Full text available at: http://dx.doi.org/10.1561/2200000048

An Introduction to Matrix Concentration Inequalities

Joel A. Tropp
Computing & Mathematical Sciences
California Institute of Technology
jtropp@cms.caltech.edu

now
the essence of knowledge
Boston — Delft

Full text available at: http://dx.doi.org/10.1561/2200000048

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

J. A. Tropp. *An Introduction to Matrix Concentration Inequalities*. Foundations and Trends® in Machine Learning, vol. 8, no. 1-2, pp. 1–230, 2015.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-839-3
© 2015 J. A. Tropp

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1 781 871 0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Full text available at: http://dx.doi.org/10.1561/2200000048

Foundations and Trends® in Machine Learning

Volume 8, Issue 1-2, 2015

Editorial Board

Editor-in-Chief

Michael Jordan
University of California, Berkeley
United States

Editors

Peter Bartlett
UC Berkeley

Yoshua Bengio
University of Montreal

Avrim Blum
CMU

Craig Boutilier
University of Toronto

Stephen Boyd
Stanford University

Carla Brodley
Tufts University

Inderjit Dhillon
UT Austin

Jerome Friedman
Stanford University

Kenji Fukumizu
ISM, Japan

Zoubin Ghahramani
University of Cambridge

David Heckerman
Microsoft Research

Tom Heskes
Radboud University

Geoffrey Hinton
University of Toronto

Aapo Hyvarinen
HIIT, Finland

Leslie Pack Kaelbling
MIT

Michael Kearns
UPenn

Daphne Koller
Stanford University

John Lafferty
University of Chicago

Michael Littman
Brown University

Gabor Lugosi
Pompeu Fabra University

David Madigan
Columbia University

Pascal Massart
University of Paris-Sud

Andrew McCallum
UMass Amherst

Marina Meila
University of Washington

Andrew Moore
CMU

John Platt
Microsoft Research

Luc de Raedt
University of Freiburg

Christian Robert
U Paris-Dauphine

Sunita Sarawagi
IIT Bombay

Robert Schapire
Princeton University

Bernhard Schoelkopf
MPI Tübingen

Richard Sutton
University of Alberta

Larry Wasserman
CMU

Bin Yu
UC Berkeley

Full text available at: http://dx.doi.org/10.1561/2200000048

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2015, Volume 8, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Full text available at: http://dx.doi.org/10.1561/2200000048

Foundations and Trends® in Machine Learning
Vol. 8, No. 1-2 (2015) 1–230
© 2015 J. A. Tropp
DOI: 10.1561/2200000048

now
the essence of knowledge

An Introduction to Matrix Concentration Inequalities

Joel A. Tropp
Computing & Mathematical Sciences
California Institute of Technology
jtropp@cms.caltech.edu

Full text available at: http://dx.doi.org/10.1561/2200000048

Contents

Full text available at: http://dx.doi.org/10.1561/2200000048

Full text available at: http://dx.doi.org/10.1561/2200000048

Full text available at: http://dx.doi.org/10.1561/2200000048

Preface

In recent years, random matrices have come to play a major role in computational mathematics, but most of the classical areas of random matrix theory remain the province of experts. Over the last decade, with the advent of matrix concentration inequalities, research has advanced to the point where we can conquer many (formerly) challenging problems with a page or two of arithmetic.

My aim is to describe the most successful methods from this area along with some interesting examples that these techniques can illuminate. I hope that the results in these pages will inspire future work on applied random matrix theory as well as refinements of the matrix concentration inequalities discussed herein.

I have chosen to present a coherent body of results based on a generalization of the Laplace transform method for establishing scalar concentration inequalities. In the last two years, Lester Mackey and I, together with our coauthors, have developed an alternative approach to matrix concentration using exchangeable pairs and Markov chain couplings. With some regret, I have chosen to omit this theory because the ideas seem less accessible to a broad audience of researchers. The interested reader will find pointers to these articles in the annotated bibliography.

The work described in these notes reflects the influence of many researchers. These include Rudolf Ahlswede, Rajendra Bhatia, Eric

Full text available at: http://dx.doi.org/10.1561/2200000048

Carlen, Sourav Chatterjee, Edward Effros, Elliott Lieb, Roberto Imbuzeiro Oliveira, Dénes Petz, Gilles Pisier, Mark Rudelson, Roman Vershynin, and Andreas Winter. I have also learned a great deal from other colleagues and friends along the way.

I would like to thank some people who have helped me improve this work. Several readers informed me about errors in the initial version of this manuscript; these include Serg Bogdanov, Peter Forrester, Nikos Karampatziakis, and Guido Lagos. The anonymous reviewers tendered many useful suggestions, and they pointed out a number of mistakes. Sid Barman gave me feedback on the final revisions to the monograph.

Last, I want to thank Léon Nijensohn for his continuing encouragement. This work is dedicated to my family, to my wife Margot, and my son Benjamin.

I gratefully acknowledge financial support from the Office of Naval Research under awards N00014-08-1-0883 and N00014-11-1002, the Air Force Office of Strategic Research under award FA9550-09-1-0643, and an Alfred P. Sloan Fellowship. Some of this research was completed at the Institute of Pure and Applied Mathematics at UCLA. I would also like to thank the California Institute of Technology and the Moore Foundation.

Joel A. Tropp
Pasadena, CA
December 2012
Revised, March 2014, December 2014, and April 2015

Full text available at: http://dx.doi.org/10.1561/2200000048

Abstract

Random matrices now play a role in many areas of theoretical, applied, and computational mathematics. Therefore, it is desirable to have tools for studying random matrices that are flexible, easy to use, and powerful. Over the last fifteen years, researchers have developed a remarkable family of results, called *matrix concentration inequalities*, that achieve all of these goals.

This monograph offers an invitation to the field of matrix concentration inequalities. It begins with some history of random matrix theory; it describes a flexible model for random matrices that is suitable for many problems; and it discusses the most important matrix concentration results. To demonstrate the value of these techniques, the presentation includes examples drawn from statistics, machine learning, optimization, combinatorics, algorithms, scientific computing, and beyond.

J. A. Tropp. *An Introduction to Matrix Concentration Inequalities*. Foundations and Trends® in Machine Learning, vol. 8, no. 1-2, pp. 1–230, 2015.
DOI: 10.1561/2200000048.

Full text available at: http://dx.doi.org/10.1561/2200000048

1

Introduction

Random matrix theory has grown into a vital area of probability, and it has found applications in many other fields. To motivate the results in this monograph, we begin with an overview of the connections between random matrix theory and computational mathematics. We introduce the basic ideas underlying our approach, and we state one of our main results on the behavior of random matrices. As an application, we examine the properties of the sample covariance estimator, a random matrix that arises in statistics. Afterward, we summarize the other types of results that appear in this monograph, and we assess the novelties in this presentation.

1.1 Historical Origins

Random matrix theory sprang from several different sources in the first half of the 20th century.

Geometry of Numbers. Peter Forrester [65, p. v] traces the field of random matrix theory to work of Hurwitz, who defined the invariant integral over a Lie group. Specializing this analysis to the

Full text available at: http://dx.doi.org/10.1561/2200000048

orthogonal group, we can reinterpret this integral as the expectation of a function of a uniformly random orthogonal matrix.

Multivariate Statistics. Another early example of a random matrix appeared in the work of John Wishart [192]. Wishart was studying the behavior of the sample covariance estimator for the covariance matrix of a multivariate normal random vector. He showed that the estimator, which is a random matrix, has the distribution that now bears his name. Statisticians have often used random matrices as models for multivariate data [123, 133].

Numerical Linear Algebra. In their remarkable work [189, 74] on computational methods for solving systems of linear equations, von Neumann and Goldstine considered a random matrix model for the floating-point errors that arise from an LU decomposition.[1] They obtained a high-probability bound for the norm of the random matrix, which they interpreted as an estimate for the error the procedure might typically incur. Curiously, in subsequent years, numerical linear algebraists became very suspicious of probabilistic techniques, and only in recent years have randomized algorithms reappeared in this field. See the surveys [120, 80, 193] for more details and references.

Nuclear Physics. In the early 1950s, physicists had reached the limits of deterministic analytical techniques for studying the energy spectra of heavy atoms undergoing slow nuclear reactions. Eugene Wigner was the first researcher to surmise that a random matrix with appropriate symmetries might serve as a suitable model for the Hamiltonian of the quantum mechanical system that describes the reaction. The eigenvalues of this random matrix model the possible energy levels of the system. See Mehta's book [128, §1.1] for an account of all this.

In each area, the motivation was quite different and led to distinct sets of questions. Later, random matrices began to percolate into other

[1] von Neumann and Goldstine invented and analyzed this algorithm before they had any digital computer on which to implement it! See [76] for an historical account.

Full text available at: http://dx.doi.org/10.1561/2200000048

fields such as graph theory (the Erdős–Rényi model [61] for a random graph) and number theory (as a model for the spacing of zeros of the Riemann zeta function [131]).

1.2 The Modern Random Matrix

By now, random matrices are ubiquitous. They arise throughout modern mathematics and statistics, as well as in many branches of science and engineering. Random matrices have several different purposes that we may wish to distinguish. They can be used within randomized computer algorithms; they serve as models for data and for physical phenomena; and they are subjects of mathematical inquiry. This section offers a taste of these applications. Note that the ideas and references here reflect the author's interests, and they are far from comprehensive!

1.2.1 Algorithmic Applications

The striking mathematical properties of random matrices can be harnessed to develop algorithms for solving many different problems.

Computing Matrix Approximations. Random matrices can be used to develop fast algorithms for computing a truncated singular-value decomposition. In this application, we multiply a large input matrix by a smaller random matrix to extract information about the dominant singular vectors of the input matrix. The seed of this idea appears in [68, 53]. The survey [80] explains how to implement this method in practice, while the two monographs [120, 193] cover more theoretical aspects.

Sparsification. One way to accelerate spectral computations on large matrices is to replace the original matrix by a sparse proxy that has similar spectral properties. An elegant way to produce the sparse proxy is to zero out entries of the original matrix at random while rescaling the entries that remain. This approach was proposed in [3, 4], and the papers [2, 105] contain recent innovations. Related ideas play an important role in Spielman and Teng's work [171] on fast algorithms for solving linear systems.

Full text available at: http://dx.doi.org/10.1561/2200000048

Subsampling of Data. In large-scale machine learning, one may need to subsample data randomly to reduce the computational costs of fitting a model. For instance, we can combine random sampling with the Nyström decomposition to obtain a randomized approximation of a kernel matrix. This method was introduced to machine learning by Williams & Seeger [191]. The paper [56] provides the first theoretical analysis, and the survey [70] contains more complete results.

Dimension Reduction. To reduce the size of a computational problem, one may apply a randomized projection to the problem data. This idea is used heavily in the theory of algorithms. Many types of dimension reduction are based on properties of random matrices. The two papers [92, 24] established the mathematical foundations of this approach. The earliest applications in computer science appear in the work [112]. Many contemporary variants depend on ideas from [6] and [45].

Combinatorial Optimization. One approach to solving a computationally difficult optimization problem is to relax (i.e., enlarge) the constraint set so the problem becomes tractable, to solve the relaxed problem, and then to use a randomized procedure to map the solution back to the original constraint set [17, §4.3]. This technique is called *relaxation and rounding*. For hard optimization problems involving a matrix variable, the analysis of the rounding procedure often involves ideas from random matrix theory [170, 134].

Compressed Sensing. When acquiring data about an object with relatively few degrees of freedom as compared with the ambient dimension, we may be able to sieve out the important information from the object by taking a small number of random measurements, where the number of measurements is comparable with the number of degrees of freedom [69, 32, 52]. This observation is now referred to as *compressed sensing*. Random matrices play a central role in the design and analysis of measurement procedures. For example, see [66, 36, 9, 185].

Full text available at: http://dx.doi.org/10.1561/2200000048

1.2.2 Modeling

Random matrices also appear as models for multivariate data or multivariate phenomena. By studying the properties of these models, we may hope to understand the typical behavior of a data-analysis algorithm or a physical system.

Sparse Approximation for Random Signals. Sparse approximation has become an important problem in statistics, signal processing, machine learning and other areas. One model for a "typical" sparse signal poses the assumption that the nonzero coefficients that generate the signal are chosen at random. When analyzing methods for identifying the sparse set of coefficients, we must study the behavior of a random column submatrix drawn from the model matrix [177, 176].

Demixing of Structured Signals. In data analysis, it is common to encounter a mixture of two structured signals, and the goal is to extract the two signals using prior information about the structures [178, 37]. A common model for this problem assumes that the signals are randomly oriented with respect to each other, which means that it is usually possible to discriminate the underlying structures. Random orthogonal matrices arise in the analysis of estimation techniques for this problem [125, 9, 126].

Stochastic Block Model. One probabilistic framework for describing community structure in a network assumes that each pair of individuals in the same community has a relationship with high probability, while each pair of individuals drawn from different communities has a relationship with lower probability. This is referred to as the *stochastic block model* [88]. It is quite common to analyze algorithms for extracting community structure from data by positing that this model holds. See [1] for a recent contribution, as well as a summary of the extensive literature.

High-Dimensional Data Analysis. More generally, random models are pervasive in the analysis of statistical estimation proce-

Full text available at: http://dx.doi.org/10.1561/2200000048

dures for high-dimensional data. Random matrix theory plays a key role in this field [123, 133, 101, 31].

Wireless Communication. Random matrices are commonly used as models for wireless channels. See the book of Tulino and Verdú for more information [187].

In these examples, it is important to recognize that random models may not coincide very well with reality, but they allow us to get a sense of what might be possible in some generic cases.

1.2.3 Theoretical Aspects

Random matrices are frequently studied for their intrinsic mathematical interest. In some fields, they provide examples of striking phenomena. In other areas, they furnish counterexamples to "intuitive" conjectures. Here are a few disparate problems where random matrices play a role.

Combinatorics. An expander graph has the property that every small set of vertices has edges linking it to a large proportion of the other vertices. The expansion property is closely related to the spectral behavior of the adjacency matrix of the graph. The easiest construction of an expander involves a random matrix argument [8, §9.2].

Numerical Analysis. For worst-case examples, the Gaussian elimination method for solving a linear system is not numerically stable. In practice, however, stability problems rarely arise. One explanation for this phenomenon is that, with high probability, a small random perturbation of any fixed matrix is well conditioned. As a consequence, it can be shown that Gaussian elimination is stable for most matrices [163].

High-Dimensional Geometry. Dvoretzky's Theorem states that, when N is large, the unit ball of each N-dimensional Banach space has a slice of dimension $n \approx \log N$ that is close to a Euclidean ball with dimension n. It turns out that a *random* slice of

Full text available at: http://dx.doi.org/10.1561/2200000048

dimension n realizes this property [129]. This result can be framed as a statement about spectral properties of a random matrix [75].

Quantum Information Theory. Random matrices appear as counterexamples for a number of conjectures in quantum information theory. Here is one instance. In classical information theory, the total amount of information that we can transmit through a pair of channels equals the sum of the information we can send through each channel separately. It was conjectured that the same property holds for quantum channels. In fact, a pair of quantum channels can have strictly larger capacity than a single channel because of entanglement. This result depends on a random matrix construction [84]. See [85] for related work.

1.3 Random Matrices for the People

Historically, random matrix theory has been regarded as a very challenging field. Even now, many well-established methods are only comprehensible to researchers with significant experience, and it may take months of intensive effort to prove new results. There are a small number of classes of random matrices that have been studied so completely that we know almost everything about them. Yet, moving beyond this *terra firma*, one quickly encounters examples where classical methods are brittle.

We hope to democratize random matrix theory. This monograph describes tools that deliver useful information about a wide range of random matrices. In many cases, a modest amount of straightforward arithmetic leads to strong results. The methods here should be accessible to computational scientists working in a variety of fields. Indeed, the techniques in this work have already found an extensive number of applications.

1.4 Basic Questions in Random Matrix Theory

Random matrices merit special attention because they have spectral properties that are quite different from familiar deterministic matrices.

Full text available at: http://dx.doi.org/10.1561/2200000048

Here are some of the questions we might want to investigate.

- What is the expectation of the maximum eigenvalue of a random Hermitian matrix? What about the minimum eigenvalue?
- How is the maximum eigenvalue of a random Hermitian matrix distributed? What is the probability that it takes values substantially different from its mean? What about the minimum eigenvalue?
- What is the expected spectral norm of a random matrix? What is the probability that the norm takes a value substantially different from its mean?
- What about the other eigenvalues or singular values? Can we say something about the "typical" spectrum of a random matrix?
- Can we say anything about the eigenvectors or singular vectors? For instance, is each one distributed almost uniformly on the sphere?
- We can also ask questions about the operator norm of a random matrix acting as a map between two normed linear spaces. In this case, the geometry of the domain and codomain play a role.

In this work, we focus on the first three questions above. We study the expectation of the extreme eigenvalues of a random Hermitian matrix, and we attempt to provide bounds on the probability that they take an unusual value. As an application of these results, we can control the expected spectral norm of a general matrix and bound the probability of a large deviation. These are the most relevant problems in many (but not all!) applications. The remaining questions are also important, but we will not touch on them here. We recommend the book [173] for a friendly introduction to other branches of random matrix theory.

1.5 Random Matrices as Independent Sums

Our approach to random matrices depends on a fundamental principle:

Full text available at: http://dx.doi.org/10.1561/2200000048

> **In applications, it is common that a random matrix can be expressed as a sum of independent random matrices.**

The examples that appear in these notes should provide ample evidence for this claim. For now, let us describe a specific problem that will serve as an illustration throughout the Introduction. We hope this example is complicated enough to be interesting but simple enough to elucidate the main points.

1.5.1 Example: The Sample Covariance Estimator

Let $\boldsymbol{x} = (X_1, \dots, X_p)$ be a complex random vector with zero mean: $\mathbb{E}\,\boldsymbol{x} = \mathbf{0}$. The *covariance matrix* $\boldsymbol{A}$ of the random vector $\boldsymbol{x}$ is the positive-semidefinite matrix

$$\boldsymbol{A} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*) = \sum_{j,k=1}^{p} \mathbb{E}\left(X_j X_k^*\right) \cdot \mathbf{E}_{jk} \tag{1.5.1}$$

The star * refers to the conjugate transpose operation, and the standard basis matrix $\mathbf{E}_{jk}$ has a one in the (j,k) position and zeros elsewhere. In other words, the (j,k) entry of the sample covariance matrix $\boldsymbol{A}$ records the covariance between the jth and kth entry of the vector $\boldsymbol{x}$.

One basic problem in statistical practice is to estimate the covariance matrix from data. Imagine that we have access to n independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, each distributed the same way as $\boldsymbol{x}$. The *sample covariance estimator* $\boldsymbol{Y}$ is the random matrix

$$\boldsymbol{Y} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^*. \tag{1.5.2}$$

The random matrix $\boldsymbol{Y}$ is an unbiased estimator[2] for the sample covariance matrix: $\mathbb{E}\,\boldsymbol{Y} = \boldsymbol{A}$. Observe that the sample covariance estimator $\boldsymbol{Y}$ fits neatly into our paradigm:

[2] The formula (1.5.2) supposes that the random vector $\boldsymbol{x}$ is known to have zero mean. Otherwise, we have to make an adjustment to incorporate an estimate for the sample mean.

Full text available at: http://dx.doi.org/10.1561/2200000048

> **The sample covariance estimator can be expressed as a sum of independent random matrices.**

This is precisely the type of decomposition that allows us to apply the tools in these notes.

1.6 Exponential Concentration Inequalities for Matrices

An important challenge in probability theory is to study the probability that a real random variable Z takes a value substantially different from its mean. That is, we seek a bound of the form

$$\mathbb{P}\left\{|Z - \mathbb{E}\, Z| \geq t\right\} \leq \underline{\quad ??? \quad} \tag{1.6.1}$$

for a positive parameter t. When Z is expressed as a sum of independent random variables, the literature contains many tools for addressing this problem. See [23] for an overview.

For a random matrix $\boldsymbol{Z}$, a variant of (1.6.1) is the question of whether $\boldsymbol{Z}$ deviates substantially from its mean value. We might frame this question as

$$\mathbb{P}\left\{\|\boldsymbol{Z} - \mathbb{E}\, \boldsymbol{Z}\| \geq t\right\} \leq \underline{\quad ??? \quad}. \tag{1.6.2}$$

Here and elsewhere, $\|\cdot\|$ denotes the spectral norm of a matrix, also known as the ℓ_2 operator norm. As noted, it is frequently possible to decompose $\boldsymbol{Z}$ as a sum of independent random matrices. We might even dream that established methods for studying the scalar concentration problem (1.6.1) extend to (1.6.2).

1.6.1 The Bernstein Inequality

To explain what kind of results we have in mind, let us return to the scalar problem (1.6.1). First, to simplify formulas, we assume that the real random variable Z has zero mean: $\mathbb{E}\, Z = 0$. If not, we can center the random variable by subtracting its mean. Second, and more restrictively, we suppose that Z can be expressed as a sum of independent, real random variables.

To control Z, we rely on two types of information: global properties of the sum (such as its mean and variance) and local properties of

Full text available at: http://dx.doi.org/10.1561/2200000048

the summands (such as their maximum fluctuation). These pieces of data are usually easy to obtain. Together, they determine how well Z concentrates around zero, its mean value.

Theorem 1.6.1 (Bernstein Inequality)**.** Let $S_1, \dots, S_n$ be independent, centered, real random variables, and assume that each one is uniformly bounded:

$$\mathbb{E}\, S_k = 0 \quad \text{and} \quad |S_k| \leq L \quad \text{for each } k = 1, \dots, n.$$

Introduce the sum $Z = \sum_{k=1}^{n} S_k$, and let $v(Z)$ denote the variance of the sum:

$$v(Z) = \mathbb{E}\, Z^2 = \sum_{k=1}^{n} \mathbb{E}\, S_k^2.$$

Then

$$\mathbb{P}\,\{|Z| \geq t\} \leq 2 \exp\left(\frac{-t^2/2}{v(Z) + Lt/3} \right) \quad \text{for all } t \geq 0.$$

See [23, §2.8] for a proof of this result. We refer to Theorem 1.6.1 as an *exponential concentration inequality* because it yields exponentially decaying bounds on the probability that Z deviates substantially from its mean.

1.6.2 The Matrix Bernstein Inequality

What is truly astonishing is that the scalar Bernstein inequality, Theorem 1.6.1, lifts directly to matrices. Let us emphasize this remarkable fact:

> **There are exponential concentration inequalities for the spectral norm of a sum of independent random matrices.**

As a consequence, once we decompose a random matrix as an independent sum, we can harness global properties (such as the mean and the variance) and local properties (such as a uniform bound on the summands) to obtain detailed information about the norm of the sum. As in the scalar case, it is usually easy to acquire the input data for the inequality. But the output of the inequality is highly nontrivial.

Full text available at: http://dx.doi.org/10.1561/2200000048

To illustrate these claims, we will state one of the major results from this monograph. This theorem is a matrix extension of Bernstein's inequality that was developed independently in the two papers [138, 183]. After presenting the result, we give some more details about its interpretation. In the next section, we apply this result to study the covariance estimation problem.

Theorem 1.6.2 (Matrix Bernstein)**.** Let $\boldsymbol{S}_1, \dots, \boldsymbol{S}_n$ be independent, centered random matrices with common dimension $d_1 \times d_2$, and assume that each one is uniformly bounded

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le L \quad \text{for each } k = 1, \dots, n.$$

Introduce the sum

$$\boldsymbol{Z} = \sum_{k=1}^{n} \boldsymbol{S}_k, \tag{1.6.3}$$

and let $v(\boldsymbol{Z})$ denote the matrix variance statistic of the sum:

$$\begin{aligned} v(\boldsymbol{Z}) &= \max\left\{ \|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|, \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\| \right\} \\ &= \max\left\{ \left\| \sum_{k=1}^{n} \mathbb{E}\left(\boldsymbol{S}_k\boldsymbol{S}_k^*\right) \right\|, \left\| \sum_{k=1}^{n} \mathbb{E}\left(\boldsymbol{S}_k^*\boldsymbol{S}_k\right) \right\| \right\}. \end{aligned} \tag{1.6.4}$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left\{\|\boldsymbol{Z}\| \ge t\right\} \le (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{v(\boldsymbol{Z}) + Lt/3} \right). \tag{1.6.5}$$

Furthermore,

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{2v(\boldsymbol{Z})\log(d_1 + d_2)} + \frac{1}{3} L \log(d_1 + d_2). \tag{1.6.6}$$

The proof of this result appears in Chapter 6.

To appreciate what Theorem 1.6.2 means, it is valuable to make a direct comparison with the scalar version, Theorem 1.6.1. In both cases, we express the object of interest as an independent sum, and we instate a uniform bound on the summands. There are three salient changes:

- The variance $v(\boldsymbol{Z})$ in the result for matrices can be interpreted as the magnitude of the expected squared deviation of $\boldsymbol{Z}$ from its

Full text available at: http://dx.doi.org/10.1561/2200000048

mean. The formula reflects the fact that a general matrix $\boldsymbol{B}$ has *two* different squares $\boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{B}^*\boldsymbol{B}$. For an Hermitian matrix, the two squares coincide.

- The tail bound has a dimensional factor $d_1 + d_2$ that depends on the size of the matrix. This factor reduces to two in the scalar setting. In the matrix case, it limits the range of t where the tail bound is informative.
- We have included a bound for $\mathbb{E}\,\|\boldsymbol{Z}\|$. This estimate is not particularly interesting in the scalar setting, but it is usually quite challenging to prove results of this type for matrices. In fact, the expectation bound is often more useful than the tail bound.

The latter point deserves amplification:

> **The expectation bound** (1.6.6) **is the most important aspect of the matrix Bernstein inequality.**

For further discussion of this result, turn to Chapter 6. Chapters 4 and 7 contain related results and interpretations.

1.6.3 Example: The Sample Covariance Estimator

We will apply the matrix Bernstein inequality, Theorem 1.6.2, to measure how well the sample covariance estimator approximates the true covariance matrix. As before, let $\boldsymbol{x}$ be a zero-mean random vector with dimension p. Introduce the $p \times p$ covariance matrix $\boldsymbol{A} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*)$. Suppose we have n independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ with the same distribution as $\boldsymbol{x}$. Form the $p \times p$ sample covariance estimator

$$\boldsymbol{Y} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^*.$$

Our goal is to study how the spectral-norm distance $\|\boldsymbol{Y} - \boldsymbol{A}\|$ between the sample covariance and the true covariance depends on the number n of samples.

For simplicity, we will perform the analysis under the extra assumption that the ℓ_2 norm of the random vector is bounded: $\|\boldsymbol{x}\|^2 \leq B$. This

Full text available at: http://dx.doi.org/10.1561/2200000048

hypothesis can be relaxed if we apply a variant of the matrix Bernstein inequality that reflects the typical magnitude of a summand $\boldsymbol{S}_k$. One such variant appears in the formula (6.1.6).

We are in a situation where it is quite easy to see how the matrix Bernstein inequality applies. Define the random deviation $\boldsymbol{Z}$ of the estimator $\boldsymbol{Y}$ from the true covariance matrix $\boldsymbol{A}$:

$$\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{A} = \sum_{k=1}^{n} \boldsymbol{S}_k \quad \text{where} \quad \boldsymbol{S}_k = \frac{1}{n}(\boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{A}).$$

The random matrices $\boldsymbol{S}_k$ are independent, identically distributed, and centered. To apply Theorem 1.6.2, we need to find a uniform bound L for the summands, and we need to control the matrix variance statistic $v(\boldsymbol{Z})$.

First, let us develop a uniform bound on the spectral norm of each summand. We may calculate that

$$\|\boldsymbol{S}_k\| = \frac{1}{n}\,\|\boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{A}\| \leq \frac{1}{n}\big(\,\|\boldsymbol{x}_k \boldsymbol{x}_k^*\| + \|\boldsymbol{A}\|\,\big) \leq \frac{2B}{n}.$$

The first relation is the triangle inequality. The second follows from the assumption that $\boldsymbol{x}$ is bounded and the observation that

$$\|\boldsymbol{A}\| = \|\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*)\| \leq \mathbb{E}\,\|\boldsymbol{x}\boldsymbol{x}^*\| = \mathbb{E}\,\|\boldsymbol{x}\|^2 \leq B.$$

This expression depends on Jensen's inequality and the hypothesis that $\boldsymbol{x}$ is bounded.

Second, we need to bound the matrix variance statistic $v(\boldsymbol{Z})$ defined in (1.6.4). The matrix $\boldsymbol{Z}$ is Hermitian, so the two squares in this formula coincide with each other:

$$v(\boldsymbol{Z}) = \|\mathbb{E}\,\boldsymbol{Z}^2\| = \left\|\sum_{k=1}^{n} \mathbb{E}\,\boldsymbol{S}_k^2\right\|.$$

We need to determine the variance of each summand. By direct calculation,

$$\begin{aligned}\mathbb{E}\,\boldsymbol{S}_k^2 &= \frac{1}{n^2}\,\mathbb{E}\,(\boldsymbol{x}_k\boldsymbol{x}_k^* - \boldsymbol{A})^2 \\ &= \frac{1}{n^2}\,\mathbb{E}\left[\,\|\boldsymbol{x}_k\|^2 \cdot \boldsymbol{x}_k\boldsymbol{x}_k^* - (\boldsymbol{x}_k\boldsymbol{x}_k^*)\boldsymbol{A} - \boldsymbol{A}(\boldsymbol{x}_k\boldsymbol{x}_k^*) + \boldsymbol{A}^2\right]\end{aligned}$$

Full text available at: http://dx.doi.org/10.1561/2200000048

$$\preccurlyeq \frac{1}{n^2}\left[B \cdot \mathbb{E}\left(\boldsymbol{x}_k\boldsymbol{x}_k^*\right) - \boldsymbol{A}^2 - \boldsymbol{A}^2 + \boldsymbol{A}^2\right]$$
$$\preccurlyeq \frac{B}{n^2} \cdot \boldsymbol{A}$$

The expression $\boldsymbol{H} \preccurlyeq \boldsymbol{T}$ means that $\boldsymbol{T} - \boldsymbol{H}$ is positive semidefinite. We used the norm bound for the random vector $\boldsymbol{x}$ and the fact that expectation preserves the semidefinite order. In the last step, we dropped the negative-semidefinite term $-\boldsymbol{A}^2$. Summing this relation over k, we reach

$$\mathbf{0} \preccurlyeq \sum_{k=1}^{n} \mathbb{E}\, \boldsymbol{S}_k^2 \preccurlyeq \frac{B}{n} \cdot \boldsymbol{A}.$$

The sum is a positive-semidefinite matrix because $\boldsymbol{S}_k^2$ is positive semidefinite. Extract the spectral norm to arrive at

$$v(\boldsymbol{Z}) = \left\| \sum_{k=1}^{n} \mathbb{E}\, \boldsymbol{S}_k^2 \right\| \leq \frac{B\,\|\boldsymbol{A}\|}{n}.$$

We have now collected the information we need to analyze the sample covariance estimator.

We can invoke the estimate (1.6.6) from the matrix Bernstein inequality, Theorem 1.6.2, with the uniform bound $L = 2B/n$ and the variance bound $v(\boldsymbol{Z}) \leq B\,\|\boldsymbol{A}\|/n$. We attain

$$\mathbb{E}\,\|\boldsymbol{Y} - \boldsymbol{A}\| = \mathbb{E}\,\|\boldsymbol{Z}\| \leq \sqrt{\frac{2B\,\|\boldsymbol{A}\|\log(2p)}{n}} + \frac{2B\log(2p)}{3n}.$$

In other words, the error in approximating the sample covariance matrix is not too large when we have a sufficient number of samples. If we wish to obtain a relative error on the order of ε, we may take

$$n \geq \frac{2B\log(2p)}{\varepsilon^2\,\|\boldsymbol{A}\|}.$$

This selection yields

$$\mathbb{E}\,\|\boldsymbol{Y} - \boldsymbol{A}\| \leq (\varepsilon + \varepsilon^2) \cdot \|\boldsymbol{A}\|.$$

It is often the case that $\|\boldsymbol{A}\| = \text{Const}$ and $B = \text{Const} \cdot p$, so we discover that $n = \text{Const} \cdot \varepsilon^{-2} p \log p$ samples are sufficient for the sample covariance estimator to provide a relatively accurate estimate of the true

Full text available at: http://dx.doi.org/10.1561/2200000048

covariance matrix $\boldsymbol{A}$. This bound is qualitatively sharp for worst-case distributions.

The analysis in this section applies to many other examples. We encapsulate the argument in Corollary 6.2.1, which we use to study several more problems.

1.6.4 History of this Example

Covariance estimation may be the earliest application of matrix concentration tools in random matrix theory. Mark Rudelson [158], building on a suggestion of Gilles Pisier, showed how to use the noncommutative Khintchine inequality [116, 117, 29, 30] to obtain essentially optimal bounds on the sample covariance estimator of a bounded random vector. The tutorial [188] of Roman Vershynin offers an overview of this problem as well as many results and references. The analysis of the sample covariance matrix here is adapted from the technical report [72]. It leads to a result similar with the one Rudelson obtained in [158].

1.6.5 Optimality of the Matrix Bernstein Inequality

Theorem 1.6.2 can be sharpened very little because it applies to every random matrix $\boldsymbol{Z}$ of the form (1.6.3). Let us say a few words about optimality now, postponing the details to §6.1.2.

Suppose that $\boldsymbol{Z}$ is a random matrix of the form (1.6.3) where the summands $\boldsymbol{S}_k$ are independent and have zero mean. Introduce the quantity

$$L_\star^2 = \mathbb{E}\,\max_k \|\boldsymbol{S}_k\|^2 .$$

In §6.1.2, we will argue that these assumptions imply

$$\begin{aligned} \text{const} \cdot \left[v(\boldsymbol{Z}) + L_\star^2\right] \quad &\leq \quad \mathbb{E}\,\|\boldsymbol{Z}\|^2 \\ &\leq \quad \text{Const} \cdot \left[v(\boldsymbol{Z}) \log(d_1+d_2) + L_\star^2 \log^2(d_1+d_2)\right]. \end{aligned} \tag{1.6.7}$$

In other words, the scale of $\mathbb{E}\,\|\boldsymbol{Z}\|^2$ must depend on the matrix variance statistic $v(\boldsymbol{Z})$ and the average upper bound $L_\star^2$ for the summands. The quantity $L = \sup \|\boldsymbol{S}_k\|$ that appears in the matrix Bernstein inequality always exceeds $L_\star$, sometimes by a large margin, but they capture the same type of information.

Full text available at: http://dx.doi.org/10.1561/2200000048

The difference between the lower and upper bound in (1.6.7) comes from the dimensional factor $\log(d_1 + d_2)$. There are random matrices $\boldsymbol{Z}$ for which the lower bound gives a more accurate reflection of $\mathbb{E}\,\|\boldsymbol{Z}\|^2$, but there are also many random matrices where the upper bound describes the behavior correctly. At present, we do not understand how to distinguish between the two extremes, but some recent progress appears in [186].

The tail bound (1.6.5) provides a useful tool in practice, but it is not necessarily the best way to collect information about large deviation probabilities. To obtain more precise results, we recommend using the expectation bound (1.6.6) to control $\mathbb{E}\,\|\boldsymbol{Z}\|$ and then applying scalar concentration inequalities to estimate $\mathbb{P}\,\{\|\boldsymbol{Z}\| \geq \mathbb{E}\,\|\boldsymbol{Z}\| + t\}$. The book [23] offers a good treatment of the methods that are available for establishing scalar concentration.

1.7 The Arsenal of Results

The Bernstein inequality is probably the most familiar exponential tail bound for a sum of independent random variables, but there are many more. It turns out that essentially all of these scalar results admit extensions that hold for random matrices. In fact, many of the established techniques for scalar concentration have analogs in the matrix setting.

1.7.1 What's Here...

This monograph focuses on a few key exponential concentration inequalities for a sum of independent random matrices, and it describes some specific applications of these results.

Matrix Gaussian Series. A matrix Gaussian series is a random matrix that can be expressed as a sum of fixed matrices, each weighted by an independent standard normal random variable. This formulation includes a surprising number of examples. The most important are undoubtedly Wigner matrices and rectangular Gaussian matrices. Another interesting case is a Toeplitz matrix with Gaussian entries. The analysis of matrix Gaussian series appears in Chapter 4.

Full text available at: http://dx.doi.org/10.1561/2200000048

Matrix Rademacher Series. A matrix Rademacher series is a random matrix that can be written as a sum of fixed matrices, each weighted by an independent Rademacher random variable.[3] This construction includes things like random sign matrices, as well as a fixed matrix whose entries are modulated by random signs. There are also interesting examples that arise in combinatorial optimization. We treat these problems in Chapter 4.

Matrix Chernoff Bounds. The matrix Chernoff bounds apply to a random matrix that can be decomposed as a sum of independent, random positive-semidefinite matrices whose maximum eigenvalues are subject to a uniform bound. These results allow us to obtain information about the norm of a random submatrix drawn from a fixed matrix. They are also appropriate for studying the Laplacian matrix of a random graph. See Chapter 5.

Matrix Bernstein Bounds. The matrix Bernstein inequality concerns a random matrix that can be expressed as a sum of independent, centered random matrices that admit a uniform spectral-norm bound. This result has many applications, including the analysis of randomized algorithms for matrix sparsification and matrix multiplication. It can also be used to study the random features paradigm for approximating a kernel matrix. Chapter 6 contains this material.

Intrinsic Dimension Bounds. Some matrix concentration inequalities can be improved when the random matrix has limited spectral content in most dimensions. In this situation, we may be able to obtain bounds that do not depend on the ambient dimension. See Chapter 7 for details.

We have chosen to present these results because they are illustrative, and they have already found concrete applications.

[3] A *Rademacher random variable* takes the two values ±1 with equal probability.

Full text available at: http://dx.doi.org/10.1561/2200000048

1.7.2 What's Not Here...

The program of extending scalar concentration results to the matrix setting has been quite fruitful, and there are many useful results beyond the ones that we detail. Let us mention some of the other tools that are available. For further information, see the annotated bibliography.

First, there are additional exponential concentration inequalities for a sum of independent random matrices. All of the following results can be established within the framework of this monograph.

Matrix Hoeffding. This result concerns a sum of independent random matrices whose squares are subject to semidefinite upper bounds [183, §7].

Matrix Bennett. This estimate sharpens the tail bound from the matrix Bernstein inequality [183, §6].

Matrix Bernstein, Unbounded Case. The matrix Bernstein inequality extends to the case where the moments of the summands grow at a controlled rate. See [183, §6] or [101].

Matrix Bernstein, Nonnegative Summands. The lower tail of the Bernstein inequality can be improved when the summands are positive semidefinite [124]; this result extends to the matrix setting. By a different argument, the dimensional factor can be removed from this bound for a class of interesting examples [141, Thm. 3.1].

The approach in this monograph can be adapted to obtain exponential concentration for matrix-valued martingales. Here are a few results from this category:

Matrix Azuma. This is the martingale version of the matrix Hoeffding bound [183, §7].

Matrix Bounded Differences. The matrix Azuma inequality gives bounds for the spectral norm of a matrix-valued function of independent random variables [183, §7].

Full text available at: http://dx.doi.org/10.1561/2200000048

Matrix Freedman. This result can be viewed as the martingale extension of the matrix Bernstein inequality [138, 180].

The technical report [182] explains how to extend other bounds for a sum of independent random matrices to the martingale setting.

Polynomial moment inequalities provide bounds for the expected trace of a power of a random matrix. Moment inequalities for a sum of independent random matrices can provide useful information when the summands have heavy tails or else a uniform bound does not reflect the typical size of the summands.

Matrix Khintchine. The matrix Khintchine inequality is the polynomial version of the exponential bounds for matrix Gaussian series and matrix Rademacher series. This result is presented in (4.7.1). See the papers [116, 29, 30] or [118, Cor. 7.3] for proofs.

Matrix Moment Inequalities. The matrix Chernoff inequality admits a polynomial variant; the simplest form appears in (5.1.9). The matrix Bernstein inequality also has a polynomial variant, stated in (6.1.6). These bounds are drawn from [40, App.].

The methods that lead to polynomial moment inequalities differ substantially from the techniques in this monograph, so we cannot include the proofs here. The annotated bibliography includes references to the large literature on moment inequalities for random matrices.

Recently, Lester Mackey and the author, in collaboration with Daniel Paulin and several other researchers [118, 143], have developed another framework for establishing matrix concentration. This approach extends a scalar argument, introduced by Chatterjee [38, 39], that depends on exchangeable pairs and Markov chain couplings. The *method of exchangeable pairs* delivers both exponential concentration inequalities and polynomial moment inequalities for random matrices, and it can reproduce many of the bounds mentioned above. It also leads to new results:

Polynomial Efron–Stein Inequality for Matrices. This bound is a matrix version of the polynomial Efron–Stein inequality [21,

Full text available at: http://dx.doi.org/10.1561/2200000048

Thm. 1]. It controls the polynomial moments of a centered random matrix that is a function of independent random variables [143, Thm. 4.2].

Exponential Efron–Stein Inequality for Matrices. This bound is the matrix extension of the exponential Efron–Stein inequality [22, Thm. 1]. It leads to exponential concentration inequalities for a centered random matrix constructed from independent random variables [143, Thm. 4.3].

Another significant advantage is that the method of exchangeable pairs can sometimes handle random matrices built from dependent random variables. Although the simplest version of the exchangeable pairs argument is more elementary than the approach in this monograph, it takes a lot of effort to establish the more useful inequalities. With some regret, we have chosen not to include this material because the method and results are accessible to a narrower audience.

Finally, we remark that the modified logarithmic Sobolev inequalities of [22, 21] also extend to the matrix setting [42]. Unfortunately, the matrix variants do not seem to be as useful as the scalar results.

In addition to the matrix concentration tools mentioned above, the field of random matrix theory offers a wide range of other methods. For an introduction to some of the major ideas in this area, see the book [173].

1.8 About This Monograph

This monograph is intended for graduate students and researchers in computational mathematics who want to learn some modern techniques for analyzing random matrices. The preparation required is minimal. We assume familiarity with calculus, applied linear algebra, the basic theory of normed spaces, and classical probability theory up through the elementary concentration inequalities (such as Markov and Bernstein). Beyond the basics, which can be gleaned from any good textbook, we include all the required background in Chapter 2.

Full text available at: http://dx.doi.org/10.1561/2200000048

The material here is based primarily on the paper "User-Friendly Tail Bounds for Sums of Random Matrices" by the present author [183]. There are several significant revisions to this earlier work:

Examples and Applications. Many of the papers on matrix concentration give limited information about how the results can be used to solve problems of interest. A major part of these notes consists of worked examples and applications that indicate how matrix concentration inequalities apply to practical questions.

Expectation Bounds. This work collects bounds for the expected value of the spectral norm of a random matrix and bounds for the expectation of the smallest and largest eigenvalues of a random symmetric matrix. Some of these useful results have appeared piecemeal in the literature [40, 118], but they have not been included in a unified presentation.

Optimality. We explain why each matrix concentration inequality is (nearly) optimal. This presentation includes examples to show that each term in each bound is necessary to describe some particular phenomenon.

Intrinsic Dimension Bounds. Over the last few years, there have been some refinements to the basic matrix concentration bounds that improve the dependence on dimension [91, 130]. We describe a new framework that allows us to prove these results with ease.

Lieb's Theorem. The matrix concentration inequalities in this monograph depend on a deep theorem [109, Thm. 6] from matrix analysis due to Elliott Lieb. We provide a complete proof of this result, along with all the background required to understand the argument.

Annotated Bibliography. We have included a list of the major works on matrix concentration, including a short summary of the main contributions of these papers. We hope this catalog will be a valuable guide for further reading.

Full text available at: http://dx.doi.org/10.1561/2200000048

The organization of the notes is straightforward. Chapter 2 contains background material that is needed for the analysis. Chapter 3 describes the framework for developing exponential concentration inequalities for matrices. Chapter 4 presents the first set of results and examples, concerning matrix Gaussian and Rademacher series. Chapter 5 introduces the matrix Chernoff bounds and their applications, and Chapter 6 expands on our discussion of the matrix Bernstein inequality. Chapter 7 shows how to sharpen some of the results so that they depend on an intrinsic dimension parameter. Chapter 8 contains the proof of Lieb's theorem. We conclude with resources on matrix concentration and a bibliography.

To make the presentation smoother, we have not followed all of the conventions for scholarly articles in journals. In particular, almost all the citations appear in the notes at the end of each chapter. Our aim has been to explain the ideas as clearly as possible, rather than to interrupt the narrative with an elaborate genealogy of results.

Full text available at: http://dx.doi.org/10.1561/2200000048

References

[1] E. Abbé and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. Available at `http://arXiv.org/abs/1503.00609`, Mar. 2015.

[2] D. Achlioptas, Z. Karnin, and E. Liberty. Near-optimal entrywise sampling for data matrices. In *Advances in Neural Information Processing Systems 26*, 2013.

[3] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 611–618 (electronic). ACM, New York, 2001.

[4] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *J. Assoc. Comput. Mach.*, 54(2):Article 10, 2007. (electronic).

[5] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, Mar. 2002.

[6] N. Ailon and B. Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[7] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B*, 28:131–142, 1966.

Full text available at: http://dx.doi.org/10.1561/2200000048

[8] N. Alon and J. H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. With an appendix on the life and work of Paul Erdős.

[9] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *Inform. Inference*, 3(3):224–294, 2014. Preprint available at `http://arXiv.org/abs/1303.6672`.

[10] T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra Appl.*, 26:203–241, 1979.

[11] S. Arora, E. Hazan, and S. Kale. A fast random sampling algorithm for sparsifying matrices. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 272–279, 2006.

[12] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[13] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[14] A. S. Bandeira and R. Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Available at `http://arXiv.org/abs/1408.6185`, Aug. 2014.

[15] S. Barman. An approximate version of Carathéodory's theorem with applications to approximating Nash equilibria and dense bipartite subgraphs. Available at `http://arXiv.org/abs/1406.2296`, June 2014.

[16] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, May 1993.

[17] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2001. Analysis, algorithms, and engineering applications.

[18] J. Bendat and S. Sherman. Monotone and convex operator functions. *Trans. Amer. Math. Soc.*, 79:58–71, 1955.

[19] R. Bhatia. *Matrix Analysis*. Number 169 in Graduate Texts in Mathematics. Springer, Berlin, 1997.

Full text available at: http://dx.doi.org/10.1561/2200000048

[20] R. Bhatia. *Positive Definite Matrices.* Princeton Univ. Press, Princeton, NJ, 2007.

[21] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005.

[22] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003.

[23] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities.* Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[24] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.

[25] J. Bourgain and L. Tzafriri. Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math.*, 57(2):137–224, 1987.

[26] J. Bourgain and L. Tzafriri. On a problem of Kadison and Singer. *J. Reine Angew. Math.*, 420:1–43, 1991.

[27] L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

[28] W. Bryc, A. Dembo, and T. Jiang. Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Ann. Probab.*, 34(1):1–38, 2006.

[29] A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Math. Ann.*, 319:1–16, 2001.

[30] A. Buchholz. Optimal constants in Khintchine-type inequalities for Fermions, Rademachers and q-Gaussian operators. *Bull. Pol. Acad. Sci. Math.*, 53(3):315–321, 2005.

[31] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[32] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[33] B. Carl. Inequalities of Bernstein–Jackson-type and the degree of compactness in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.

Full text available at: http://dx.doi.org/10.1561/2200000048

[34] E. Carlen. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010.

[35] E. A. Carlen and E. H. Lieb. A Minkowski type trace inequality and strong subadditivity of quantum entropy. II. Convexity and concavity. *Lett. Math. Phys.*, 83(2):107–126, 2008.

[36] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[37] V. Chandrasekaran, S. Sanghavi, P. A. Parillo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. Available at `http://arXiv.org/abs/0906.2220`, Jun. 2009.

[38] S. Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)–Stanford University.

[39] S. Chatterjee. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138:305–321, 2007.

[40] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis using matrix concentration inequalities. *Inform. Inference*, 1(1), 2012. `doi:10.1093/imaiai/ias001`.

[41] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis using matrix concentration inequalities. ACM Report 2012-01, California Inst. Tech., Pasadena, CA, Feb. 2012.

[42] R. Y. Chen and J. A. Tropp. Subadditivity of matrix φ-entropy and concentration of random matrices. *Electron. J. Probab.*, 19(27):1–30, 2014.

[43] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952.

[44] S. Chrétien and S. Darses. Invertibility of random submatrices via tail-decoupling and a matrix Chernoff inequality. *Statist. Probab. Lett.*, 82(7):1479–1487, 2012.

[45] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 81–90. ACM, New York, 2013.

[46] A. Connes and E. Størmer. Entropy for automorphisms of II_1 von Neumann algebras. *Acta Math.*, 134(3-4):289–306, 1975.

Full text available at: http://dx.doi.org/10.1561/2200000048

[47] D. Cristofides and K. Markström. Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures Algs.*, 32(8):88–100, 2008.

[48] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[49] A. d'Asprémont. Subsampling algorithms for semidefinite programming. *Stoch. Syst.*, 1(2):274–305, 2011.

[50] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of Banach Space Geometry*, pages 317–366. Elsevier, Amsterdam, 2002.

[51] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM J. Matrix Anal. Appl.*, 29(4):1120–1146, 2007.

[52] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, Apr. 2006.

[53] P. Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD, 1999)*, pages 291–299. ACM, New York, 1999.

[54] P. Drineas and R. Kannan. Fast Monte Carlo algorithms for approximate matrix multiplication. In *Proc. 42nd IEEE Symp. Foundations of Computer Science (FOCS)*, pages 452–259, 2001.

[55] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices. I. Approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.

[56] P. Drineas and M. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, 2005.

[57] P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Inform. Process. Lett.*, 111(8):385–389, 2011.

[58] A. Ebadian, I. Nikoufar, and M. E. Gordji. Perspectives of matrix convex functions. *Proc. Natl. Acad. Sci. USA*, 108(18):7313–7314, 2011.

[59] E. G. Effros. A matrix convexity approach to some celebrated quantum inequalities. *Proc. Natl. Acad. Sci. USA*, 106(4):1006–1008, Jan. 2009.

Full text available at: http://dx.doi.org/10.1561/2200000048

[60] H. Epstein. Remarks on two theorems of E. Lieb. *Comm. Math. Phys.*, 31:317–325, 1973.

[61] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[62] W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[63] W. Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

[64] X. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École d'Été de Probabilités de Saint-Flour, IV-1974*, pages 1–96. Lecture Notes in Math., Vol. 480. Springer, Berlin, 1975.

[65] P. J. Forrester. *Log-gases and random matrices*, volume 34 of *London Mathematical Society Monographs Series.* Princeton University Press, Princeton, NJ, 2010.

[66] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing.* Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[67] D. A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, Feb. 1975.

[68] A. Frieze, R. Kannan, and S. Vempala. Fast Monte Carlo algorithms for finding low-rank approximations. In *Proc. 39th Ann. IEEE Symp. Foundations of Computer Science (FOCS)*, pages 370–378, 1998.

[69] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss. Near-optimal sparse Fourier representations via sampling. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pages 152–161. ACM, New York, 2002.

[70] A. Gittens and M. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 2014. To appear. Preprint available at `http://arXiv.org/abs/1303.1849`.

[71] A. Gittens and J. A. Tropp. Error bounds for random matrix approximation schemes. ACM Report 2014-01, California Inst. Tech., Nov. 2009. Available at `http://arXiv.org/abs/0911.4108`.

[72] A. Gittens and J. A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. ACM Report 2014-02, California Inst. Tech., 2014. Available at `http://arXiv.org/abs/1104.4513`.

Full text available at: http://dx.doi.org/10.1561/2200000048

[73] C. Godsil and G.F. Royle. *Algebraic Graph Theory.* Number 207 in Graduate Texts in Mathematics. Springer, 2001.

[74] H. H. Goldstine and J. von Neumann. Numerical inverting of matrices of high order. II. *Proc. Amer. Math. Soc.*, 2:188–202, 1951.

[75] Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50(4):265–289, 1985.

[76] J. F. Grcar. John von Neumann's analysis of Gaussian elimination and the origins of modern numerical analysis. *SIAM Rev.*, 53(4):607–682, 2011.

[77] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes.* Oxford University Press, New York, third edition, 2001.

[78] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, Mar. 2011.

[79] D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. Available at `http://arXiv.org/abs/1001.2738`.

[80] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, June 2011.

[81] R. Hamid, Y. Xiao, A. Gittens, and D. DeCoste. Compact random feature maps. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, July 2014.

[82] F. Hansen and G. K. Pedersen. Jensen's inequality for operators and Löwner's theorem. *Math. Ann.*, 258(3):229–241, 1982.

[83] F. Hansen and G. K. Pedersen. Jensen's operator inequality. *Bull. London Math. Soc.*, 35(4):553–564, 2003.

[84] M. B. Hastings. Superadditivity of communication complexity using entangled inputs. *Nature Phys.*, 5:255–257, 2009.

[85] P. Hayden and A. Winter. Counterexamples to the maximal p-norm multiplicity conjecture for all $p > 1$. *Comm. Math. Phys.*, 284(1):263–280, 2008.

[86] F. Hiai and D. Petz. *Introduction to Matrix Analysis and Applications.* Springer, Feb. 2014.

[87] N. J. Higham. *Functions of Matrices: Theory and Computation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

Full text available at: http://dx.doi.org/10.1561/2200000048

[88] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[89] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[90] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge Univ. Press, 2nd edition, 2013.

[91] D. Hsu, S. M. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.*, 17:no. 14, 13, 2012.

[92] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

[93] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities. *Ann. Probab.*, 31(2):948–995, 2003.

[94] M. Junge and Q. Xu. On the best constants in some non-commutative martingale inequalities. *Bull. London Math. Soc.*, 37:243–253, 2005.

[95] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities II: Applications. *Israel J. Math.*, 167:227–282, 2008.

[96] M. Junge and Q. Zeng. Noncommutative Bennett and Rosenthal inequalities. *Ann. Probab.*, 41(6):4287–4316, 2013.

[97] M. Junge and Q. Zeng. Noncommutative martingale deviation and Poincaré type inequalities with applications. *Probab. Theory Related Fields*, Feb. 2014. Preprint available at `http://arXiv.org/abs/1211.3209`.

[98] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *Proc. 15th Intl. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2012.

[99] B. Kashin and L. Tzafriri. Some remarks on coordinate restriction of operators to coordinate subspaces. Insitute of Mathematics Preprint 12, Hebrew University, Jerusalem, 1993–1994.

[100] T. Kemp. Math 247a: Introduction to random matrix theory. Available at `http://www.math.ucsd.edu/~tkemp/247A/247A.Notes.pdf`, 2013.

Full text available at: http://dx.doi.org/10.1561/2200000048

[101] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[102] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. Available at `http://arXiv.org/abs/1312.3580`, Dec. 2013.

[103] F. Kraus. Über konvexe Matrixfunktionen. *Math. Z.*, 41(1):18–42, 1936.

[104] F. Kubo and T. Ando. Means of positive linear operators. *Math. Ann.*, 246(3):205–224, 1979/80.

[105] A. Kundu and P. Drineas. A note on randomized element-wise matrix sparsification. Available at `http://arXiv.org/abs/1404.0320`, Apr. 2014.

[106] R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282, 2005.

[107] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

[108] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6):2118–2132, Nov. 1996.

[109] E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.*, 11:267–288, 1973.

[110] E. H. Lieb and R. Seiringer. Stronger subadditivity of entropy. *Phys. Rev. A*, 71:062329–1–9, 2005.

[111] G. Lindblad. Entropy, information and quantum measurements. *Comm. Math. Phys.*, 33:305–322, 1973.

[112] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[113] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf. Randomized nonlinear component analysis. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, July 2014.

[114] K. Löwner. Über monotone Matrixfunktionen. *Math. Z.*, 38(1):177–216, 1934.

[115] G. Lugosi. Concentration-of-measure inequalities. Available at `http://www.econ.upf.edu/~lugosi/anu.pdf`, 2009.

Full text available at: http://dx.doi.org/10.1561/2200000048

[116] F. Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, 303(7):289–292, 1986.

[117] F. Lust-Piquard and G. Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

[118] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangable pairs. *Ann. Probab.*, 42(3):906–945, 2014. Preprint available at `http://arXiv.org/abs/1201.6002`.

[119] A. Magen and A. Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1422–1436. SIAM, Philadelphia, PA, 2011.

[120] M. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learning*, 3(2):123–224, Feb. 2011.

[121] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967.

[122] A. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison–Singer problem. *Ann. Math.*, June 2014. To appear. Preprint available at `http://arXiv.org/abs/1306.3969`.

[123] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis.* Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont., 1979. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

[124] A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *JIPAM. J. Inequal. Pure Appl. Math.*, 4(1):Article 15, 6 pp. (electronic), 2003.

[125] M. McCoy and J. A. Tropp. Sharp recovery thresholds for convex deconvolution, with applications. *Found. Comput. Math.*, Apr. 2014. Preprint available at `http://arXiv.org/abs/1205.1580`.

[126] M. B. McCoy and J. A. Tropp. The achievable performance of convex demixing. Available at `http://arXiv.org/abs/1309.7478`, Sep. 2013.

[127] M. W. Meckes. On the spectral norm of a random Toeplitz matrix. *Electron. Comm. Probab.*, 12:315–325 (electronic), 2007.

[128] M. L. Mehta. *Random matrices*, volume 142 of *Pure and Applied Mathematics (Amsterdam).* Elsevier/Academic Press, Amsterdam, third edition, 2004.

Full text available at: http://dx.doi.org/10.1561/2200000048

[129] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971.

[130] S. Minsker. Some extensions of Bernstein's inequality for self-adjoint operators. Available at `http://arXiv.org/abs/1112.5448`, Nov. 2011.

[131] H. L. Montgomery. The pair correlation of zeros of the zeta function. In *Analytic number theory (Proc. Sympos. Pure Math., Vol. XXIV, St. Louis Univ., St. Louis, Mo., 1972)*, pages 181–193. Amer. Math. Soc., Providence, R.I., 1973.

[132] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge Univ. Press, Cambridge, 1995.

[133] R. J. Muirhead. *Aspects of multivariate statistical theory.* John Wiley & Sons Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[134] A. Naor, O. Regev, and T. Vidick. Efficient rounding for the noncommutative Grothendieck inequality (extended abstract). In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 71–80. ACM, New York, 2013.

[135] D. Needell and J. A. Tropp. Paved with good intentions: analysis of a randomized block Kaczmarz method. *Linear Algebra Appl.*, 441:199–221, 2014.

[136] A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Prog. Ser. B*, 109:283–317, 2007.

[137] A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*, volume 335 of *London Mathematical Society Lecture Note Series.* Cambridge University Press, Cambridge, 2006.

[138] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at `http://arXiv.org/abs/0911.0600`, Feb. 2010.

[139] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15:203–212, 2010.

[140] R. I. Oliveira. The spectrum of random k-lifts of large graphs (with possibly large k). *J. Combinatorics*, 1(3/4):285–306, 2011.

[141] R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. Available at `http://arXiv.org/abs/1312.2903`, Dec. 2013.

Full text available at: http://dx.doi.org/10.1561/2200000048

[142] B. N. Parlett. *The symmetric eigenvalue problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[143] D. Paulin, L. Mackey, and J. A. Tropp. Efron–Stein inequalities for random matrices. Available at `http://arXiv.org/abs/1408.3470`, Aug. 2014.

[144] D. Petz. Quasi-entropies for finite quantum systems. *Rep. Math. Phys.*, 23(1):57–65, 1986.

[145] D. Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory (Warsaw, 1992)*, volume 30 of *Banach Center Publ.*, pages 287–298. Polish Acad. Sci., Warsaw, 1994.

[146] D. Petz. From f-divergence to quantum quasi-entropies and their use. *Entropy*, 12(3):304–325, 2010.

[147] D. Petz. Matrix analysis with some applications. Available at `bolyai.cs.elte.hu/~petz/matrixbme.pdf`, Feb. 2011.

[148] I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

[149] G. Pisier. Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis, 1980–1981*, pages Exp. No. V, 13. École Polytech., Palaiseau, 1981.

[150] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[151] W. Pusz and S. L. Woronowicz. Functional calculus for sesquilinear forms and the purification map. *Rep. Mathematical Phys.*, 8(2):159–170, 1975.

[152] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184, Vancouver, Dec. 2007.

[153] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, 2008.

[154] B. Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, 2011.

[155] M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12:731–817, 2011.

Full text available at: http://dx.doi.org/10.1561/2200000048

[156] S. Riemer and C. Schütt. On the expectation of the norm of random matrices with non-identically distributed entries. *Electron. J. Probab.*, 18:no. 29, 13, 2013.

[157] H. P. Rosenthal. On subspaces of L_p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

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

[159] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proc. 40th Ann. Conf. Information Sciences and Systems (CISS)*, Mar. 2006.

[160] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. Assoc. Comput. Mach.*, 54(4):Article 21, 19 pp., Jul. 2007. (electronic).

[161] M. B. Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *J. Math. Phys.*, 43(9):4358–4375, Sep. 2002.

[162] M. B. Ruskai. Erratum: Inequalities for quantum entropy: A review with conditions for equality [*J. Math. Phys.* 43, 4358 (2002)]. *J. Math. Phys.*, 46(1):0199101, 2005.

[163] A. Sankar, D. A. Spielman, and S.-H. Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Anal. Appl.*, 28(2):446–476, 2006.

[164] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proc. 47th Ann. IEEE Symp. Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[165] B. Schölkopf and S. Smola. *Learning with Kernels.* MIT Press, 1998.

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

[167] A. Sen and B. Virág. The top eigenvalue of the random Toeplitz matrix and the sine kernel. *Ann. Probab.*, 41(6):4050–4079, 2013.

[168] S. Shalev-Shwartz and N. Srebro. Low ℓ_1-norm and guarantees on sparsifiability. In *ICML/COLT/UAI Sparse Optimization and Variable Selection Workshop*, July 2008.

[169] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, second edition, 1996. Translated from the first (1980) Russian edition by R. P. Boas.

[170] A. M.-C. So. Moment inequalities for sums of random matrices and their applications in optimization. *Math. Prog. Ser. A*, Dec. 2009.

Full text available at: http://dx.doi.org/10.1561/2200000048

[171] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 81–90 (electronic), New York, 2004. ACM.

[172] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, Berkeley, 1972. Univ. California Press.

[173] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[174] W. Thirring. *Quantum mathematical physics*. Springer-Verlag, Berlin, second edition, 2002. Atoms, molecules and large systems, Translated from the 1979 and 1980 German originals by Evans M. Harrell II.

[175] N. Tomczak-Jaegermann. 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.

[176] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Math. Acad. Sci. Paris*, 346(23-24):1271–1274, 2008.

[177] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25:1–24, 2008.

[178] J. A. Tropp. On the linear independence of spikes and sines. *J. Fourier Anal. Appl.*, 14(5-6):838–858, 2008.

[179] J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008.

[180] J. A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011.

[181] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1-2):115–126, 2011.

[182] J. A. Tropp. User-friendly tail bounds for matrix martingales. ACM Report 2011-01, California Inst. Tech., Pasadena, CA, Jan. 2011.

[183] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, August 2011.

[184] J. A. Tropp. From joint convexity of quantum relative entropy to a concavity theorem of Lieb. *Proc. Amer. Math. Soc.*, 140(5):1757–1760, 2012.

Full text available at: http://dx.doi.org/10.1561/2200000048

[185] J. A. Tropp. Convex recovery of a structured signal from independent random measurements. In *Sampling Theory, a Renaissance*. Birkhäuser Verlag, 2014. To appear. Available at `http://arXiv.org/abs/1405.1102`.

[186] J. A. Tropp. Second-order matrix concentration inequalities. Available at `http://www.arXiv.org/abs/1504.05919`, Apr. 2015.

[187] A. M. Tulino and S. Verdú. *Random matrix theory and wireless communications*. Number 1(1) in Foundations and Trends in Communications and Information Theory. Now Publ., 2004.

[188] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[189] J. von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, 53:1021–1099, 1947.

[190] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)*, 62:548–564, 1955.

[191] C. K. I. Williams and M. Seeger. Using the Nyström method to spped up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, Vancouver, 2001.

[192] J. Wishart. The generalised product moment distribution in samples from a multivariate normal population. *Biometrika*, 20A(1–2):32–52, 1928.

[193] D. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, 2014.

[194] A. Zouzias. *Randomized primitives for linear algebra and applications*. PhD thesis, Univ. Toronto, 2013.