frac{X_i}{n}\right) \ge (1+s)\,\mathrm{e}^{\mu}\right\} \\ &\le \mathbb{P}\left\{\exp\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \ge (1+s)\,\mathrm{e}^{\mu}\right\} \\ &= \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}\, X_i) \ge \log(1+s)\right\}. \end{aligned}$$

The inequality follows from the numerical fact $1 + a \le \mathrm{e}^a$, valid for $a \in \mathbb{R}$. Hoeffding's inequality furnishes the bound

$$\mathbb{P}\{Z_n \ge (1+s)\,\mathrm{e}^{\mu}\} \le \exp\left(\frac{-n\log^2(1+s)}{2b^2}\right). \tag{2.1}$$

At the small scale $s \le \mathrm{e}$, in which case $\log(1+s) \ge s/\mathrm{e}$, the growth bound (2.1) implies a subgaussian tail behavior:

$$\mathbb{P}\{Z_n - \mathbb{E}\, Z_n \ge t\,\mathrm{e}^{\mu}\} \le \mathbb{P}\{Z_n - \mathrm{e}^{\mu} \ge t\,\mathrm{e}^{\mu}\} \le \exp\left(\frac{-nt^2}{2\mathrm{e}^2 b^2}\right) \quad \text{for } t \le \mathrm{e}. \tag{2.2}$$

A similar inequality holds for the lower tail.

2.2 A Product of Random Perturbations of the Identity

We might hope that products of random matrices exhibit a similar behavior. Consider an independent family $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ of $d \times d$ matrices that satisfy

$$\mathbb{E}\,\boldsymbol{X}_i = \boldsymbol{A} \quad \text{and} \quad \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\|^2 \leq b^2 \quad \text{almost surely.} \tag{2.3}$$

Here are elsewhere, $\|\cdot\|$ is the spectral norm, that is, the ℓ_2 operator norm. Form a product of random perturbations of the identity and compute its mean:

$$\boldsymbol{Z}_n = \left(\mathbf{I} + \frac{\boldsymbol{X}_n}{n}\right) \cdots \left(\mathbf{I} + \frac{\boldsymbol{X}_1}{n}\right) \quad \text{and} \quad \mathbb{E}\,\boldsymbol{Z}_n = \left(\mathbf{I} + \frac{\boldsymbol{A}}{n}\right)^n \approx \mathrm{e}^{\boldsymbol{A}}. \tag{2.4}$$

Is it true that the spectral norm $\|\boldsymbol{Z}_n\|$ is proportional to e^{μ}, where $\mu = \|\boldsymbol{A}\|$? Does the random product $\boldsymbol{Z}_n$ concentrate near its mean $\mathbb{E}\,\boldsymbol{Z}_n$?

These speculations are correct. Moreover, we can obtain bounds that parallel the scalar inequalities announced in the last subsection. Here is one particular result that follows from our analysis.

Theorem 2.1 (Products of Perturbations of the Identity—Special Case) *Consider an independent family $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ of random matrices that satisfy the hypotheses* (2.2). *Define $\mu := \|\boldsymbol{A}\|$. The matrix product $\boldsymbol{Z}_n$ introduced in* (2.2) *satisfies the bounds*

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \geq (1+s)\,\mathrm{e}^{\mu}\right\} \leq d \cdot \exp\left(\frac{-n\log^2(1+s)}{2b^2}\right) \qquad \textit{when } \log(1+s) \geq 2b^2/n;$$

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \geq t\mathrm{e}^{\mu}\right\} \leq \max\{d, \mathrm{e}\} \cdot \exp\left(\frac{-nt^2}{2\mathrm{e}^2 b^2}\right) \qquad \textit{when } t \leq \mathrm{e}.$$

Theorem 2.1 follows from Corollary 6.1.

As compared with the scalar bounds (2.1) and (2.1), the results in Theorem 2.1 feature an additional dimensional factor d in front of the exponential. This term leads to a dependency of $\log d$ in the bounds for products of random matrices. Otherwise, everything is the same, including the constants.

2.3 Proof Strategy

How might one establish a result like Theorem 2.1? The derivation in Section 2.1 is valid only for products of random scalars. We cannot even begin to make this argument for matrices because the exponential of a sum of matrices generally does not equal the product of the exponentials.

In this paper, we take a completely different approach. The key is to observe that multiplying a random product $\boldsymbol{Z} \in \mathbb{M}_d$ by a statistically independent factor $\boldsymbol{Y} \in \mathbb{M}_d$ creates a predictable change plus a random perturbation:

$$\boldsymbol{Y}\boldsymbol{Z} = (\mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z} + (\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z}.$$

Since the second term has zero mean, conditional on $\boldsymbol{Z}$, we can exploit this orthogonality property to estimate the size of the product:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Y}\boldsymbol{Z}\|_2^2 &= \mathbb{E}\,\|(\mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z}\|_2^2 + \mathbb{E}\,\|(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z}\|_2^2 \\ &\leq \left(\|\mathbb{E}\,\boldsymbol{Y}\|^2 + \mathbb{E}\,\|\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}\|^2\right)\left(\mathbb{E}\,\|\boldsymbol{Z}\|_2^2\right) =: (1+v)\,m^2 \cdot \left(\mathbb{E}\,\|\boldsymbol{Z}\|_2^2\right)\end{aligned}$$

The notation $\|\cdot\|_2$ refers to the Schatten 2-norm, also known as the Frobenius norm. The last step introduces data about the random matrix $\boldsymbol{Y}$: the mean $m = \|\mathbb{E}\,\boldsymbol{Y}\|$ and the relative variance $v = \mathbb{E}\,\|\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}\|^2 / \|\mathbb{E}\,\boldsymbol{Y}\|^2$. We can apply the same argument recursively to decompose the matrix $\boldsymbol{Z}$ into its own factors.

The approach in the last paragraph depends on the fact that $\|\cdot\|_2$ is the norm induced by the trace inner product. To undertake the same action for the spectral norm $\|\cdot\|$, we first need to approximate the spectral norm by the Schatten p-norm for $p \approx \log d$. Then we can invoke a remarkable geometric property of the

Schatten p-norm, called *uniform smoothness*, as a substitute for the orthogonality law. See the paper [29] for an introduction to this circle of ideas. Section 4 executes this method.

2.4 Additional Results

We establish a family of norm inequalities for products of random matrices. The main result, Theorem 5.1, gives a bound for the moments of a Schatten p-norm of a random product and a centered random product. From this fact, we derive expectation bounds, tail bounds, and matrix concentration inequalities. Many of these results hold under weaker assumptions than Theorem 2.1, addressing cases where the random matrices have different means, are unbounded, or form an adapted sequence.

To give a better indication of what we can prove, let us give an informal presentation of one of our main results, Corollary 5.1. The statement concerns a general product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$ of independent random matrices of dimension d. Abbreviating $p = 1 + 2\log d$, we have the inequality

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \mathrm{e}\sqrt{pv}\prod_{i=1}^{n}\|\mathbb{E}\,\boldsymbol{Y}_i\| \quad \text{when} \quad v := \sum_{i=1}^{n}\frac{\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^2}{\|\mathbb{E}\,\boldsymbol{Y}_i\|^2} \le \frac{1}{p}.$$

We can interpret v as the accumulated relative variance in the product.

For example, in the setting of Theorem 2.1, the quantity $v = O(b^2/n)$. It follows that

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| = O\left(\sqrt{\frac{b^2\log d}{n}}\,\|\mathbb{E}\,\boldsymbol{Z}_n\|\right).$$

In particular, $\|\boldsymbol{Z}_n\|$ is much closer to e^{μ} than to the worst-case bound e^{b}.

2.5 Roadmap

We continue with an overview of related work in Section 3. Section 4 presents background results from matrix theory and high-dimensional probability. We establish our main results for general matrix products in Section 5. Afterward, Section 6 draws corollaries for a product of perturbations of the identity. Finally, we describe some refinements and extensions in Section 7.

3 Related Work

Products of random matrices have been studied for decades, primarily within the fields of ergodic theory, control theory, random matrix theory, and free probability. More recently, applied mathematicians have developed results that are tailored to problems arising in data science. Almost all prior work is either asymptotic in the length of the product or asymptotic in the dimension of the matrices. This section contains an overview of these inquiries.

3.1 Direct Connections

The most immediate precedent for our research is the recent paper of Henriksen and Ward [21]. They were motivated by the problem of understanding streaming algorithms for covariance estimation. Their work gives, perhaps, the first explicit nonasymptotic bounds for a somewhat general product of random matrices with fixed dimension. The argument is based on the matrix Bernstein inequality and a combinatorial fact about set partitions.

Henriksen and Ward focus on the setting of Theorem 2.1, and they establish a bound of the form

$$\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \frac{b\mathrm{e}^b}{\sqrt{n}} \cdot \operatorname{polylog}(n, d, 1/\delta) \quad \text{with probability at least } 1-\delta. \tag{3.1}$$

In contrast, our new result reported in Theorem 2.1 implies that for $d \geq 3$ and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \leq \frac{b\mathrm{e}^{\mu}}{\sqrt{n}} \cdot \sqrt{2\mathrm{e}^2 \log(d/\delta)} \quad \text{when } 2b^2\log(d/\delta) \leq n. \tag{3.2}$$

Furthermore, if we assume that $\|\boldsymbol{X}_i\| \leq b$ almost surely for each i, then (5.5) implies that (3.1) holds without restriction. Compared with previous work, the salient improvement in (3.1) stems from the reduction of the factor e^b to e^{μ}. This difference is most pronounced when $\mathbb{E}\,\boldsymbol{X}_i = 0$ for each i, in which case the bound (3.1) removes the exponential factor entirely. Even under the assumption that $\boldsymbol{X}_i \succcurlyeq \mathbf{0}$ for all each i, it can happen that $b \geq d\mu$, so this refinement can make a big difference.

Also in the setting of Theorem 2.1, several works obtain results on the asymptotic behavior of $\boldsymbol{Z}_n$. Berger [9] establishes, via a semigroup argument based on the Chernoff product formula, that $\boldsymbol{Z}_n \longrightarrow \mathrm{e}^{\boldsymbol{A}}$ in probability as $n \longrightarrow \infty$. Emme and Hubert [14] recently obtained a refinement of this result: motivated by a problem in ergodic theory, they show that $\boldsymbol{Z}_n \longrightarrow \mathrm{e}^{\boldsymbol{A}}$ as $n \longrightarrow \infty$ under the sole assumptions that $\sum_{i=1}^n \boldsymbol{X}_i/n \longrightarrow \boldsymbol{A}$ and $\sum_{i=1}^n \|\boldsymbol{X}_i\|/n < \infty$. Their argument expands the product and computes the limit of the kth order term using an induction. We can recover a special case of their results by applying (3.1): given a triangular array $\{\boldsymbol{X}_i^{(n)} : i \leq n \text{ and } n \in \mathbb{N}\}$ of independent random matrices, form the products

$$\boldsymbol{Z}^{(n)} = \left(\mathbf{I} + \boldsymbol{X}_n^{(n)}\right) \cdots \left(\mathbf{I} + \boldsymbol{X}_1^{(n)}\right).$$

The bound (3.1), combined with the first Borel–Cantelli Lemma, guarantees that

$$\boldsymbol{Z}^{(n)} \longrightarrow \mathrm{e}^{\boldsymbol{A}} \quad \text{as } n \longrightarrow \infty, \text{ almost surely.}$$

While Emme and Hubert do not require independence, their approach does not readily yield nonasymptotic bounds. Our analysis gives a rate of convergence that matches the corresponding bound (2.1) for scalar random variables.

After the first version of our work appeared, Kathuria et al. [23] released a preprint containing an independent proof of a weaker version of Theorem 2.1. Though their bound removes several logarithmic factors from the bound (3.1) proved by Henriksen and Ward, their results still depend on e^b rather than on e^{μ}. Like us, they use a martingale decomposition of $\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n$, but they employ the matrix Freedman inequality [41] to analyze the martingale.

3.2 Other Recent Applications

Some applied work on random matrix products has been driven by the empirical observation that stochastic gradient descent converges faster when the gradient approximations are sampled *without* replacement, rather than sampled *with* replacement. Some papers that investigate this question from the point of view of (nonasymptotic) matrix inequalities include [35, 22, 2]. This specific problem has been solved by Gürbüzbalaban et al. [18] using optimization theory. However, none of these results directly address the questions at hand.

Researchers studying randomly initialized deep neural networks have also developed theoretical analysis for products of random matrices; see [19, 49]. These results involve operations on matrices with independent entries, and they focus on the large-matrix limit.

3.3 Ergodic Theory and Control Theory

Products of random matrices describe the evolution of a linear stochastic dynamical system. For this reason, they have been a subject of perennial interest within the literatures on ergodic theory and on control theory. For the most part, this research is concerned with properties of the asymptotics of infinite products of matrices (of fixed size). Let us give a few more details.

Consider a finite family $\mathscr{A} = \{\boldsymbol{A}_1, \dots, \boldsymbol{A}_s\} \subset \mathbb{M}_d$ of fixed matrices. Construct a random matrix $\boldsymbol{X} \in \mathbb{M}_d$ with the distribution

$$\mathbb{P}\{\boldsymbol{X} = \boldsymbol{A}_i\} = \frac{1}{s} \quad \text{for each } i = 1, \dots, s.$$

The *Lyapunov exponent* of the set $\mathscr{A}$ is the quantity

$$\lambda(\mathscr{A}) := \lim_{n\to\infty} \frac{1}{n} \log \|\boldsymbol{X}_n \cdots \boldsymbol{X}_1\| \quad \text{where } \boldsymbol{X}_i \sim \boldsymbol{X} \text{ iid}.$$

The Furstenberg–Kesten theorem [16] establishes that $\lambda(\mathscr{A})$ exists almost surely, but approximating $\lambda(\mathscr{A})$ is algorithmically undecidable [46, Thm. 2]. As a consequence, we must be pessimistic about finding a completely satisfactory solution to the matrix concentration problem for products.

To learn more about Lyapunov exponents and to find additional references, see the paper [3] for work in control theory and the paper [48] for work in ergodic theory. Another major application of random products is to study the asymptotic behavior of a random walk on a group; we refer the reader to [26,15,8] for more information.

3.4 Random Matrix Theory and Free Probability

Products of random matrices have also been considered within random matrix theory and free probability. This connection is natural, but matrix products have received somewhat less attention than other kinds of random matrix models. In these contexts, it is common to study a product of a small number of matrices (two or three, say) in the limit as the dimension of the matrices grows.

Bai and Silverstein [5, Chap. 4] present a limit law for the sequence of products of a random matrix with iid entries and a random matrix whose spectral distribution has a deterministic limit. This theorem is motivated by a statistical application, multivariate analysis of variance. Note, however, that convergence of the spectral distribution does not determine the limit of the spectral norm.

Free probability gives a complete description of the spectral distribution of a product of two freely independent elements as the "multiplicative free convolution" of the spectral distributions of the factors. The connection to random matrix theory stems from the fact that a family of "adequately random" matrices becomes freely independent in the limit as the dimension of the matrices tends to infinity. See the book of Nica & Speicher [30] for a digestible introduction; some other good treatments include [34,37,38]. Free probability has significant applications in wireless communications [47].

For highly structured random matrices (invariant ensembles), it may be possible to obtain more detailed formulas for products. See [24,13] for some work in this direction. Recently, Hanin and Paouris [20] developed a non-asymptotic version of the argument of Furstenberg and Kesten [16] to prove concentration bounds for the singular values of products matrices with independent Gaussian entries, though they conjecture that their results continue to hold under weaker assumptions.

4 Random Matrix Inequalities via Uniform Smoothness

To analyze products of random matrices, we exploit classic methods that were developed to study the evolution of a martingale taking values in a uniformly smooth Banach space. These ideas are relevant for us because the matrix Schatten classes (with power $2 \le p < \infty$) enjoy a remarkable uniform smoothness property.

In this section, we outline the required background from matrix analysis and high-dimensional probability. Naor's tutorial paper [29] serves as a model for our presentation; it contains a more general treatment but does not give the sharp constants. See Section 4.6 for additional discussion about the history of these ideas.

4.1 Notation and Background

We work in the complex field $\mathbb{C}$; identical results hold for the real field $\mathbb{R}$. We often use the infix notation for the minimum ($\wedge$) and the maximum ($\vee$) of two real numbers.

The operator $\mathbb{P}$ computes the probability on an event. The operator $\mathbb{E}$ computes the expectation of a random variable. Nonlinear functions, such as powers, bind before the expectation.

The linear space $\mathbb{C}^{d\times r}$ contains all $d \times r$ matrices with complex entries. The algebra $\mathbb{M}_d$ consists of all $d \times d$ matrices with complex entries. We use the standard definitions of scalar multiplication, matrix addition, matrix multiplication, and the adjoint (i.e., conjugate transpose). Any statement about matrices that is not qualified with specific dimensions holds for all matrices with compatible dimensions. Nonlinear functions, such as matrix powers, bind before the trace. The matrix absolute value $|\boldsymbol{A}| := (\boldsymbol{A}^*\boldsymbol{A})^{1/2}$, where $(\cdot)^{1/2}$ is the positive-semidefinite square root of a positive-semidefinite matrix.

We write $\|\cdot\|$ for the spectral norm on matrices; the spectral norm coincides with the maximum singular value, and it is also known as the ℓ_2 operator norm. For each $p \ge 1$, the symbol $\|\cdot\|_p$ refers to the Schatten p-norm, which returns the ℓ_p norm of the singular values of its argument. The symbol S_p refers to a linear space of matrices (of fixed dimension), equipped with the Schatten p-norm.

For parameters $p, q \ge 1$, we define the $L_q(S_p)$ norm of a random matrix $\boldsymbol{X}$ as

$$|||\boldsymbol{X}|||_{p,q} := \|\boldsymbol{X}\|_{L_q(S_p)} := \left(\mathbb{E}\, \|\boldsymbol{X}\|_p^q \right)^{1/q}.$$

The $L_q(S_p)$ norm is an operator ideal norm, in the sense that

$$|||\boldsymbol{A}\boldsymbol{X}|||_{p,q} \le \|\boldsymbol{A}\| \cdot |||\boldsymbol{X}|||_{p,q} \quad \text{for fixed } \boldsymbol{A} \text{ and random } \boldsymbol{X}. \tag{4.1}$$

This statement follows instantly from the analogous property of the Schatten p-norm.

We sometimes use the following simple inequalities for the moments of a random matrix $\boldsymbol{X}$:

$$\mathbb{E}\, \|\boldsymbol{X}\| \le \inf_{p\ge 1} \mathbb{E}\, \|\boldsymbol{X}\|_p = \inf_{p,q\ge 1} |||\boldsymbol{X}|||_{p,q}. \tag{4.2}$$

The equality follows from Lyapunov's inequality, combined with the fact that $|||\boldsymbol{X}|||_{p,1} = \mathbb{E}\, \|\boldsymbol{X}\|_p$ for all $p \ge 1$.

4.2 Uniform Smoothness for Matrices

Uniform smoothness [1] is a property of a normed space that describes how much the norm of a point changes under symmetric perturbation. Since the Schatten-2 space S_2 is an inner-product space, the parallelogram law gives an exact description of this phenomenon:

$$\frac{1}{2}\left[\|\boldsymbol{X}+\boldsymbol{Y}\|_2^2 + \|\boldsymbol{X}-\boldsymbol{Y}\|_2^2 \right] = \|\boldsymbol{X}\|_2^2 + \|\boldsymbol{Y}\|_2^2 .$$

Remarkably, in other Schatten classes, the parallelogram law is replaced by an inequality.

Fact 4.1 (Uniform Smoothness for Schatten Classes) *Let $\boldsymbol{A}, \boldsymbol{B}$ be matrices of the same size. For $p \ge 2$,*

$$\left[\frac{1}{2}\left(\|\boldsymbol{A}+\boldsymbol{B}\|_p^p + \|\boldsymbol{A}-\boldsymbol{B}\|_p^p \right) \right]^{2/p} \le \|\boldsymbol{A}\|_p^2 + \mathrm{C}_p \|\boldsymbol{B}\|_p^2 . \tag{4.3}$$

The optimal constant $\mathrm{C}_p := p - 1$. The inequality is reversed when $1 \le p \le 2$.

Fact 4.1 was first established by Tomczak-Jaegermann [40]; she obtained the sharp constant C_p when p is an even number. Ball, Carlen, and Lieb [6, Thm. 1] determined that C_p is the optimal constant for all values of p. Throughout the paper, we will continue to write $\mathrm{C}_p = p - 1$.

4.3 Uniform Smoothness for Random Matrices

Much as the Schatten class S_p of matrices enjoys a uniform smoothness property, the normed space $L_q(S_p)$ of random matrices is also uniformly smooth. When $2 \le q \le p$, this statement follows as an easy consequence of Fact 4.1.

[1] More precisely, we are considering uniformly smooth spaces whose modulus of smoothness has power type 2.

Corollary 4.1 (Uniform Smoothness for Random Matrices) *Let X, Y be random matrices of the same size. When $2 \le q \le p$,*

$$\left[\frac{1}{2}\left(|||X+Y|||_{p,q}^{q} + |||X-Y|||_{p,q}^{q}\right)\right]^{2/q} \le |||X|||_{p,q}^{2} + \mathrm{C}_p|||Y|||_{p,q}^{2}.$$

Proof Apply Lyapunov's inequality to the left-hand side of (4.1) to pass from the pth power to the qth power, and then transfer the exponent to the right-hand side to obtain the pointwise bound

$$\frac{1}{2}\left(\|X+Y\|_p^q + \|X-Y\|_p^q\right) \le \left[\|X\|_p^2 + \mathrm{C}_p\,\|Y\|_p^2\right]^{q/2}.$$

Take the expectation, and use the triangle inequality for the $L_{q/2}$ norm:

$$\frac{1}{2}\left(\mathbb{E}\,\|X+Y\|_p^q + \mathbb{E}\,\|X-Y\|_p^q\right) \le \left[\left(\mathbb{E}\,\|X\|_p^q\right)^{2/q} + \mathrm{C}_p\left(\mathbb{E}\,\|Y\|_p^q\right)^{2/q}\right]^{q/2}.$$

Reinterpret the latter display using the $L_q(S_p)$ norm $|||\cdot|||_{p,q}$.

4.4 Subquadratic Averages for Random Matrices

Corollary 4.1 admits a powerful extension that controls how the norm of a matrix changes if we add a random matrix that has zero mean. This result is the main tool that we employ in our study of random products.

Proposition 4.1 (Subquadratic Averages) *Consider random matrices X, Y of the same size that satisfy $\mathbb{E}[Y|X] = \mathbf{0}$. When $2 \le q \le p$,*

$$|||X+Y|||_{p,q}^{2} \le |||X|||_{p,q}^{2} + \mathrm{C}_p|||Y|||_{p,q}^{2}.$$

The constant $\mathrm{C}_p = p-1$ is the best possible.

Ricard and Xu [36] obtained a version of Proposition 4.1 in the more general setting of a von Neumann algebra. In their work, the expectation implicit in the L_q norm is replaced by the projection onto a subalgebra. They emphasize that the key feature of their work is the determination of the sharp constant.

Here, we offer a very short proof of Proposition 4.1 with a suboptimal constant. The method is drawn from Naor's paper [29]. Lemma A.1, in the appendix, unspools an elementary argument that delivers the sharp constant.

Proof By Jensen's inequality, applied conditionally on X,

$$\begin{aligned}\frac{1}{2}\left(|||X+Y|||_{p,q}^{2} + |||X|||_{p,q}^{2}\right) &\le \frac{1}{2}\left(|||X+Y|||_{p,q}^{2} + |||X-Y|||_{p,q}^{2}\right)\\ &\le \left[\frac{1}{2}\left(|||X+Y|||_{p,q}^{q} + |||X-Y|||_{p,q}^{q}\right)\right]^{2/q} \le |||X|||_{p,q}^{2} + \mathrm{C}_p|||Y|||_{p,q}^{2}.\end{aligned}$$

The second inequality is Lyapunov's; the third is Corollary 4.1. Upon rearranging, we find that

$$|||X+Y|||_{p,q}^{2} \le |||X|||_{p,q}^{2} + 2\mathrm{C}_p|||Y|||_{p,q}^{2}. \tag{4.4}$$

This is the stated result, with a spurious factor of 2.

4.5 Matrix-Valued Martingales

To demonstrate the value of Proposition 4.1, let us explain how it leads to moment bounds for a matrix-valued martingale sequence. Consider a null matrix martingale $\{X_1, \dots, X_n\} \subset \mathbb{M}_d$ with difference sequence $\{\Delta_1, \dots, \Delta_n\} \subset \mathbb{M}_d$. That is,

$$X_0 = \mathbf{0} \quad \text{and} \quad X_i = X_{i-1} + \Delta_i \quad \text{where} \quad \mathbb{E}[\Delta_i | X_0, \dots, X_{i-1}] = \mathbf{0} \quad \text{for } i = 1, \dots, n.$$

Applying Proposition 4.1 repeatedly, we arrive at the bound

$$|||\boldsymbol{X}_n|||_{p,q}^2 \leq \mathrm{C}_p \sum_{i=1}^{n} |||\boldsymbol{\Delta}_i|||_{p,q}^2. \tag{4.5}$$

In words, the squared norm of the martingale is controlled by the sum of the squares of the norms of the martingale differences. The inequality (4.5) is a powerful extension of the orthogonality of the increments of a martingale taking values in an inner-product space, say S_2. The uniform smoothness constant C_p shows how the geometry of the matrix space intermediates.

In this work, we will develop bounds for random matrix products by applying a similar technique to appropriately chosen decompositions of the product.

4.6 History

The approach in this section has a long history. Let us summarize the contributions that are most relevant to our development.

For real numbers, the (sharp) uniform smoothness property in Fact 4.1 is known as the *two-point inequality*; it was established independently by Leonard Gross [17] and Aline Bonami [11] in the early 1970s, with later contributions by William Beckner [7]. In 1974, the uniform smoothness property for the Schatten classes was obtained by Nicole Tomczak-Jaegermann [40]. It took another 20 years before Ball, Carlen, and Lieb [6] obtained the sharp uniform smoothness constants for all Schatten classes. The property dual to uniform smoothness is called *uniform convexity*. See [6] for a detailed exposition.

Tomczak-Jaegermann [40, Thm. 3.1] also demonstrated that Rademacher averages are subquadratic in each Schatten space S_p with $p \geq 2$; that is, the Banach space S_p has the *type 2* property [25, Chap. 9]. This fact is a prototype for the more general result stated in Proposition 4.1. Tropp [42, Sec. 4.8] points out that parts of the Ahlswede–Winter [1, App.] theory of sums of independent random matrices already follow from Tomczak-Jaegermann's work. (In contrast, Tropp's matrix concentration inequalities [42] are more closely related to a fact from operator theory, the noncommutative Khintchine inequality of Françoise Lust-Piquard [28]; Tropp's results are derived using a theorem [27, Thm. 6] of Elliot Lieb.)

Assaf Naor [29] traces the application of uniform convexity inequalities in the study of martingales to a 1975 paper of Gilles Pisier [33]. Naor [29] gives a nice introduction to this circle of ideas, which he uses to derive a general version of the Azuma inequality that holds in any uniformly smooth Banach space.

At least as early as 1988, Donald Burkholder [12] applied closely related convexity inequalities to derive sharp inequalities for martingales taking values in a Hilbert space. The paper [36] of Éric Ricard and Quanhua Xu is a recent entry in this line of research.

5 A Product of Independent Random Matrices

In this section, we obtain our main results on the growth and concentration of a product of independent random matrices. Section 5.1 shows how to decompose a random product into pieces that we can control using a recursive argument. Based on these ideas, we derive Theorem 5.1, a general bound on the moments of the norm of the matrix product. The moment estimate leads to a family of expectation bounds (Corollary 5.1) and probability bounds (Corollary 5.2).

The balance of the paper contains applications of these results (Section 6) and extensions of the method to other settings (Section 7).

5.1 Decomposition of Random Products

Our approach is based on a recursive argument that describes how the product evolves as we include more factors. At each step, we decompose the product into a nonrandom term and a random term with mean zero. This formulation allows us to apply Proposition 4.1 on subquadratic averages.

Consider a fixed matrix $\boldsymbol{Z}_0 \in \mathbb{M}_d$ and an independent family $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices. We can recursively construct products of these random matrices:

$$\boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} \quad \text{for } i = 1, \ldots, n.$$

Evidently, the last element of the sequence takes the form $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1 \boldsymbol{Z}_0$. By independence, $\mathbb{E}\, \boldsymbol{Z}_n = (\mathbb{E}\, \boldsymbol{Y}_n) \cdots (\mathbb{E}\, \boldsymbol{Y}_1) \boldsymbol{Z}_0$.

The random product $\boldsymbol{Z}_i$ admits a simple decomposition into a mean term and a fluctuation term:

$$\boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} = (\mathbb{E}\, \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} + (\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} \quad \text{for each } i = 1, \ldots, n. \tag{5.1}$$

Since $\boldsymbol{Y}_i$ is independent from $\boldsymbol{Z}_{i-1}$, the second term is conditionally zero mean:

$$\mathbb{E}\,[(\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} | \boldsymbol{Z}_{i-1}] = \boldsymbol{0}. \tag{5.2}$$

The property (5.1) supports the use of Proposition 4.1, which gives a bound on the squared norm of $\boldsymbol{Z}_i$ as a sum of the squared norms of the mean and fluctuation terms. The the norm of the mean term admits a simple bound:

$$|||(\mathbb{E}\, \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q} \le \|\mathbb{E}\, \boldsymbol{Y}_i\| \cdot |||\boldsymbol{Z}_{i-1}|||_{p,q}. \tag{5.3}$$

The inequality follows from the operator ideal property of the Schatten p-norm. We likewise have an explicit bound on the norm of the random fluctuation term:

$$|||(\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q} \le \left(\mathbb{E}\, \|\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i\|^q \cdot \mathbb{E}\, \|\boldsymbol{Z}_{i-1}\|_p^q\right)^{1/q} = (\mathbb{E}\, \|\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i\|^q)^{1/q}\, |||\boldsymbol{Z}_{i-1}|||_{p,q}. \tag{5.4}$$

The first relation follows from the operator ideal property and the statistical independence of the random matrices $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_{i-1}$. Combining (5.1) and (5.1) gives a recursive bound on $|||\boldsymbol{Z}_i|||_{p,q}$ in terms of $|||\boldsymbol{Z}_{i-1}|||_{p,q}$.

We can study the concentration properties of the product $\boldsymbol{Z}_i$ using a related decomposition:

$$\boldsymbol{Z}_i - \mathbb{E}\, \boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} - (\mathbb{E}\, \boldsymbol{Y}_i)(\mathbb{E}\, \boldsymbol{Z}_{i-1}) = (\mathbb{E}\, \boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \mathbb{E}\, \boldsymbol{Z}_{i-1}) + (\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}. \tag{5.5}$$

As in (5.1), the second term is a fluctuation that is conditionally zero mean, and applying Proposition 4.1 once again gives a suitable recursive bound.

We carry out the details of these arguments in Theorem 5.1.

5.2 Growth and Concentration

Our main result controls the growth of the moments of a product of independent random matrices. It also describes how well the random product concentrates around its expectation.

Theorem 5.1 (Growth and Concentration of Random Products) *Consider a fixed matrix $\boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}$ and an independent family $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices. Form the product*

$$\boldsymbol{Z}_n = \boldsymbol{Y}_n \boldsymbol{Y}_{n-1} \cdots \boldsymbol{Y}_2 \boldsymbol{Y}_1 \boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}.$$

For parameters $2 \le q \le p$, assume that

$$\|\mathbb{E}\, \boldsymbol{Y}_i\| \le m_i \quad \textit{and} \quad \left(\mathbb{E}\, \|\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i\|^q\right)^{1/q} \le \sigma_i m_i \quad \textit{for } i = 1, \ldots, n.$$

Define the product of means and the accumulated relative fluctuation

$$M = \prod_{i=1}^n m_i \quad \textit{and} \quad v = \sum_{i=1}^n \sigma_i^2.$$

Then the random product Z_n satisfies the growth bound and the concentration bound

$$|||Z_n|||_{p,q} \le \mathrm{e}^{C_p v/2} \left\|Z_0\right\|_p \cdot M; \tag{5.6}$$

$$|||Z_n - \mathbb{E}\, Z_n|||_{p,q} \le \left(\mathrm{e}^{C_p v} - 1\right)^{1/2} \left\|Z_0\right\|_p \cdot M. \tag{5.7}$$

Proof (Proof of Theorem 5.1, relation (5.1)*)* By the homogeneity of (5.1), we may assume that $m_i = 1$ for each index i, so that also $M = 1$. As in (5.1), we have the decomposition

$$Z_i := Y_i Z_{i-1} = (\mathbb{E}\, Y_i) Z_{i-1} + (Y_i - \mathbb{E}\, Y_i) Z_{i-1} \quad \text{for each } i = 1, \dots, n.$$

Now, Proposition 4.1 implies that

$$\begin{aligned} |||Z_i|||_{p,q}^2 &\le |||(\mathbb{E}\, Y_i) Z_{i-1}|||_{p,q}^2 + C_p \cdot |||(Y_i - \mathbb{E}\, Y_i) Z_{i-1}|||_{p,q}^2 \\ &\le \|\mathbb{E}\, Y_i\|^2 \cdot |||Z_{i-1}|||_{p,q}^2 + C_p \left(\mathbb{E}\, \|Y_i - \mathbb{E}\, Y_i\|^q\right)^{2/q} \cdot |||Z_{i-1}|||_{p,q}^2 \\ &\le (1 + C_p \sigma_i^2) \cdot |||Z_{i-1}|||_{p,q}^2 \\ &\le \exp(C_p \sigma_i^2) \cdot |||Z_{i-1}|||_{p,q}^2. \end{aligned}$$

The second line follows from (5.1), and the third depends on our hypotheses about the factors Y_i. The last relation requires the numerical inequality $1 + a \le \mathrm{e}^a$, valid for all $a \in \mathbb{R}$. By iteration,

$$|||Z_i|||_{p,q}^2 \le \exp\left(C_p \sum_{k=1}^{i} \sigma_k^2\right) \cdot \left\|Z_0\right\|_p^2. \tag{5.8}$$

In the final step, we use the assumption that Z_0 is not random to see that $|||Z_0|||_{p,q} = \|Z_0\|_p$. For $i = n$, the formula (5.2) is the advertised result.

Proof (Proof of Theorem 5.1, relation (5.1)*)* The pattern of argument is similar with the proof of (5.1). By the homogeneity of (5.1), we may assume that all $m_i = 1$ and that $M = 1$. As in (5.1), we have the decomposition

$$Z_i - \mathbb{E}\, Z_i = Y_i Z_{i-1} - (\mathbb{E}\, Y_i)(\mathbb{E}\, Z_{i-1}) = (\mathbb{E}\, Y_i)(Z_{i-1} - \mathbb{E}\, Z_{i-1}) + (Y_i - \mathbb{E}\, Y_i) Z_{i-1}.$$

Again, we invoke Proposition 4.1 to ascertain that

$$\begin{aligned} |||Z_i - \mathbb{E}\, Z_i|||_{p,q}^2 &\le |||(\mathbb{E}\, Y_i)(Z_{i-1} - \mathbb{E}\, Z_{i-1})|||_{p,q}^2 + C_p \cdot |||(Y_i - \mathbb{E}\, Y_i) Z_{i-1}|||_{p,q}^2 \\ &\le |||Z_{i-1} - \mathbb{E}\, Z_{i-1}|||_{p,q}^2 + C_p \sigma_i^2 \cdot |||Z_{i-1}|||_{p,q}^2 \\ &\le |||Z_{i-1} - \mathbb{E}\, Z_{i-1}|||_{p,q}^2 + C_p \sigma_i^2 \exp\left(\sum_{k=1}^{i-1} C_p \sigma_i^2\right) \cdot \left\|Z_0\right\|_p^2. \end{aligned}$$

The last inequality is our growth bound (5.2). This recurrence relation delivers

$$\begin{aligned} |||Z_n - \mathbb{E}\, Z_n|||_{p,q}^2 &\le |||Z_0 - \mathbb{E}\, Z_0|||_{p,q}^2 + \left[\sum_{i=1}^{n} C_p \sigma_i^2 \exp\left(\sum_{k=1}^{i-1} C_p \sigma_k^2\right)\right] \cdot \left\|Z_0\right\|_p^2 \\ &= \left[\sum_{i=1}^{n} C_p \sigma_i^2 \exp\left(\sum_{k=1}^{i-1} C_p \sigma_k^2\right)\right] \cdot \left\|Z_0\right\|_p^2 \\ &\le \left[\exp\left(\sum_{i=1}^{n} C_p \sigma_i^2\right) - 1\right] \cdot \left\|Z_0\right\|_p^2. \end{aligned}$$

The equality holds because Z_0 is not random. The last relation is a numerical inequality, whose proof appears in Lemma A.2.

Observe that the difference between the bounds (5.1) and (5.1) is only visible when $\mathrm{C}_p v$ is small, in which case

$$\mathrm{e}^{\mathrm{C}_p v/2} \approx 1 \quad \text{and} \quad \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \approx \sqrt{\mathrm{C}_p v}.$$

This is the setting where the concentration result may be nontrivial.

In applications of Theorem 5.1, we often take p large enough that the term $\|\boldsymbol{Z}_0\|_p$ is a constant. For instance, when $\boldsymbol{Z}_0 = \mathbf{I}$, it suffices to take $p \approx \log d$. The most natural choices for q are $q = 2$ and $q = p$. The former is appropriate when only bounds on the variance $\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^2$ are available, whereas the latter applies when we assume an almost-sure bound on the fluctuations of $\boldsymbol{Y}_i$.

The next remark contains a minor extension of Theorem 5.1. Similar extensions are possible at other points in this paper. For the most part, we omit these developments.

Remark 5.1 (Growth from Concentration) In some instances, we can improve over the growth bound (5.1) by applying the triangle inequality to the decomposition $\boldsymbol{Z}_n = (\mathbb{E}\,\boldsymbol{Z}_n) + (\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n)$ and invoking the concentration bound (5.1):

$$|||\boldsymbol{Z}_n|||_{p,q} \le \|\mathbb{E}\,\boldsymbol{Z}_n\|_p + \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \|\boldsymbol{Z}_0\|_p \cdot M.$$

Similarly, we can apply Proposition 4.1 together with (5.1) to obtain

$$|||\boldsymbol{Z}_n|||_{p,q}^2 \le \|\mathbb{E}\,\boldsymbol{Z}_n\|_p^2 + \mathrm{C}_p \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right) \|\boldsymbol{Z}_0\|_p^2 \cdot M^2.$$

Neither of these bounds represents a strict improvement over the other or over the growth bound (5.1).

5.3 Expectation Bounds for the Spectral Norm

In many cases, we just need to know the expected value of the product $\|\boldsymbol{Z}_n\|$ or the expected value of the fluctuation $\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\|$. We can obtain bounds for these quantities as an easy consequence of Theorem 5.1.

Corollary 5.1 (Expectation Bounds) *Consider an independent sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices, and form the product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. Assume that*

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le m_i \quad \text{and} \quad \left(\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^2\right)^{1/2} \le \sigma_i m_i \quad \text{for } i = 1, \dots, n.$$

Let $M = \prod_{i=1}^n m_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \exp\left(\sqrt{2v\,(2v \vee \log d)}\right) \cdot M. \tag{5.9}$$

Provided that $v\,(1 + 2\log d) \le 1$, then also

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \sqrt{\mathrm{e}^2 v\,(1 + 2\log d)} \cdot M. \tag{5.10}$$

Proof To apply Theorem 5.1, we set $\boldsymbol{Z}_0 = \mathbf{I}$ and choose the power $q = 2$.

To obtain the growth bound (5.1), consider the Schatten norm of order $p = \sqrt{2(2v \vee \log d)/v}$. Note that $p \ge 2$ and that $\|\boldsymbol{Z}_0\|_p \le d^{1/p} \le \mathrm{e}^{pv/2}$. Invoke Theorem 5.1, relation (5.1), to see that

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le |||\boldsymbol{Z}_n|||_{p,2} \le \mathrm{e}^{\mathrm{C}_p v/2} \|\boldsymbol{Z}_0\|_p \cdot M \le \mathrm{e}^{pv/2} \cdot \mathrm{e}^{pv/2} \cdot M = \mathrm{e}^{pv} \cdot M.$$

We used the fact that $\mathrm{C}_p = p - 1 < p$. This is the stated result.

To obtain the concentration bound (5.1), consider the Schatten norm $p = 2(1 + \log d)$. Note that $p \ge 2$ and that $\|\boldsymbol{Z}_0\|_p \le d^{1/p} \le \sqrt{\mathrm{e}}$. Now, we use Theorem 5.1, relation (5.1), in a similar fashion. Assuming that $\mathrm{C}_p v \le 1$,

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le |||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,2} \le \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \|\boldsymbol{Z}_0\|_p \cdot M \le \mathrm{e}\sqrt{\mathrm{C}_p v} \cdot M.$$

The last bound is the numerical inequality $\mathrm{e}^a - 1 \le \mathrm{e}a$, valid when $a \in [0, 1]$. Finally, note that $\mathrm{C}_p = p - 1 = 1 + 2\log d$.

The inequality (5.1) shows its power when each σ_i is small. Assume that each $m_i = 1$ and $\sigma_i \le b/n$ for a constant b. If we assume that $\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| \le \sigma_i m_i$ almost surely, then it is not hard to check that

$$\|\mathbb{E}\,\boldsymbol{Z}_n\| \le 1 \quad \text{while} \quad \|\boldsymbol{Z}_n\| \le (1 + (b/n))^n \le \mathrm{e}^b \quad \text{almost surely.}$$

If $b\sqrt{(2\log d)/n}$ is close to zero, then (5.1) implies

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \mathrm{e}^{b\sqrt{(2\log d)/n}} \approx 1.$$

That is, $\mathbb{E}\,\|\boldsymbol{Z}_n\|$ is much closer to $\|\mathbb{E}\,\boldsymbol{Z}_n\|$ than to the worst-case value e^b.

5.4 Tail Bounds for the Spectral Norm

The moment bounds in Theorem 5.1 can also be upgraded to obtain tail bounds for $\|\boldsymbol{Z}_n\|$ and $\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\|$.

Corollary 5.2 (Tail Bounds) *Consider an independent sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices, and form the product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. Assume that*

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le m_i \quad \textit{and} \quad \|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| \le \sigma_i m_i \quad \textit{almost surely for } i = 1, \dots, n.$$

Let $M = \prod_{i=1}^n m_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \ge tM\right\} \le d \cdot \exp\left(\frac{-\log^2 t}{2v}\right) \quad \textit{when } \log t \ge 2v. \tag{5.11}$$

Furthermore,

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \ge tM\right\} \le (d \vee \mathrm{e}) \cdot \exp\left(\frac{-t^2}{2\mathrm{e}^2 v}\right) \quad \textit{when } t \le \mathrm{e}. \tag{5.12}$$

Proof We begin with the proof of (5.2). By homogeneity, we may assume that $m_i = 1$ for each i, so also $M = 1$. Apply Markov's inequality and (4.1) to obtain

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \ge t\right\} \le \inf_{p \ge 2} t^{-p} \cdot \mathbb{E}\,\|\boldsymbol{Z}_n\|^p \le \inf_{p \ge 2} t^{-p} \cdot \|\!|\boldsymbol{Z}_n|\!\|_{p,p}^p.$$

To bound the $L_p(S_p)$ norm, we will use Theorem 5.1 with $\boldsymbol{Z}_0 = \mathbf{I}$ and with $q = p$. Relation (5.1) gives

$$t^{-p} \cdot \|\!|\boldsymbol{Z}_n|\!\|_{p,p}^p \le t^{-p} \cdot \mathrm{e}^{p\mathrm{C}_p v/2} \, \|\boldsymbol{Z}_0\|_p^p = d \cdot \left(t^{-2} \mathrm{e}^{\mathrm{C}_p v}\right)^{p/2}.$$

We have used the fact that $\|\boldsymbol{Z}_0\|_p^p = \|\mathbf{I}\|_p^p = d$. Under the assumption that $\log t \ge 2v$, we may select $p = (\log t)/v \ge 2$. This choice yields

$$d \cdot \left(t^{-2} \mathrm{e}^{pv}\right)^{p/2} = d \cdot \exp\left(\frac{-\log^2 t}{2v}\right).$$

Sequence the last three displays to arrive at the bound (5.2).

We establish (5.2) in an analogous fashion. The same argument, using relation (5.1), implies that

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \ge t\right\} \le \inf_{p \ge 2} d \cdot \left[t^{-2}\left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)\right]^{p/2}.$$

Supposing that $t^2/(\mathrm{e}^2 v) < 2$, the bound (5.2) holds trivially because $\mathrm{e} \cdot \exp(-t^2/(2\mathrm{e}^2 v)) \ge 1$. Otherwise, we may select the parameter $p = t^2/(\mathrm{e}^2 v) \ge 2$. Under the assumption that $t \le \mathrm{e}$, $\mathrm{C}_p v \le pv \le (t/\mathrm{e})^2 \le 1$, so that

$e^{C_p v} - 1 \le eC_p v \le t^2/e$. Therefore,

$$d \cdot \left[t^{-2}\left(e^{C_p v} - 1\right)\right]^{p/2} \le d \cdot e^{-p/2} = d \cdot \exp\left(\frac{-t^2}{2e^2 v}\right).$$

The last two displays imply (5.2).

Let us demonstrate the optimality of our tail bounds in Corollary 5.2 by showing that they are tight in the commutative case. Consider a simple scenario where all the factors of the product $\boldsymbol{Z}_n$ are diagonal matrices:

$$\boldsymbol{Y}_i = \operatorname{diag}(y_i^{(1)}, y_i^{(2)}, \cdots, y_i^{(d)}), \quad i = 1, 2, \ldots, n.$$

The scalar random variables $y_i^{(j)}$, $1 \le i \le n$, $1 \le j \le d$ are independent of each other and satisfy

$$\mathbb{E}\, y_i^{(j)} = m_i \ge 0 \quad \text{and} \quad |y_i^{(j)} - m_i| \le \sigma_i m_i \quad \text{almost surely for } 1 \le i \le n, 1 \le j \le d.$$

Fix $\boldsymbol{Z}_0 = \mathbf{I}$. It follows that

$$\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1 = \operatorname{diag}\left(\prod_{i=1}^n y_i^{(1)}, \cdots, \prod_{i=1}^n y_i^{(d)}\right).$$

For each diagonal entry of $\boldsymbol{Z}_n$, a calculation similar to that in Section 2.1 gives

$$\begin{aligned}\mathbb{P}\left\{\prod_{i=1}^n y_i^{(j)} \ge (1+s)M\right\} &= \mathbb{P}\left\{\prod_{i=1}^n \left(1 + \frac{y_i^{(j)} - m_i}{m_i}\right) \ge 1 + s\right\} \\ &\le \mathbb{P}\left\{\exp\left(\sum_{i=1}^n \frac{y_i^{(j)} - m_i}{m_i}\right) \ge 1 + s\right\} = \mathbb{P}\left\{\sum_{i=1}^n \frac{y_i^{(j)} - m_i}{m_i} \ge \log(1+s)\right\}.\end{aligned}$$

As usual, $M = \prod_{i=1}^n m_i$. The inequality follows from the numerical fact $1 + a \le e^a$ for all $a \in \mathbb{R}$. We then use Hoeffding's inequality to obtain

$$\mathbb{P}\left\{\prod_{i=1}^n y_i^{(j)} \ge (1+s)M\right\} \le \exp\left(\frac{-\log^2(1+s)}{2v}\right),$$

where $v = \sum_{i=1}^n \sigma_i^2$. At the small scale $s \le e$, in which case $\log(1+s) \ge s/e$, the above bound implies

$$\mathbb{P}\left\{\prod_{i=1}^n y_i^{(j)} - \prod_{i=1}^n \mathbb{E}\, y_i^{(j)} \ge tM\right\} \le \exp\left(\frac{-t^2}{2e^2 v}\right),$$

Taking a uniform bound over the d diagonal entries of $\boldsymbol{Z}_n$ yields

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Z}_n) \ge (1+s)M\right\} = \mathbb{P}\left\{\exists\, i : \prod_{i=1}^n y_i^{(j)} \ge (1+s)M\right\} \le d \cdot \exp\left(\frac{-\log^2(1+s)}{2v}\right); \tag{5.13}$$

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n) \ge tM\right\} = \mathbb{P}\left\{\exists\, i : \prod_{i=1}^n y_i^{(j)} - M \ge tM\right\} \le d \cdot \exp\left(\frac{-t^2}{2e^2 v}\right) \quad \text{for } t \le e. \tag{5.14}$$

Moreover, these bounds are easily seen to be essentially tight.

To assess our results, we apply Corollary 5.2 to see what it predicts for this model. The assumptions translate to the matrix bounds

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| = \|m_i \mathbf{I}\| = m_i \quad \text{and} \quad \|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| \le \sigma_i m_i \quad \text{almost surely for } 1 \le i \le n.$$

Therefore, Corollary 5.2 implies that

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Z}_n) \geq (1+s)M\right\} \leq d \cdot \exp\left(\frac{-\log^2(1+s)}{2v}\right) \quad \text{for } \log(1+s) \geq 2v; \tag{5.15}$$

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n) \geq tM\right\} \leq d \cdot \exp\left(\frac{-t^2}{2\mathrm{e}^2 v}\right) \quad \text{for } t \leq \mathrm{e}. \tag{5.16}$$

By comparing (5.4) and (5.4) with (5.4) and (5.4), we see that our estimates obtained from the uniform smoothness argument are tight in the commutative scenario, except that we require $\log(1+s) \geq 2v$ for the growth bound (5.4).

5.5 Uniform Bounds on Factors

In some circumstances, it is reasonable to assume that the factors are bounded in norm almost surely. For example, the randomized Kaczmarz algorithm [39] can be expressed as the repeated application of random *contractions*, that is, matrices whose singular values are bounded by one. Other randomized linear fixed-point iterations take a similar form.

A modification of the proof of Theorem 5.1 offers some potential improvements in this setting. Fix parameters $2 \leq q \leq p$. Suppose that $\|\boldsymbol{Y}_i\| \leq b_i$ almost surely and $|||\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|||_{p,q} \leq \sigma_i b_i$ for each index i. Define $B = \prod_{i=1}^n b_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$|||\boldsymbol{Z}_n|||_{p,q} \leq \|\boldsymbol{Z}_0\|_p \cdot B; \tag{5.17}$$

$$|||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,q} \leq \sqrt{\mathrm{C}_p v}\;\|\boldsymbol{Z}_0\|_p \cdot B. \tag{5.18}$$

The growth bound (5.5) is an immediate consequence of the definition $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1 \boldsymbol{Z}_0$. The concentration result (5.5) follows if we repeat the proof of (5.1), using homogeneity to assume that $b_i = 1$ for each i and employing the growth bound (5.5) in place of (5.1).

These bounds yield strengthenings of Corollaries 5.1 and 5.2. If we assume $|||\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|||_{p,2} \leq \sigma_i b_i$ for each i, then applying the argument of Section 5.5 with (5.5) implies that

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \leq \sqrt{\mathrm{e}v\,(1 + 2\log d)}B. \tag{5.19}$$

This improves the constant in (5.1) by a factor of $\sqrt{\mathrm{e}}$, and it removes the condition that $v\,(1+2\log d) \leq 1$.

Similarly, if we assume $\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| \leq \sigma_i b_i$ almost surely, then we obtain an unconditional variant of the concentration bound (5.2):

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \geq t \cdot B\right\} \leq (d \vee \mathrm{e}) \cdot \exp\left(\frac{-t^2}{2\mathrm{e}v}\right) \quad \text{for all } t > 0. \tag{5.20}$$

Both (5.5) and (5.5) scale with $B = \prod_{i=1}^n b_i$. In the important special case where each factor in the product is a random contraction, we may take $b_i = 1$ for each i, so that $B = 1$. This bound may be pessimistic, as the next theorem shows.

Theorem 5.2 (Product of Random Contractions) *Consider an independent family $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random contractions; that is, $\|\boldsymbol{Y}_i\| \leq 1$. Form the random product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. For a parameter $p \geq 2$, assume that*

$$\|\mathbb{E}\,|\boldsymbol{Y}_i|^p\|^{1/p} \leq \overline{m}_i \leq 1 \quad \textit{and} \quad \|\mathbb{E}\,|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|^p\|^{1/p} \leq \sigma_i \overline{m}_i \quad \textit{for } i = 1, \dots, n.$$

Define $\overline{M} := \prod_{i=1}^n \overline{m}_i$ and $v := \sum_{i=1}^n \sigma_i^2$. Then

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \leq 1 \wedge (d^{1/p} \cdot \overline{M}); \tag{5.21}$$

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \leq \sqrt{C_p v}\, d^{1/p} \cdot \overline{M}. \tag{5.22}$$

Unlike (5.5), (5.5) and (5.5), the bounds of Theorem 5.2 scale with $\overline{M}$. This quantity can be substantially smaller than B, particularly if each factor has low rank. For example, if $\boldsymbol{Y}_i$ is a projection to a uniformly random rank-one subspace, then $\|\boldsymbol{Y}_i\| = 1$ but $\left\|\mathbb{E}\,|\boldsymbol{Y}_i|^2\right\| = 1/d$.

It is also possible to prove a tail bound, assuming an almost-sure bound on the spectral norm of the fluctuations; we omit the details. For convenience, we have focused on products of random contractions, but a homogeneity argument implies that similar inequalities hold for any product of uniformly bounded random matrices.

To prove Theorem 5.2, we require a lemma that isolates the influence of each factor in the product. This result gives an improvement in the setting where the factor is a contraction, but it may give inferior results in other settings.

Lemma 5.1 (Random Product: Absolute Values) *Let $\boldsymbol{Y} \in \mathbb{M}_d$ be a random matrix, and let $\boldsymbol{Z} \in \mathbb{M}_d$ be a random matrix that is independent from $\boldsymbol{Y}$. For $2 \le q \le p$,*

$$|||\boldsymbol{YZ}|||_{p,q} \le \|\mathbb{E}\,|\boldsymbol{Y}|^p\|^{1/p} \cdot |||\boldsymbol{Z}|||_{p,q}.$$

Proof Write out the $L_q(S_p)$ norm, and introduce matrix absolute values:

$$|||\boldsymbol{YZ}|||_{p,q}^q = \mathbb{E}\,\|\boldsymbol{YZ}\|_p^q = \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{Z}^*\boldsymbol{Y}^*\boldsymbol{YZ}\right)^{p/2}\right]^{q/p} = \mathbb{E}\left[\operatorname{tr}\left(\,|\boldsymbol{Z}^*|\cdot|\boldsymbol{Y}|^2\cdot|\boldsymbol{Z}^*|\,\right)^{p/2}\right]^{q/p}.$$

The last relation can be verified using polar factorizations. Apply the Araki–Lieb–Thirring inequality [10, Thm. IX.2.20] to distribute the power onto the factors in the trace. Conditioning on the random matrix $\boldsymbol{Z}$, we obtain

$$\begin{aligned} |||\boldsymbol{YZ}|||_{p,q}^q &\le \mathbb{E}\,\mathbb{E}\left[\left[\operatorname{tr}\left(\,|\boldsymbol{Z}^*|^{p/2}\cdot|\boldsymbol{Y}|^p\cdot|\boldsymbol{Z}^*|^{p/2}\right)\right]^{q/p}\,\middle|\,\boldsymbol{Z}\right] \\ &\le \mathbb{E}\left[\operatorname{tr}\left(\,|\boldsymbol{Z}^*|^{p/2}\cdot\left(\mathbb{E}\,|\boldsymbol{Y}|^p\right)\cdot|\boldsymbol{Z}^*|^{p/2}\right)\right]^{q/p}. \end{aligned}$$

The inequality is Jensen's, which is justified because $q/p \le 1$. Bounding the matrix in the center by its norm,

$$|||\boldsymbol{YZ}|||_{p,q}^q \le \|\mathbb{E}\,|\boldsymbol{Y}|^p\|^{q/p}\cdot\mathbb{E}\left[\operatorname{tr}|\boldsymbol{Z}^*|^p\right]^{q/p} = \|\mathbb{E}\,|\boldsymbol{Y}|^p\|^{q/p}\cdot|||\boldsymbol{Z}|||_{p,q}^q.$$

Take the qth root to complete the proof.

With this result at hand, Theorem 5.2 follows from familiar arguments.

Proof (Proof of Theorem 5.2) Define $\boldsymbol{Z}_0 = \mathbf{I}$ and $\boldsymbol{Z}_i = \boldsymbol{Y}_i\boldsymbol{Z}_{i-1}$ for each index $i = 1, \dots, n$. We begin with the proof of (5.2). Since each factor is a contraction, it is clear that

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \mathbb{E}\prod_{k=1}^n \|\boldsymbol{Y}_k\| \le 1.$$

To obtain a less trivial bound on the expectation, we apply Lemma 5.1 repeatedly. For $p \ge 2$,

$$\mathbb{E}\,\|\boldsymbol{Z}_i\| \le |||\boldsymbol{Z}_i|||_{p,p} \le \prod_{k=1}^i \|\mathbb{E}\,|\boldsymbol{Y}_k|^p\|^{1/p}\cdot|||\mathbf{I}|||_{p,p} = d^{1/p}\prod_{k=1}^i \overline{m}_k. \tag{5.23}$$

The statement (5.2) combines these two observations for the choice $i = n$.

Let us continue with the proof of (5.2), which is analogous to the argument in Theorem 5.1(5.1). First, by expanding the inequality $\mathbb{E}\,|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|^2 \succcurlyeq \mathbf{0}$, we see that $\mathbf{0} \preccurlyeq |\mathbb{E}\,\boldsymbol{Y}_i|^2 \preccurlyeq \mathbb{E}\,|\boldsymbol{Y}_i|^2$. As a consequence, for $p \ge 2$,

$$\|\mathbb{E}\,\boldsymbol{Y}_i\|^2 \le \left\|\mathbb{E}\,|\boldsymbol{Y}_i|^2\right\| \le \|\mathbb{E}\,|\boldsymbol{Y}_i|^p\|^{2/p} \le \overline{m}_i^2.$$

The second inequality follows from a matrix extension of Lyapunov's inequality [4, Cor. 4.2(i)]. Now, we may calculate that

$$\begin{aligned}|||\boldsymbol{Z}_i - \mathbb{E}\,\boldsymbol{Z}_i|||_{p,p}^2 &\le |||(\mathbb{E}\,\boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1})|||_{p,p}^2 + \mathrm{C}_p \cdot |||(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,p}^2 \\ &\le \overline{m}_i^2 \cdot |||\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}|||_{p,p}^2 + \mathrm{C}_p\sigma_i^2\overline{m}_i^2 \cdot |||\boldsymbol{Z}_{i-1}|||_{p,p}^2 \\ &\le \overline{m}_i^2 \cdot |||\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}|||_{p,p}^2 + \mathrm{C}_p\sigma_i^2 \cdot d^{2/p}\prod_{k=1}^{i}\overline{m}_k^2.\end{aligned}$$

The second inequality uses (5.1), Lemma 5.1, and the last display. The third inequality requires the preliminary bound (5.5). Unrolling the recursion,

$$|||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,p}^2 \le \mathrm{C}_p d^{2/p}\left(\prod_{i=1}^{n}\overline{m}_i^2\right)\left(\sum_{i=1}^{n}\sigma_i^2\right) = \mathrm{C}_p d^{2/p}\overline{M}^2 v.$$

This result implies the advertised bound (5.2).

6 Application: Random Perturbations of the Identity

This section treats the fundamental case where the factors $\boldsymbol{Y}_i$ in the product are independent, random perturbations of the identity. That is, $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i$ where $\{\boldsymbol{X}_i\} \subset \mathbb{M}_d$ is an independent family. Applying our main theorems in this setting yields the promised improvements to the tail bounds developed by Henriksen–Ward [21].

6.1 Iterative Algorithms

To motivate this development, observe that random perturbations of the identity arise from the analysis of the iterative scheme

$$\boldsymbol{u}^{(i+1)} = \boldsymbol{u}^{(i)} + \boldsymbol{X}_i\boldsymbol{u}^{(i)} \quad \text{for } i = 1, 2, 3 \dots. \tag{6.1}$$

where $\boldsymbol{X}_i\boldsymbol{u}^{(i)}$ is a linear update to the current iterate $\boldsymbol{u}^{(i)}$. In this application, the norm of each $\boldsymbol{X}_i$ is proportional to the step size of the scheme, so it is typically small and it is controlled by the user. For example, the updates in Oja's algorithm [31] take the form (6.1).

For now, we do not permit the random matrix $\boldsymbol{X}_i$ to depend on the sequence $\{\boldsymbol{u}^{(i)}\}$ of iterates. Later, in Section 7.2, we describe an extension of our approach to the setting where $\{\boldsymbol{X}_i\}$ is an adapted sequence. This variant allows for the study of a wider class of iterative algorithms.

6.2 Bounds for the Product

First, we develop bounds for the growth and concentration of a product of perturbations of the identity. In Section 6.3, we develop results for the inverse of the product.

Corollary 6.1 (Perturbations of the Identity) *Consider an independent family $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ of random matrices, and form the product $\boldsymbol{Z}_n = (\mathbf{I} + \boldsymbol{X}_n)\cdots(\mathbf{I} + \boldsymbol{X}_1)$. Assume that*

$$\|\mathbb{E}\,\boldsymbol{X}_i\| \le \xi_i \quad \textit{and} \quad \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \le \sigma_i \quad \textit{almost surely for } i = 1, \dots, n.$$

Define $\xi = \sum_{i=1}^n \xi_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}_n\| &\le \exp\left(\xi + \sqrt{2v\log d}\right) && \textit{when } 2v \le \log d; \\ \mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| &\le \mathrm{e}^{\xi+1}\sqrt{v\,(1 + 2\log d)} && \textit{when } v\,(1 + 2\log d) \le 1.\end{aligned}$$

Moreover,

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \ge t\mathrm{e}^{\xi}\right\} \le d \cdot \exp\left(\frac{-\log^2 t}{2v}\right) \qquad \textit{when } \log t \ge 2v;$$

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \ge t\mathrm{e}^{\xi}\right\} \le (d \vee \mathrm{e}) \cdot \exp\left(\frac{-t^2}{2\mathrm{e}^2 v}\right) \qquad \textit{when } t \le \mathrm{e}.$$

Proof Let $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i$ for each index i. Then

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le 1 + \|\mathbb{E}\,\boldsymbol{X}_i\| \le \mathrm{e}^{\xi_i} =: m_i.$$

Furthermore, since $m_i \ge 1$,

$$\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| = \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \le \sigma_i \le \sigma_i m_i.$$

The results follow instantly from Corollary 5.1 and Corollary 5.2.

6.3 Bounds for the Inverse of a Product

In some applications, it is valuable to have a lower bound for the minimum singular value of a random product. Equivalently, we can seek an upper bound for the spectral norm of the inverse of the product. This section describes a situation where clean results are possible.

Consider the case where the factors $\boldsymbol{Y}_i$ are perturbations of the identity: $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i$, where $\boldsymbol{X}_i$ is small enough to ensure that $\boldsymbol{Y}_i$ is invertible with probability 1. In this setting, we can easily study the inverse of the product using Corollary 6.1.

Corollary 6.2 (Perturbations of the Identity: Inverses) *Frame the same hypotheses as in Corollary 6.1. Assume that $\xi_i + \sigma_i < 1$ for each index i, and define*

$$\bar{\xi} = \sum_{i=1}^{n}\left[\xi_i + \frac{(\xi_i+\sigma_i)^2}{1-(\xi_i+\sigma_i)}\right] \quad \textit{and} \quad \bar{v} = \sum_{i=1}^{n}\left[\sigma_i + \frac{2(\xi_i+\sigma_i)^2}{1-(\xi_i+\sigma_i)}\right]^2.$$

Then

$$\mathbb{E}\left\|\boldsymbol{Z}_n^{-1}\right\| \le \exp\left(\bar{\xi} + \sqrt{2\bar{v}\log d}\right) \qquad \textit{when } 2\bar{v} \le \log d;$$

$$\mathbb{E}\left\|\boldsymbol{Z}_n^{-1} - \mathbb{E}\,\boldsymbol{Z}_n^{-1}\right\| \le \mathrm{e}^{\bar{\xi}}\sqrt{\mathrm{e}^2\bar{v}\,(1+2\log d)} \qquad \textit{when } \bar{v}\,(1+2\log d) \le 1.$$

Proof With the same notation as in Corollary 6.1, observe that $\boldsymbol{Z}_n^{-1} = (\mathbf{I}+\boldsymbol{X}_1)^{-1}\cdots(\mathbf{I}+\boldsymbol{X}_n)^{-1}$. This is an independent product that can be bounded by applying the corollary. To do so, we simply need to express $(\mathbf{I}+\boldsymbol{X}_i)^{-1} = \mathbf{I} + \bar{\boldsymbol{X}}_i$ for suitable random matrices $\bar{\boldsymbol{X}}_i$. The perturbation terms $\bar{\boldsymbol{X}}_i$ are obtained from the calculation

$$(\mathbf{I}+\boldsymbol{X}_i)^{-1} = \mathbf{I} + \sum_{k=1}^{\infty}(-1)^k \boldsymbol{X}_i^k = \mathbf{I} - \boldsymbol{X}_i + \boldsymbol{X}_i^2(\mathbf{I}+\boldsymbol{X}_i)^{-1} =: \mathbf{I} + \bar{\boldsymbol{X}}_i.$$

It remains to develop estimates for the size of the perturbation.

The uniform bound $\|\boldsymbol{X}_i\| \le \|\mathbb{E}\,\boldsymbol{X}_i\| + \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \le \xi_i + \sigma_i < 1$ implies that

$$\left\|(\mathbf{I}+\boldsymbol{X}_i)^{-1}\right\| \le \left(1 - \|\boldsymbol{X}_i\|\right)^{-1} \le \frac{1}{1-(\xi_i+\sigma_i)}.$$

Therefore, the norm of the expected perturbation satisfies

$$\left\|\mathbb{E}\,\bar{\boldsymbol{X}}_i\right\| \le \|\mathbb{E}\,\boldsymbol{X}_i\| + \left\|\mathbb{E}\left[\boldsymbol{X}_i^2(\mathbf{I}+\boldsymbol{X}_i)^{-1}\right]\right\| \le \xi_i + \frac{(\xi_i+\sigma_i)^2}{1-(\xi_i+\sigma_i)} =: \bar{\xi}_i.$$

The fluctuations of the perturbation satisfy

$$\left\|\bar{X}_i - \mathbb{E}\,\bar{X}_i\right\| \le \|X_i - \mathbb{E}\,X_i\| + 2\left\|X_i^2(\mathbf{I}+X_i)^{-1}\right\| \le \sigma_i + \frac{2(\xi_i+\sigma_i)^2}{1-(\xi_i+\sigma_i)} =: \bar{\sigma}_i.$$

The results follow when we apply Corollary 6.1 with the random matrices $\bar{X}_i$ in place of the X_i.

7 Improvements and Extensions

The argument underlying Theorem 5.1 has several natural extensions. In Section 7.1, we derive better estimates for a matrix product where the initial term is rectangular. In Section 7.2, we document the changes that are necessary in case the factors in the product are not independent but form an adapted sequence. Last, in Section 7.3, we explain how to develop a bound on the spectral radius of a product.

7.1 Low-Rank Products

So far, we have focused on the setting where the initial matrix $Z_0 = \mathbf{I}$. In many applications, we are interested in the action of the random product $Y_n \cdots Y_1 \in \mathbb{M}_d$ on a specific matrix $Z_0 \in \mathbb{C}^{d\times r}$ with relatively few columns. In this case, the terms that control the behavior of the product may be significantly smaller. Here is an example of the kinds of results one can achieve.

Theorem 7.1 (Growth and Concentration of Low-Rank Random Products) *Consider a fixed matrix $Z_0 \in \mathbb{C}^{d\times r}$ and an independent sequence $\{Y_1,\dots,Y_n\} \subset \mathbb{M}_d$ of random matrices. Form the product $Z_n = Y_n \cdots Y_1 Z_0$. Assume that*

$$\|\mathbb{E}\,Y_i\| \le m_i \quad \textit{and} \quad \sup_{P\in\mathscr{P}_r}\left(\mathbb{E}\,\|(Y_i - \mathbb{E}\,Y_i)P\|^2\right)^{1/2} \le \sigma_i m_i \quad \textit{for } i = 1,\dots,n,$$

where $\mathscr{P}_r \subset \mathbb{M}_d$ is the set of rank-r orthogonal projectors. Define $M = \prod_{i=1}^n m_i$ and $v = \sum_{i=1}^n \sigma_i^2$. For each $p \ge 2$,

$$\mathbb{E}\,\|Z_n\| \le \mathrm{e}^{\mathrm{C}_p v/2} \cdot \|Z_0\|_p \cdot M.$$
$$\mathbb{E}\,\|Z_n - \mathbb{E}\,Z_n\| \le \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \cdot \|Z_0\|_p \cdot M.$$

Proof Define $Z_i = Y_i Z_{i-1}$ for each index i. Since $Z_0 \in \mathbb{C}^{d\times r}$, the rank of each matrix Z_i is at most r. Thus, we can write $Z_i = P_i Z_i$, where P_i is a rank-r orthogonal projector that only depends on $Y_i,\dots,Y_1$ and Z_0. As a consequence,

$$\begin{aligned}
|||(Y_i - \mathbb{E}\,Y_i)Z_{i-1}|||_{p,2} &= |||(Y_i - \mathbb{E}\,Y_i)P_{i-1}Z_{i-1}|||_{p,2} \\
&\le \left(\mathbb{E}\left[\,\|(Y_i - \mathbb{E}\,Y_i)P_{i-1}\|^2 \cdot \|Z_{i-1}\|_p^2\,\right]\right)^{1/2} \\
&\le \sup_{P\in\mathscr{P}_r}\left(\mathbb{E}\,\|(Y_i - \mathbb{E}\,Y_i)P\|^2\right)^{1/2} \cdot \left(\mathbb{E}\,\|Z_{i-1}\|_p^2\right)^{1/2} \le \sigma_i m_i \cdot |||Z_{i-1}|||_{p,2}.
\end{aligned}$$

We have used the fact that Y_i is independent from P_{i-1} and from Z_{i-1} to pass to the last line.

The rest of the proof runs along the same lines as the argument in Theorem 5.1, using the last display in place of the bound (5.1).

Let us offer a simple example to illustrate why Theorem 7.1 can produce better outcomes than Theorem 5.1. Consider a random matrix $X \in \mathbb{M}_d$ with the distribution $\mathbb{P}\{X = \mathbf{e}_j\mathbf{e}_j{}^*\} = d^{-1}$ for each $j = 1,\dots,d$. As usual, $\mathbf{e}_j \in \mathbb{C}^d$ is the jth standard basis vector. Construct the random matrix $Y = \mathbf{I} + \varepsilon X$, where ε is a Rademacher random variable that is independent from X. Clearly, $\mathbb{E}\,Y = \mathbf{I}$. For any rank-$r$ orthogonal projector P,

$$\mathbb{E}\,\|(Y - \mathbb{E}\,Y)P\|^2 = \mathbb{E}\,\|PX^*XP\| = \frac{1}{d}\sum_{i=1}^d \operatorname{tr}[P\mathbf{e}_i\mathbf{e}_i^*P] = \frac{1}{d}\operatorname{tr} P = \frac{r}{d}.$$

Therefore,

$$\sup_{P \in \mathcal{P}_r} \left(\mathbb{E} \, \|(Y - \mathbb{E}\, Y)P\|^2 \right)^{1/2} = \sqrt{r/d} \leq 1.$$

By contrast, $\mathbb{E}\,\|Y - \mathbb{E}\,Y\|^2 = \mathbb{E}\,\|X\|^2 = 1$. This bound hence offers a significant improvement when $r \ll d$.

7.2 Adapted Sequences

We can easily generalize our results on a product of independent random matrices to a product of adapted random matrices. This kind of extension is valuable for studying iterative algorithms where the choices made by the algorithm at a given step depend on the history of the iteration.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \mathcal{F}$ be a filtration. For each index $i = 1, \dots, n$, we write $\mathbb{E}_i$ for the expectation conditioned on the σ-algebra $\mathcal{F}_i$. The operator $\mathbb{E}_0 := \mathbb{E}$ is the unconditional expectation.

We consider an adapted sequence $\{Y_1, \dots, Y_n\} \subset \mathbb{M}_d$ of random matrices; that is, each Y_i is measurable with respect to $\mathcal{F}_i$. The next result provides information about the growth and concentration properties of the product $Z_n = Y_n \cdots Y_1$. Note that the natural concentration result compares Z_n with a product of conditional expectations, rather than the expectation of the product.

Theorem 7.2 (Products of Adapted Random Matrices) *Consider a fixed matrix $Z_0 \in \mathbb{M}_d$ and an adapted sequence $\{Y_1, \dots, Y_n\} \subset \mathbb{M}_d$ of random matrices. Form the products*

$$Z_n = Y_n \cdots Y_1 Z_0 \quad \text{and} \quad F_n = (\mathbb{E}_{n-1}\, Y_n) \cdots (\mathbb{E}_1\, Y_2)(\mathbb{E}_0\, Y_1) Z_0.$$

Assume that

$$\|\mathbb{E}_{i-1}\, Y_i\| \leq m_i \quad \text{and} \quad \|Y_i - \mathbb{E}_{i-1}\, Y_i\| \leq \sigma_i m_i \quad \text{almost surely for } i = 1, \dots, n.$$

Define $M = \prod_{i=1}^n m_i$ and $v = \sum_{i=1}^n \sigma_i^2$. For $2 \leq q \leq p$, the random product Z_n satisfies the growth and concentration bounds

$$|||Z_n|||_{p,q} \leq \mathrm{e}^{C_p v/2} \, \|Z_0\|_p \cdot M; \tag{7.1}$$

$$|||Z_n - F_n|||_{p,q} \leq \left(\mathrm{e}^{C_p v} - 1\right)^{1/2} \|Z_0\|_p \cdot M. \tag{7.2}$$

Proof Recursively construct the products

$$Z_i = Y_i Z_{i-1} \quad \text{and} \quad F_i = (\mathbb{E}_{i-1}\, Y_i) F_{i-1} \quad \text{for } i = 1, \dots, n.$$

To bound the growth of Z_i and the concentration of $Z_i - F_i$, we simply need to update the argument from Theorem 5.1.

To obtain (7.2), decompose

$$Z_i = (\mathbb{E}_{i-1}\, Y_i) Z_{i-1} + (Y_i - \mathbb{E}_{i-1}\, Y_i) Z_{i-1}.$$

Since $\mathbb{E}_{i-1}\, Y_i$ and Z_{i-1} are both measurable with respect to $\mathcal{F}_{i-1}$ and $\mathbb{E}_{i-1}(Y_i - \mathbb{E}_{i-1}\, Y_i) = \mathbf{0}$, the obvious variant of Proposition 4.1 implies that

$$\begin{aligned} |||Z_i|||_{p,q}^2 &\leq |||(\mathbb{E}_{i-1}\, Y_i) Z_{i-1}|||_{p,q}^2 + C_p |||(Y_i - \mathbb{E}_{i-1}\, Y_i) Z_{i-1}|||_{p,q}^2 \\ &\leq m_i^2 |||Z_{i-1}|||_{p,q} + C_p m_i^2 \sigma_i^2 |||Z_{i-1}|||_{p,q}^2. \end{aligned}$$

The second inequality follows from (4.1). This is the same recurrence we obtain in the proof of Theorem 5.1, relation (5.1). The rest of the argument is the same.

To obtain (7.2), decompose

$$Z_i - F_i = Y_i Z_{i-1} - (\mathbb{E}_{i-1}\, Y_i) F_{i-1} = (\mathbb{E}_{i-1}\, Y_i)(Z_{i-1} - F_{i-1}) + (Y_i - \mathbb{E}_{i-1}\, Y_i) Z_{i-1}.$$

As before, Proposition 4.1 implies that

$$\begin{aligned} |||\boldsymbol{Z}_i - \boldsymbol{F}_i|||_{p,q}^2 &\le |||(\mathbb{E}_{i-1}\boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \boldsymbol{F}_{i-1})|||_{p,q}^2 + \mathrm{C}_p|||(\boldsymbol{Y}_i - \mathbb{E}_{i-1}\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q}^2 \\ &\le m_i^2 |||\boldsymbol{Z}_{i-1} - \boldsymbol{F}_{i-1}|||_{p,q} + \mathrm{C}_p m_i^2 \sigma_i^2 |||\boldsymbol{Z}_{i-1}|||_{p,q}^2. \end{aligned}$$

This is the same recurrence that arose when we established Theorem 5.1, relation (5.1). The balance of the argument is identical.

7.3 The Spectral Radius

Products of matrices are closely related to the evolution of discrete-time linear dynamical systems. In this context, it may be more natural to study the *spectral radius* of the matrix product, rather than its spectral norm. Bounds for the spectral radius follow as corollary of our work, owing to the following classical fact.

Fact 7.3 (Schur) *Let $\boldsymbol{M} \in \mathbb{M}_d$ be a square matrix. The spectral radius $\varrho(\boldsymbol{M})$ is defined as the maximum absolute value of an eigenvalue of $\boldsymbol{M}$. It satisfies the variational principle*

$$\varrho(\boldsymbol{M}) = \inf_{\boldsymbol{S} \in \mathbb{M}_d} \left\| \boldsymbol{S}^{-1}\boldsymbol{M}\boldsymbol{S} \right\|.$$

The infimum takes place over all invertible matrices $\boldsymbol{S}$. In particular $\varrho(\boldsymbol{M}) \le \|\boldsymbol{M}\|$.

Let us give an indication of the kinds of results that are possible.

Corollary 7.1 (Expectation Bounds for the Spectral Radius) *Consider an independent sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices, and form the product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. Let $\boldsymbol{S} \in \mathbb{M}_d$ be a fixed invertible matrix, and assume that*

$$\left\| \boldsymbol{S}^{-1}(\mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{S} \right\| \le m_i \quad \text{and} \quad \left(\mathbb{E} \left\| \boldsymbol{S}^{-1}(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{S} \right\|^2 \right)^{1/2} \le \sigma_i m_i \quad \text{for } i = 1, \dots, n.$$

Let $M = \prod_{i=1}^n m_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$\mathbb{E}\,\varrho(\boldsymbol{Z}_n) \le \exp\left(\sqrt{2v\,(2v \vee \log d)} \right) \cdot M.$$

Proof Combine Corollary 5.1 and Fact 7.3.

7.4 Prospects

We have developed a collection of nonasymptotic bounds for products of random matrices. These results hold under simple and easily verifiable conditions, and they give accurate predictions about the behavior of some particular instances (e.g., products of iid random perturbations of the identity). The proofs are based on foundational results about the geometry of the Schatten classes, and they can easily be adapted to treat variants of the problems under consideration.

A disappointing feature of our results is that they do not account for interactions between the matrix factors. For example, when $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i/n$ for bounded, independent matrix perturbations $\boldsymbol{X}_i$, we have shown that

$$\log \mathbb{E}\,\|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| \le \frac{1}{n} \sum_{i=1}^n \|\,\mathbb{E}\,\boldsymbol{X}_i\| + O\left(\sqrt{\frac{\log d}{n}} \right).$$

However, when the matrices $\boldsymbol{X}_i$ are Hermitian and commute almost surely, it is easy to show the sharper bound

$$\log \mathbb{E}\,\|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| \le \frac{1}{n} \left\| \sum_{i=1}^n \mathbb{E}\,\boldsymbol{X}_i \right\| + O\left(\sqrt{\frac{\log d}{n}} \right).$$

The results of Emme and Hubert [14] establish that $\lim_{n\to\infty} \log \mathbb{E} \|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| = \lim_{n\to\infty} \left\| \sum_{i=1}^n \mathbb{E}\, \boldsymbol{X}_i \right\| / n$. It therefore seems reasonable to conjecture that a refined bound of the latter type exists in more generality. The growth bounds discussed in Remark 5.1 imply a statement of the form

$$\log \mathbb{E} \|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| \le \frac{1}{n} \left\| \sum_{i=1}^n \mathbb{E}\, \boldsymbol{X}_i \right\| + \text{error},$$

but the error term is not sharp. This type of bound would echo Tropp's improvements [42] to the Ahlswede–Winter results [1] for a sum of independent random matrices. At present, it is not clear whether this refinement is possible, nor what technical arguments would lead there.

A Supplementary Proofs

This appendix collects a few additional arguments. First, we establish the sharp form of the result on subquadratic averages, Proposition 4.1, using an elementary method.

Lemma A.1 (Sharp Subquadratic Averages) *Let $\boldsymbol{X}, \boldsymbol{Y}$ be random matrices of the same size that satisfy $\mathbb{E}[\boldsymbol{Y} | \boldsymbol{X}] = \boldsymbol{0}$. When $2 \le q \le p$,*

$$|||\boldsymbol{X} + \boldsymbol{Y}|||_{p,q}^2 \le |||\boldsymbol{X}|||_{p,q}^2 + \mathrm{C}_p\, |||\boldsymbol{Y}|||_{p,q}^2,$$

where the optimal constant $\mathrm{C}_p := p - 1$.

Proof Fix a natural number n, and set $\boldsymbol{Z} = n^{-1}\boldsymbol{Y}$. Inequality (4.4) states that

$$D_1 := |||\boldsymbol{X} + \boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X}|||_{p,q}^2 - 2\mathrm{C}_p\, |||\boldsymbol{Z}|||_{p,q}^2 \le 0.$$

For a parameter $2 \le k \le n$, Corollary 4.1 and Lyapunov's inequality imply that

$$|||\boldsymbol{X} + k\boldsymbol{Z}|||_{p,q}^2 + |||\boldsymbol{X} + (k-2)\boldsymbol{Z}|||_{p,q}^2 \le 2\, |||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 + 2\mathrm{C}_p\, |||\boldsymbol{Z}|||_{p,q}^2.$$

Rearranging the last display, we see that

$$\begin{aligned} D_k &:= |||\boldsymbol{X} + k\boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 - 2\mathrm{C}_p k\, |||\boldsymbol{Z}|||_{p,q}^2 \\ &\le |||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X} + (k-2)\boldsymbol{Z}|||_{p,q}^2 - 2\mathrm{C}_p (k-1)\, |||\boldsymbol{Z}|||_{p,q}^2 = D_{k-1}. \end{aligned}$$

In particular, $D_k \le D_1 \le 0$. Using a telescoping sum,

$$\begin{aligned} |||\boldsymbol{X} + \boldsymbol{Y}|||_{p,q}^2 - |||\boldsymbol{X}|||_{p,q}^2 &= \sum_{k=1}^n \left(|||\boldsymbol{X} + k\boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 \right) \\ &= \sum_{k=1}^n \left(D_k + 2\mathrm{C}_p k\, |||\boldsymbol{Z}|||_{p,q}^2 \right) \le \sum_{k=1}^n 2\mathrm{C}_p k\, |||\boldsymbol{Z}|||_{p,q}^2 = \mathrm{C}_p \frac{n+1}{n}\, |||\boldsymbol{Y}|||_{p,q}^2. \end{aligned}$$

Take the limit as $n \to \infty$ to arrive at the stated result.

Second, we present a basic numerical inequality for weighted sums of exponentials.

Lemma A.2 *Let $a_1, a_2, \dots, a_n$ be a sequence of real numbers. Then*

$$\sum_{i=1}^n a_i \exp\left(\sum_{k=1}^{i-1} a_k \right) \le \exp\left(\sum_{i=1}^n a_i \right) - 1.$$

Proof The elementary inequality $a \le \mathrm{e}^a - 1$, valid for $a \in \mathbb{R}$, implies that

$$a_i \exp\left(\sum_{k=1}^{i-1} a_k \right) \le \exp\left(\sum_{k=1}^{i} a_k \right) - \exp\left(\sum_{k=1}^{i-1} a_k \right).$$

Sum the displayed equation over $i = 1, \dots, n$ to verify the claim.

References

1. Ahlswede, R., Winter, A.: Strong converse for identification via quantum channels. IEEE Trans. Inform. Theory **48**(3), 569–579 (2002). doi:10.1109/18.985947
2. Albar, W., Junge, M., Zhao, M.: Noncommutative versions of the arithmetic-geometric mean inequality (2017)
3. Altschuler, J.M., Parrilo, P.A.: Lyapunov exponent of rank one matrices: Ergodic formula and inapproximability of the optimal distribution (2019)
4. Ando, T.: Concavity of certain maps on positive definite matrices and applications to Hadamard products. Linear Algebra Appl. **26**, 203–241 (1979). doi:10.1016/0024-3795(79)90179-4
5. Bai, Z., Silverstein, J.W.: Spectral analysis of large dimensional random matrices, second edn. Springer Series in Statistics. Springer, New York (2010). doi:10.1007/978-1-4419-0661-8
6. Ball, K., Carlen, E.A., Lieb, E.H.: Sharp uniform convexity and smoothness inequalities for trace norms. Inventiones mathematicae **115**(1), 463–482 (1994). doi:10.1007/BF01231769
7. Beckner, W.: Inequalities in Fourier analysis. Ann. of Math. (2) **102**(1), 159–182 (1975). doi:10.2307/1970980
8. Benoist, Y., Quint, J.F.: Random walks on reductive groups, vol. 62. Springer, Cham (2016). doi:10.1007/978-3-319-47721-3
9. Berger, M.A.: Central limit theorem for products of random matrices. Trans. Amer. Math. Soc. **285**(2), 777–803 (1984). doi:10.2307/1999463
10. Bhatia, R.: Matrix analysis, *Graduate Texts in Mathematics*, vol. 169. Springer-Verlag, New York (1997). doi:10.1007/978-1-4612-0653-8
11. Bonami, A.: étude des coefficients de fourier des fonctions de $l^p(g)$. Ann. Inst. Fourier (Grenoble) **20**(2), 335–402 (1970). doi:10.5802/aif.357
12. Collectif: Sharp inequalities for martingales and stochastic integrals. In: Colloque Paul Lévy sur les processus stochastiques, no. 157-158 in Astérisque, pp. 75–94. Société mathématique de France (1988)
13. Dartois, S., Forrester, P.J.: Schwinger-dyson and loop equations for a product of square ginibre random matrices. Journal of Physics A: Mathematical and Theoretical (2020). doi:10.1088/1751-8121/ab6fc4
14. Emme, J., Hubert, P.: Limit laws for random matrix products (2017)
15. Furman, A.: Random walks on groups and random transformations. In: Handbook of dynamical systems, Vol. 1A, pp. 931–1014. North-Holland, Amsterdam (2002). doi:10.1016/S1874-575X(02)80014-5
16. Furstenberg, H., Kesten, H.: Products of random matrices. Ann. Math. Statist. **31**, 457–469 (1960). doi:10.1214/aoms/1177705909
17. Gross, L.: Existence and uniqueness of physical ground states. J. Functional Analysis **10**, 52–109 (1972). doi:10.1016/0022-1236(72)90057-2
18. Gürbüzbalaban, M., Ozdaglar, A., Parrilo, P.: Why random reshuffling beats stochastic gradient descent. Mathematical Programming pp. 1–36 (2019). doi:10.1007/s10107-019-01440-w
19. Hanin, B., Nica, M.: Products of many large random matrices and gradients in deep neural networks. Communications in Mathematical Physics pp. 1–36 (2019). doi:10.1007/s00220-019-03624-z
20. Hanin, B., Paouris, G.: Non-asymptotic Results for Singular Values of Gaussian Matrix Products. arXiv e-prints arXiv:2005.08899 (2020)
21. Henriksen, A., Ward, R.: Concentration inequalities for random matrix products (2019)
22. Israel, A., Krahmer, F., Ward, R.: An arithmetic-geometric mean inequality for products of three matrices. Linear Algebra Appl. **488**, 1–12 (2016). doi:10.1016/j.laa.2015.09.013
23. Kathuria, T., Mukherjee, S., Srivastava, N.: On Concentration Inequalities for Random Matrix Products. arXiv e-prints arXiv:2003.06319 (2020)
24. Kieburg, M.: Products of Complex Rectangular and Hermitian Random Matrices (2019)
25. Ledoux, M., Talagrand, M.: Probability in Banach spaces. Classics in Mathematics. Springer-Verlag, Berlin (2011). doi:10.1007/978-3-642-20212-4. Isoperimetry and processes, Reprint of the 1991 edition
26. Ledrappier, F.: Some asymptotic properties of random walks on free groups. In: Topics in probability and Lie groups: boundary theory, vol. 28, pp. 117–152. Amer. Math. Soc., Providence, RI (2001). doi:10.1007/978-3-642-20212-4
27. Lieb, E.H.: Convex trace functions and the Wigner-Yanase-Dyson conjecture. Advances in Math. **11**, 267–288 (1973). doi:10.1016/0001-8708(73)90011-X
28. Lust-Piquard, F.: Inégalités de Khintchine dans C_p $(1 < p < \infty)$. C. R. Acad. Sci. Paris Sér. I Math. **303**(7), 289–292 (1986)
29. Naor, A.: On the banach-space-valued azuma inequality and small-set isoperimetry of alon–roichman graphs. Combinatorics, Probability and Computing **21**(4), 623–634 (2012). doi:10.1017/S0963548311000757
30. Nica, A., Speicher, R.: Lectures on the combinatorics of free probability, *London Mathematical Society Lecture Note Series*, vol. 335. Cambridge University Press, Cambridge (2006). doi:10.1017/CBO9780511735127
31. Oja, E.: A simplified neuron model as a principal component analyzer. J. Math. Biol. **15**(3), 267–273 (1982). doi:10.1007/BF00275687
32. Oliveira, R.I.: Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges (2009)
33. Pisier, G.: Martingales with values in uniformly convex spaces. Israel J. Math. **20**(3-4), 326–350 (1975). doi:10.1007/BF02760337
34. Rao, N.R., Edelman, A.: The polynomial method for random matrices. Found. Comput. Math. **8**(6), 649–702 (2008). doi:10.1007/s10208-007-9013-x
35. Recht, B., Re, C.: Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences (2012)
36. Ricard, É., Xu, Q.: A noncommutative martingale convexity inequality. The Annals of Probability **44**(2), 867–882 (2016). doi:10.1214/14-AOP990
37. Shlyakhtenko, D.: Random matrices and free probability. In: Random Matrices, 26. Amer. Math. Soc., Providence, RI (2019)
38. Speicher, R.: Lecture notes on "free probability theory" (2019)
39. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl. **15**(2), 262–278 (2009). doi:10.1007/s00041-008-9030-4
40. Tomczak-Jaegermann, N.: The moduli of smoothness and convexity and the Rademacher averages of trace classes $S_p (1 \leq p < \infty)$. Studia Math. **50**, 163–182 (1974)

41. Tropp, J.A.: Freedman's inequality for matrix martingales. Electronic Communications in Probability **16**, 262–270 (2011). doi:10.1214/ECP.v16-1624
42. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. Comput. Math. **12**(4), 389–434 (2012). doi:10.1007/s10208-011-9099-z
43. Tropp, J.A.: An introduction to matrix concentration inequalities. Foundations and Trends in Machine Learning **8**(1-2), 1–230 (2015). doi:10.1561/2200000048
44. Tropp, J.A.: The expected norm of a sum of independent random matrices: an elementary approach. In: High dimensional probability VII, *Progr. Probab.*, vol. 71, pp. 173–202. Springer, [Cham] (2016). doi:10.1007/978-3-319-40519-3_8
45. Tropp, J.A.: Second-order matrix concentration inequalities. Appl. Comput. Harmon. Anal. **44**(3), 700–736 (2018). doi:10.1016/j.acha.2016.07.005
46. Tsitsiklis, J.N., Blondel, V.D.: The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. Math. Control Signals Systems **10**(1), 31–40 (1997). doi:10.1007/BF01219774
47. Tulino, A.M., Verdú, S.: Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory **1**(1), 1–182 (2004). doi:10.1561/0100000001
48. Wilkinson, A.: What are Lyapunov exponents, and why are they interesting? Bull. Amer. Math. Soc. (N.S.) **54**(1), 79–105 (2017). doi:10.1090/bull/1552
49. Yang, G.: Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation (2019)