Concentration Inequalities of Random Matrices and Solving Ptychography with a Convex Relaxation

Thesis by

Richard Yuhua Chen

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2017
(Defended July 5th, 2016)

To my parents, Bingfa Chen and Lianyan Chen.

Acknowledgments

My time in Caltech has been an invaluable experience, benefiting from the tutorship and helping hands of many great minds. First all, I would like to thank my advisor, Joel Tropp, for introducing me to an exciting field of research, for holding the highest standards, and for supporting me with the resources to become a good scholar. Thanks are also in order for members of the thesis committee, Prof. Thomas Hou, Prof. Houman Owhadi, Prof. Babak Hassibi, and Prof. Changhuei Yang, for their encouragement and evaluation of my research. To the staff of the Applied and Computational Math option at Caltech, Maria Lopez, Sydney Garstang, and Carmen Nemer-Sirois in particular, thank you for taking care of all processes in the department.

In addition, I would like to thank the faculty in the Division of Humanities and Social Sciences, in particular Prof. Jaksa Cvitanic and Dr. George Georgiadis, who encouraged me to explore my interests in social sciences and introduced me to a myriad of interesting topics.

At Caltech, I am lucky to be part of the Caltech Graduate Student Council, through which I have collaborated with wonderful colleagues, such as Toni Lee, Alison Kunz, Elizabeth Jensen, and Artemis Ailianou. Thank you for your passion; I have learned tons from your dedication to making Caltech a good experience for everyone! I am also lucky to work with and learn from many wonderful staff across many organizations in Caltech. I would also like to thank Dr. Felicia Hunt, Dr. Maria Oh, Portia Harris, and Laura Flower Kim for believing in me and supporting me along the way. Their advice and help have assisted me to grow in full dimensions.

I am also indebted to all my friends. I want to thank Matteo Ronchi, Daniel Chao, Khai Chiong, Roarke Horstmeyer, Yong Sheng Soh, Christos Thramboulidis, Ming Fai Wong, Wael Halbawi, Dongyang Kang, Jeongwan Haah, Fernando de Goes, Eldar Akhmetgaliyev, Peyman Tavallali, and Tomasc Tyranowski. Whether it was a movie or hiking experience, I have benefited from our conversation and your wisdom. And Markus Hauschild and Seunghee Ye, among many things, thank you for picking the best restaurants in town for our Sunday dinners! Also, Alberto Stochino, Justin Liu, Andreas Hoenselaar, Nick Hoh, John Yong, and Thomas Ader, thank you for getting me drunk once in a while. And Yunzhang Zhu, Qin Li, Qiaomin Xie, Lele Wang, Alexis Muir Watt, and Danny Goodman, thank you for your encouragement and advice!

Finally and most importantly, I would like to thank my parents, who have made tremendous sacrifice in supporting my growth and education. I want to thank my father, who is always very patient in answering my questions about how everything works when I was little and encourages me to aim for the best. I am grateful to my mother, who perseveres through many challenges that life presents while never fails to be optimistic and smile in the end.

Abstract

Random matrix theory has seen rapid development in recent years. In particular, researchers have developed many non-asymptotic matrix concentration inequalities that parallel powerful scalar concentration inequalities. In this thesis, we focus on three topics: 1) estimating sparse covariance matrix using matrix concentration inequalities, 2) constructing the matrix φ-entropy to derive matrix concentration inequalities, 3) developing scalable algorithms to solve the phase recovery problem of ptychography based on low-rank matrix factorization.

Estimation of covariance matrix is an important subject. In the setting of high dimensional statistics, the number of samples can be small in comparison to the dimension of the problem, thus estimating the complete covariance matrix is unfeasible. By assuming that the covariance matrix satisfies some sparsity assumptions, prior work has proved that it is feasible to estimate the sparse covariance matrix of Gaussian distribution using the masked sample covariance estimator. In this thesis, we use a new approach and apply non-asymptotic matrix concentration inequalities to obtain tight sample bounds for estimating the sparse covariance matrix of subgaussian distributions.

The entropy method is a powerful approach in developing scalar concentration inequalities. The key ingredient is the subadditivity property that scalar entropy function exhibits. In this thesis, we construct a new concept of matrix φ-entropy and prove that matrix φ-entropy also satisfies a subadditivity property similar to the scalar form. We apply this new concept of matrix φ-entropy to derive non-asymptotic matrix concentration inequalities.

Ptychography is a computational imaging technique which transforms low-resolution intensity-only images into a high-resolution complex recovery of the signal. Conventional algorithms are based on alternating projection, which lacks theoretical guarantees for their performance. In this thesis, we construct two new algorithms. The first algorithm relies on a convex formulation of the ptychography problem and on low-rank matrix recovery. This algorithm improves traditional approaches' performance but has high computational cost.

The second algorithm achieves near-linear runtime and memory complexity by factorizing the objective matrix into its low-rank components and approximates the first algorithm's imaging quality.

Published Content and Contributions

R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis via the matrix laplace transform. *CALIFORNIA INST OF TECH PASADENA DEPT OF COMPUTING AND MATHEMATICAL SCIENCES*, 2012.

R. Y. Chen participated in the conception and formulation of the problem, solved the major technical challenge of bounding the matrix variance of a matrix Schur product in the semi-definite order, and participated in the writing of the manuscript.

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. doi: 10.1214/EJP.v19-2964.

R. Y. Chen proposed this project, defined the concept of matrix φ-entropy, and produced an initial argument for establishing the subadditivity property of the matrix φ-entropy, which contained errors and was corrected and improved by J. A. Tropp. R. Y. Chen also participated in the writing of the manuscript.

R. Horstmeyer, R. Y. Chen, X. Ou, B. Ames, J. A. Tropp, and C. Yang. Solving ptychography with a convex relaxation. *New Journal of Physics*, 17(5):053044, 2015. doi: 10.1088/1367-2630/17/5/053044.

R. Y. Chen participated in the design, implementation, and optimization of the algorithms for this project and the writing of the manuscript.

Contents

Acknowledgments **iv**

Abstract **vi**

Published Content and Contributions **viii**

1 Introduction and History **1**

1.1 Random Matrix Theory . . . 2

1.1.1 Asymptotic Approach . . . 2

1.1.1.1 Studying the Limiting Distribution in the Global Regime . . 3

1.1.1.2 Studying the Limiting Distribution in the Local Regime . . . 5

1.1.1.3 Free Probability . . . 7

1.1.2 Non-asymptotic Results . . . 7

1.2 Scalar Concentration Inequalities . . . 10

1.2.1 Scalar Concentration of Measure Phenomenon . . . 11

1.2.1.1 Classical Concentration Inequalities . . . 12

1.2.1.2 Geometric Isoperimetry and Gaussian Concentration Inequalities . . . 13

1.2.1.3 Talagrand's Inequality and Empirical Process . . . 14

1.2.1.4 The Entropy Method . . . 16

1.2.1.5 Transportation Cost Method . . . 18

1.2.1.6 Stein's Method . . . 19

1.2.2 Classical Concentration Inequalities via the Laplace Transform Method 20

1.2.2.1 Scalar Laplace Transform Method . . . 21

1.2.2.2 Linear Decoupling of the Moment Generating Function . . . 21

1.2.3 The Entropy Method . . . 23

1.2.3.1 Scalar Entropy, Deviation Probability Bound, and Herbst's Argument . . . 23

1.2.3.2 Gaussian Concentration From the Logarithmic-Sobolev Inequalities . . . 24

1.2.3.3 Bounded Differences Inequality from Modified Logarithmic-Sobolev Inequalities . . . 25

1.2.3.4 Φ-Entropy and Moment Bounds . . . 30

2 Context and Impact 32

2.1 Matrix Concentration Inequalities . . . 32

2.1.1 History of Matrix Concentration Inequalities . . . 32

2.1.1.1 Classification of Non-Commutative Situations . . . 33

2.1.1.2 Non-Commutative Moment Inequalities . . . 34

2.1.1.3 Non-Commutative Large Deviation Inequalities . . . 37

2.1.1.4 Other Results . . . 40

2.1.1.5 General Notations . . . 41

2.1.2 Matrix Concentration Inequalities via Lieb's Theorem . . . 41

2.1.2.1 Matrix Moment Generating Function . . . 42

2.1.2.2 Matrix Laplace Bound . . . 42

2.1.2.3 Lieb's Theorem and the Subadditivity of Matrix Cumulant Generating Function . . . 43

2.1.2.4 Probability Bounds for Sums of Independent Random Matrices 44

2.1.2.5 Matrix Concentration Inequalities via Lieb's Theorem . . . 45

2.1.2.6 Extension . . . 48

2.1.3 Method of Exchangeable Pairs . . . 50

2.1.3.1 Matrix Stein Pairs . . . 50

2.1.3.2 Constructing Exchangeable Pairs from Sums of Independent Random Matrices . . . 52

2.1.3.3 Control Matrix Moment Generating Function with Method of Exchangeable Pairs . . . 53

2.1.3.4 Probability Bounds for Matrix Stein Pairs . . . 55

2.1.3.5 Matrix Concentration Inequalities via the Method of Exchangeable Pairs 55
2.1.3.6 Extension 57
2.2 Overview of Main Results of the Thesis 58
2.2.1 Analysis of the Masked Sample Covariance Estimator via the Matrix Laplace Transform Method 59
2.2.1.1 Covariance Matrix Estimation 59
2.2.1.2 Masked Sample Covariance Estimator 60
2.2.1.3 Contributions 61
2.2.2 Subadditivity of Matrix φ-Entropy and Matrix Concentration Inequalities 62
2.2.2.1 Matrix Entropy 63
2.2.2.2 Subadditivity Property and Matrix Concentration 65
2.2.2.3 Matrix φ-Entropy 67
2.2.2.4 Concentration Inequalities from the Subadditivity Properties of Matrix Entropy 68
2.2.2.5 Generalized Subadditivity Properties 69
2.2.2.6 Discussions 70
3 The Masked Sample Covariance Estimator: An Analysis via the Matrix Laplace Transform Method 78
3.1 Introduction 78
3.1.1 Classical Covariance Estimation 78
3.1.2 Motivation for Masked Covariance Estimation 80
3.1.3 The Mask Matrix 82
3.1.4 The Masked Sample Covariance Estimator 83
3.1.5 The Complexity of a Mask 84
3.1.6 Masked Covariance Estimation for Gaussian Distributions 85
3.1.6.1 Sample Complexity Bound 86
3.1.6.2 Is This Sample Complexity Bound Optimal? 86
3.1.6.3 Application Example 87
3.1.6.4 Comparison with Bounds of Levina and Vershynin 88

3.1.6.5 Proof Techniques 88
3.1.7 Organization of the paper 89
3.2 Preliminaries 89
3.2.1 Notation and Conventions 90
3.2.2 Facts about the Schur Product 90
3.2.3 Subgaussian Random Variables 91
3.2.4 The Matrix Laplace Transform Method 92
3.3 Masked Covariance Estimation for a Subgaussian Distribution 94
3.3.1 Concentration Parameters 94
3.3.2 Main Result for Masked Covariance Estimation 95
3.3.2.1 Interpretation and Consequences 96
3.3.3 Specialization to Gaussian Distributions 97
3.3.4 Proof of Corollary 3.3.3 from Theorem 3.3.2 99
3.3.5 Proof of Theorem 3.3.2 100
3.3.5.1 The Matrix Laplace Transform Method 100
3.3.5.2 Symmetrizing the Random Sum 101
3.3.5.3 Matrix cgf Bound for the Matrix mgf 101
3.3.5.4 Probability Bound for the Estimation Error 102
3.3.5.5 Bound for the Expected Estimation Error 103
3.4 The Matrix cgf of a Schur Product 104
3.4.1 Computing the Variance 105
3.4.2 Proof of Lemma 3.4.1 107
3.4.2.1 The Truncation Argument 107
3.4.2.2 Bernstein Estimate for the Bounded Part of the Random Matrix 108
3.4.2.3 Controlling the Unbounded Part of the Random Matrix . . . 109
3.4.2.4 Combining the Results 111
4 Subadditivity of Matrix φ-Entropy and Concentration of Random Matrices 112
4.1 Introduction and Related Work 112
4.1.1 Subadditivity and Concentration 112

4.1.2 Subadditivity of Matrix Entropies 114
4.2 Main Results 115
4.2.1 Notation and Background 116
4.2.2 Subadditivity of Matrix Entropies 117
4.2.2.1 The Class of Matrix Entropies 117
4.2.2.2 Matrix φ-Entropy 118
4.2.2.3 Subadditivity of Matrix φ-Entropy 119
4.2.3 Some Matrix Concentration Inequalities 120
4.2.3.1 A Bounded Difference Inequality 121
4.2.3.2 Example: Sample covariance matrices 122
4.2.3.3 Matrix Moment Bounds 123
4.2.4 Generalized Subadditivity of Matrix φ-Entropy 123
4.3 Operators and Functions acting on Matrices 125
4.3.1 Linear Operators on Matrices 125
4.3.2 Monotonicity and Convexity 126
4.3.3 The Derivative of a Vector-Valued Function 127
4.4 Subadditivity of Matrix φ-Entropy 128
4.4.1 Representation of Matrix φ-Entropy as a Supremum 128
4.4.1.1 The Convexity Lemma 129
4.4.1.2 Proof of Lemma 4.4.1 130
4.4.2 A Conditional Jensen Inequality 132
4.4.3 Proof of Theorem 4.2.5 132
4.5 Entropy Bounds via Exchangeability 134
4.5.1 Representation of the Matrix φ-Entropy as an Infimum 134
4.5.2 Proof of Corollary 4.2.6 135
4.6 Members of the Φ_∞ function class 136
4.6.1 Tensor Product Operators 136
4.6.2 The Derivative of a Standard Matrix Function 137
4.6.3 Operator Means 138
4.6.4 Entropy 138
4.6.5 Power Functions 140
4.7 A Bounded Difference Inequality for Random Matrices 143

4.7.1 Preliminaries 143
4.7.2 Proof of Theorem 4.2.8 144
4.7.2.1 The Matrix Laplace Transform Method 144
4.7.2.2 A Differential Inequality for the Trace Mgf 144
4.7.2.3 Solving the Differential Inequality 146
4.7.2.4 The Laplace Transform Argument 147
4.8 Moment Inequalities for Random Matrices with Bounded Differences 147
4.9 Lemma 4.4.1, The General Case 149
4.10 Generalized Subadditivity of Matrix φ-Entropy 152
4.10.1 Representation of the Generalized φ-Entropy as a Supremum 152
4.10.1.1 Generalized Convexity Lemma 153
4.10.1.2 Proof of Lemma 4.10.1 154
4.10.2 Generalized Conditional Jensen Inequality 155
4.10.3 Proof of Theorem 4.2.10 156

5 Solving Ptychography with a Convex Relaxation 158
5.1 Introduction 159
5.2 Fundamentals 163
5.3 Results: Convex Lifted Ptychography (CLP) 165
5.3.1 The CLP solver 165
5.3.2 CLP simulations and noise performance 169
5.4 Results: Factorization for Low-Rank Ptychography (LRP) 171
5.4.1 The LRP solver 171
5.4.2 LRP simulations and noise performance 173
5.5 Results: Experiment 175
5.5.1 Quantitative performance 176
5.5.2 Biological sample reconstruction 178
5.6 Discussion and Conclusion 181

Bibliography 183

Chapter 1

Introduction and History

In this thesis, we study multiples aspects of modern random matrix theory. The matrix Laplace transform method is a recent and versatile approach to develop easy-to-use matrix concentration inequalities. In Chapter 3, we apply this method to analyze the masked sample covariance estimator, which estimates a sparse covariance matrix using a small number of samples. In Chapter 4, we define a concept of matrix entropy for finite-dimensional random matrices. We establish that the matrix entropy also exhibits an appealing subadditivity property and obtain several matrix concentration results. Recent works formulate the problem of phase retrieval as a convex low-rank matrix completion problem. In Chapter 5, we consider the specific phase retrieval problem of ptychography and propose two algorithms. The first algorithm is convex, achieves better signal recovery than existing iterative methods, but scales poorly when the problem size increases. The second algorithm is scalable. It takes a non-convex approach and the performance approximates the recovery quality of the first convex algorithm.

Our work builds upon a rich set of matrix concentration inequalities that are developed in recent decades. And we dedicate Chapter 2 of this thesis to a complete treatment. Matrix concentration inequalities depend upon two related fields, the random matrix theory and the field of developing scalar concentration results. We summarize these two areas in this introductory chapter. In Section 1.1, we cover both the asymptotic and nonasymptotic approaches of studying random matrices. Section 1.2 illustrates various methods of deriving scalar concentration inequalities. In particular, we demonstrate the method of obtaining classical concentration inequalities via the scalar Laplace transform method, which is the inspiration of the matrix Laplace transform method as explained in Chapter 2. In addition, we summarize the powerful scalar entropy method. The scalar entropy method stimulates us to develop a similar entropy method for random matrices in Chapter 4.

1.1 Random Matrix Theory

A random matrix is a random variable taking values in the matrix algebra. Random matrix theory is a diverse field of research that studies random matrices, in particular their spectral properties, under various distributions. The main topics include characterizing random matrices' empirical eigenvalue or singular value distribution, deriving the expected values of the extreme eigenvalues of a Hermitian random matrix, bounding the expected spectral norm of a random matrix, etc.

In this section, we review the two approaches of studying random matrices: the asymptotic approach and the non-asymptotic approach. The asymptotic approach (Section 1.1.1) arose in the 1920s and is also called the classical random matrix theory. The nonasymptotic approach (Section 1.1.2) emerged in recent years as new applications in information science posed a new set of random matrix problems. As a result, the techniques of these two approaches are very different.

1.1.1 Asymptotic Approach

In the asymptotic approach, we consider normalized structured random matrices whose entries follow a certain distribution. As we increase the dimension of the matrix to infinity, the empirical distribution of the eigenvalues of the normalized random matrix often converges to a certain continuous distribution. The goal of the asymptotic approach is to quantify the properties of the limiting spectral distribution in both the global and the local regimes. We delineate the development of this field of research and summarize the main topics and directions of work in both the global regime (Section 1.1.1.1) and the local regime (Section 1.1.1.2). We also mention briefly the framework of free probability (Section 1.1.1.3) to study random matrices. We recommend the books of Anderson et. al. [2] and Tao [217] for a good coverage of this field.

1.1.1.1 Studying the Limiting Distribution in the Global Regime

Random matrices arise in different application scenarios and researchers aggregate them into various matrix ensembles with specific structures and study their limiting spectral properties. In 1928, a random matrix model appeared in Jon Wishart's work [246] where he studied the distribution of the sample covariance matrix for a large multi-variate normal sample. Following Wishart's work, researchers refer to the Gaussian sample covariance matrix as the Wishart matrix ensemble. The Wishart distribution of a $d \times d$ Gaussian sample covariance matrix is characterized by two parameters: a $d \times d$ scale matrix $\boldsymbol{\Sigma}$ that is the true covariance matrix of the samples' Gaussian distribution $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$; the degree of freedom n that corresponds to the number of samples $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \mathbb{R}^{d \times n}$ that generate the sample covariance matrix. Thus, the distribution of the sample covariance matrix

$$\boldsymbol{M}_n = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$$

is denoted as $\boldsymbol{M}_n \sim W_d(\boldsymbol{\Sigma}, n)$. The Wishart matrix generalizes the scalar chi-square distribution to the matrix setting. Wishart matrix plays an important role in studying multivariate statistics. We study the empirical spectral distribution of $\boldsymbol{M}_n$ normalized by the the matrix dimension as follows:

$$\mu_{\frac{1}{d}\boldsymbol{M}_n} := \frac{1}{d}\sum\nolimits_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{M}_n/d)},$$

where δ_y is the Dirac delta function at $y \in R$ and $\{\lambda_i\}$ are the eigenvalues. The Marchenko–Pastur law [142] characterizes the limiting eigenvalue distribution of Gaussian sample covariance matrix as we maintain the ratio of the dimension and sample size $y = d/n \in (0, 1]$ constant and take the value of n to infinity. In the case of $\boldsymbol{\Sigma} = \mathbf{I}$, we have

$$\mu_{\frac{1}{d}\boldsymbol{M}_n} \to \frac{1}{2\pi xy} \cdot \sqrt{(b-x)(x-a)} \cdot 1_{x\in[a,b]} \quad \text{a.s.,} \quad \text{as } y = d/n \text{ and } n \to +\infty,$$

where $a = (1-\sqrt{y})^2$ and $b = (1+\sqrt{y})^2$ mark the boundaries of the limiting distribution. The value of y measures the sampling ratio and the Marchenko–Pastur law tells that when the number of samples is very large, such that y is very close to 0, the limiting distribution is very close to a Dirac delta function at 1. This is consistent with the intuition that when we have sufficient samples, the sample covariance matrix is close to the real covariance matrix $\mathbf{I}$ in this case.

Then, in 1955, Eugene Wigner [245] constructed large symmetric random matrices to model the behavior of atomic nuclei. Wigner's work motivated a line of research regarding the asymptotic behaviors of large Hermitian matrices with i.i.d. entries, or as we call them today, the Wigner matrices. A $d \times d$ Wigner matrix can be represented as $\boldsymbol{W}_d = (\xi_{ij})$ where the entries $\{\xi_{ij}\}$ are complex and satisfy the Hermitian condition $\xi_{ij} = \xi_{ji}^*$, and the upper-triangle entries $\{\xi_{ij}\}_{i>j}$ are i.i.d. complex random variables. The corresponding normalized empirical spectral distribution is

$$\mu_{\frac{1}{\sqrt{d}}\boldsymbol{W}_d} := \frac{1}{n} \cdot \sum\nolimits_{i=1}^{d} \delta_{\lambda_i(\boldsymbol{W}_d/\sqrt{d})}.$$

Special cases of the Wigner matrix include the Gaussian matrix ensembles, which include the Gaussian Unitary Ensemble (GUE), the Gaussian Orthogonal Ensemble (GOE), and the Gaussian Symplectic Ensemble (GSE), whose distributions are invariant under unitary, orthogonal, and symplectic conjugation respectively. In Wigner's work, he considered the special case when the the upper-triangle entries are independent and follow the standard Gaussian distribution. He proved that the limiting spectral distribution of $\frac{1}{\sqrt{d}}\boldsymbol{W}_d$ is the semicircle law. The semicircle law [245] states that as we increase the dimension to infinity, the empirical spectral distribution of $\frac{1}{\sqrt{d}}\boldsymbol{W}_d$ converges in probability to the semicircle distribution:

$$\mu_{\frac{1}{\sqrt{d}}\boldsymbol{W}_d} \to \frac{1}{2\pi}\sqrt{4 - |x|^2} \cdot 1_{x \in [-2,2]} \,\mathrm{d}x \quad \text{in probability.,} \quad \text{as } d \to +\infty. \tag{1.1.1}$$

After Wigner's work, various authors consider more general cases of Hermitian matrices and attempt to establish conditions for stronger convergence. The intuition is to go beyond specific structure assumptions associated with each random matrix ensemble and establish the limiting distribution of a large group of matrices sharing the same symmetry structure. This intuition is called the universality principle [62]. In [4], Arnold established that when the upper-triangular entries are independent, mean-zero, and have unit variance, a sufficient condition for the limiting spectral distribution to converge almost surely to the semicircle law (1.1.1) is that the entries have finite fourth moment. It turns out that the almost-sure convergence still holds without the finite fourth moment assumption. See the details in Bai and Silverstein's monograph [6, Chapter 2].

Bai and Yin [7] also made a connection between the sample covariance matrix and the semicircle law. They assume that the matrix $\boldsymbol{X} \in \mathbb{C}^{d\times n}$ have i.i.d. entries with mean zero, unit variance, and finite fourth-moment. Take the limit of $n \to \infty$ and $d/n \to 0$. Then the empirical spectral distribution of the following centered sample covariance matrix

$$\boldsymbol{B}_d = \frac{1}{\sqrt{nd}}(\boldsymbol{X}\boldsymbol{X}' - \mathbf{I})$$

converges to the semicircle law (1.1.1) almost surely.

Going beyong Hermitian random matrices, we have an important conjecture, the circular law. This conjecture posits that when the entries of an $n \times n$ random matrix $\boldsymbol{A}_n$ are i.i.d., centered, and have unit variance, the empirical spectral distribution of $\frac{1}{\sqrt{n}}\boldsymbol{A}_n$ converges to the uniform distribution on the complex unit disk both in probability and almost surely. Many authors contribute to proving this conjecture, including Zhou & Pan [171] and Götze & Tikhomirov [83]. Tao and Vu fully establish the circular law in their work [221].

1.1.1.2 Studying the Limiting Distribution in the Local Regime

In the asymptotic approach, studying the spectral properties in the local regime is also very important and bears many fruitful results. Research in the local regime is divided into two categories. The first category looks at the bulk statistics and studies the statistical properties such as the joint distributions of the eigenvalues that lie between the extreme eigenvalues. The focus is the k-point correlation function

$$\rho_k^{(n)}(x_1, \ldots, x_k) \quad \text{where } k \leq n \text{ and } x_1 < x_2 < \cdots x_k$$

for an $n \times n$ matrix. The k-point correlation function is the limiting probability density for the event that there is an eigenvalue in each of the disjoint intervals $[x_1, x_1 + \varepsilon], \ldots, [x_k, x_k + \varepsilon]$ as $\varepsilon \to 0$. Integrating various functions against the k-

point correlation function leads to quantities such as the eigenvalue gap, etc.

One central theme in this line of work is to establish the Wigner–Dyson–Mehta universality conjecture [67], which posits that when appropriated normalized, the k-point correlation functions of the eigenvalues of an $n \times n$ Wigner matrix in the bulk converges to the k-point correlation function of the Dyson sine process when we take the limit $n \to +\infty$. An early result is given by Ginibre [79] for the Gaussian Unitary Ensemble (GUE), in which the matrix entries are independent, standard normal variables. The distribution of a GUE matrix is invariant under conjugation of unitary matrices. Ginibre obtained explicit expressions for the joint distributions and the spacing between the eigenvalues for GUE matrices. Then in 2001, Johansson [107] proved the conjecture for a more general glass of Wigner matrix, called the Johansson ensemble, which is a linear interpolation between an arbitrary Wigner matrix and a GUE matrix. In recent years, we see important breakthroughs which establish the conjecture for a larger class of matrix ensemble. Erdos et.al. [68] prove the universality conjecture for all Wigner matrices whose entries distribute according to certain smoothness and decay conditions. In the work of Tao & Vu [218], the authors establish the conjecture based on certain moments assumption and conditions on the support of the distribution of the matrix entries. Combining their methods together, the authors [69] jointly validate the conjecture for all Wigner matrices whose entries exhibit sub-exponential distributions. Tao and Vu [219] further establish the conjecture under a stronger sense of convergence. Going beyond Hermitian matrices, Tao and Vu [220] prove that the bulk statistics of non-Hermitian matrices with independent entries whose distribution satisfies certain conditions also exhibit the asymptotic behavior of the conjecture.

The second category studies the edge, or the extreme eigenvalues of the limiting spectral distribution, which has a different asymptotic behavior compared with the bulk. The Bai–Yin law [5] describes the following limiting behavior of the extreme singular values for an $N \times n$ matrix whose entries are i.i.d. with zero mean, unit variance, and finite fourth moment:

$$s_{\min}(\boldsymbol{A}) \sim \sqrt{N} - \sqrt{n} \quad \text{and} \quad s_{\max}(\boldsymbol{A}) \sim \sqrt{N} + \sqrt{n} \quad \text{almost surely,} \tag{1.1.2}$$

as $N \to \infty$ and n/N converges to a constant. Another important result is the Tracy–Widom distributions [228, 229] which characterize the distribution of the normalized largest eigenvalue of the Wigner matrices. Tracy–Widom distributions consist of three types, which correspond to the Gaussian Orthogonal, Unitary, and Symplectic Ensembles respectively. The universality principle also exhibits in the random matrices' edge behavior. For example, El Karoui [116] extended this result to the Wishart distribution and provided the sufficient conditions for the largest eigenvalue of a nonsingular Gaussian sample covariance matrix to converge to the Tracy–Widom distribution. Onatski [165] extends El Karoui's result to the singular Wishart distribution. We also mention a recent work by Bao et. al. [9] along this direction.

1.1.1.3 Free Probability

A new theoretical framework to study random matrices asymptotically is the free probability developed by Volculescu. Instead of the traditional probability space, free probability is based upon a possibly noncommutative algebra endowed with a linear expectation functional. Voiculescu [241] showed that when the dimension goes to infinity, the behaviors of random matrices converge to those of free random variables. In addition, Voiculescu established that free random variables also have a central limit theorem where the limiting distribution is Wigner's semicircle law (1.1.1). Voiculescu defined the concept of free entropy [243] in free probability and demonstrated the connections between the free entropy and the asymptotic behavior of large random matrices. He also explained the applications of free entropy in solving problems related to von Neumann algebras. Speicher's book chapter [203], Hiai and Petz's book [95], and Voiculescu's book [242] provide a good reference for this field.

1.1.2 Non-asymptotic Results

This thesis focuses on the concentration inequalities of finite dimensional structured random matrices and this field belongs to the second nonasymptotic approach of studying random matrices. Unlike in the asymptotic approach, where the normalized empirical spectral distribution converges when the matrix's dimension is taken to infinity, the spectral distribution of finite size random matrices is far from convergent and usually depends on the specific structures of the random matrices. Thus, the goal of the nonasymptotic approach is to derive approximate characterizations of the spectral distribution of finite-size random matrices.

The first set of problems is to derive nonasymptotic versions of the asymptotic results for random matrix ensembles. Due to the accommodating properties of the Gaussian distribution, such results appeared first for random matrices with Gaussian entries. For example, Gordon's theorem [59] is the following nonasymptotic version of the Bai-Yin law (1.1.2) that bounds the extreme singular values of an $N \times n$ $(N > n)$ matrix $\boldsymbol{A}$ with independent entries that follow the standard normal distribution:

$$\sqrt{N} - \sqrt{n} \leq \mathbb{E}\, s_{\min}(\boldsymbol{A}) \leq \mathbb{E}\, s_{\max}(\boldsymbol{A}) \leq \sqrt{N} + \sqrt{n}. \tag{1.1.3}$$

Szarek [208] characterized the deviation of the interior singular values from their 'typical locations' for square matrices with independent Gaussian entries, which can be considered as a nonasymptotic version of the semicircle law. Going beyond the Gaussian distribution, an important result due to Latala [121] extends the upper bound in (1.1.3) and controls the expected maximum singular value of a random matrix $\boldsymbol{A}$ whose entries a_{ij} are independent and have zero mean:

$$\mathbb{E}\, s_{\max}(\boldsymbol{A}) \leq C \left[\max_i \left(\sum\nolimits_j \mathbb{E}\, a_{ij}^2 \right)^{1/2} + \max_j \left(\sum\nolimits_i \mathbb{E}\, a_{ij}^2 \right)^{1/2} + \left(\sum\nolimits_{i,j} \mathbb{E}\, a_{ij}^4 \right)^{1/4} \right], \tag{1.1.4}$$

where C is a universal constant. Seginer derived similar results in his work [200]. Other authors including Szarek [209], Litval et. al. [135], and Rudelson & Vershynin [194, 195] characterize the least singular values of random matrices. We point out that Vershynin's chapter [239] is a good source for reference.

Another set of problems in the nonasymptotic approach is to derive matrix concentration-of-measure results, which include the large deviation probability bounds and matrix moment bounds, for random matrix functions. The former provides upper bounds on the probability that a random matrix $\boldsymbol{X}$ of finite dimension deviates from its mean $\mathbb{E}\,\boldsymbol{X}$ by a certain threshold t:

$$\mathbb{P}\left\{ \|\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}\| \geq t \right\}.$$

These probabilistic upper bounds usually depend on the ambient dimension of the random matrix and other parameters related to the distribution. The latter provides upper bounds on the matrix moments as measured by the Shatten p-norm:

$$\mathbb{E} \left\| \boldsymbol{X} \right\|_p^p \quad \text{for all } p \in \mathbb{Z}_+,$$

where $\left\| \boldsymbol{X} \right\|_p = \left(\sum_i s_i^p(\boldsymbol{X}) \right)^{1/p}$ and $\{s_i(\boldsymbol{X})\}$ are the singular values of $\boldsymbol{X}$. These two problems are very related and usually an improved result in one implies potential refinement in the other. This set of problems parallels the scalar concentration results that we introduce in Section 1.2.1. In this section, we summarize related applications. In Chapter 2, we provide an in-depth discussion on the history and main results of matrix concentration inequalities.

The interest in deriving nonasymptotic concentration inequalities for random matrix functions stems from many modern applications that require characterizations of the spectrum of finite-size random matrices. One related field of application is compressed sensing [65], which leverages the sparsity heuristic to recover high dimensional signals that can be sparsely represented in a certain basis from a set of under-determined linear measurements. A sufficient condition for successful recovery of the sparse signal is that the sampling matrix satisfies the restricted isometry property (RIP) [43]. RIP requires that any submatrix of the sampling matrix roughly preserves the spectral norm when the submatrix multiplies with an arbitrary vector. Constructions of deterministic sampling matrices that satisfy RIP are difficult while Candes et al [43] show that certain classes of random matrices exhibit RIP with very high probability. Rauhut's chapter [185] shows that matrix concentration inequalities such as the non-commutative Khintchine's inequality play an important role in establishing the RIP properties of random matrices. The non-commutative Khintchine's inequality extends from classical scalar Khintchine's inequality and controls high-order matrix moments for a sum of deterministic matrices, instead of scalars, modulated by independent Rademacher or Gaussian random variables with second-order matrix variance. Finding the best streamlined derivation and obtaining the tightest upper bound is an active research topic and we review its development in Chapter 2. Another related field of application is high dimensional data analytics. As the size of data matrices increases, operations such as a full singular value decomposition become exorbitantly expensive. Based on the spectral properties of random matrices, various authors including Halko et al. [91] have developed approximate randomized algorithms that reduces computational complexity by dimensionality reduction.

Scalar concentration inequalities is a well-developed field. Although matrices are non-commutative, researchers have drawn inspirations from the methodologies of developing scalar concentration inequalities and have successfully developed some matrix counterparts. In the next section, we illustrate concepts and methods from scalar concentration inequalities that are relevant in the scope of the thesis.

1.2 Scalar Concentration Inequalities

In probability theory, the law of large numbers dictates that the average of n independent and identically distributed random variables converges asympotically to the mean, in probability (the weak law) or almost surely (the strong law), as the number of summands n goes to infinity. In applications, we often encounter problems that require non-asympototic characterizations on the deviation of the average of a finite sum from its mean. Concentration inequalities provide such answers. Classical large deviation inequalities, such as Hoeffding's inequality, estimate the following deviation probability for a sum of i.i.d. random variables $\{X_i\}$

$$\mathbb{P}\left\{\left|\frac{1}{n}\cdot\sum\nolimits_{i=1}^{n} X_i - \mathbb{E}\, X_i\right| \geq t\right\}, \quad t \geq 0. \tag{1.2.1}$$

The deviation probabilities decay rapidly as the value of t increases and exhibit a high concentration of measure around the expectation. Besides sums of independent random variables, the concentration-of-measure phenomenon also extends to more general scalar functions of independent random variables and developing new tools for deriving new concentration results is an active field of research.

Another set of concentration inequalities are moment bounds of random functions. The goal is to bound the following L_p norm of a random function. For example, for a sum of i.i.d. random variables in this case we want to control

$$\left\|\sum\nolimits_{i=1}^{n} X_i\right\| = \left(\mathbb{E}\left|\sum\nolimits_{i=1}^{n} X_i\right|^{p}\right)^{1/p} \tag{1.2.2}$$

from above and below. The methods of characterizing scalar large deviation probabilities and moment bounds are often the foundations based on which researchers develop matrix concentration results.

In this section, we review the scalar concentration inequalities and the relevant methods. First, we summarize the development of various approaches to study the scalar concentration of measure phenomenon in Section 1.2.1. Next, we illustrate the main ideas behind the derivation of classical concentration inequalities via the Laplace transform method in Section 1.2.2 and the entropy method in Section 1.2.3. The matrix Laplace transform method is the matrix equivalent of the scalar Laplace transform method, which is the key to multiple matrix concentration inequalities. In particular, coupled with a deep concavity result called Lieb's theorem, the matrix Laplace transform method leads to multiple matrix versions of classical concentration inequalities, which we exhibit in Chapter 2. The scalar entropy method is the inspiration for our work in constructing the matrix φ-entropy in Chapter 4.

1.2.1 Scalar Concentration of Measure Phenomenon

As Talagrand described in his paper [214], the concentration of measure is the phenomenon that with high probability, regular functions of multiple independent variables are very close to the mean. As described in [125], concentration of measure phenomenon exists in classical probability, geometric analysis, and functional analysis. The book [30] provides a comprehensive coverage for this field of research. We illustrate the historical development and the main methods of this field, many of which serve as foundations for developing matrix concentration inequalities. We start with the classical concentration inequalities for sums of independent random variables in Section 1.2.1.1. Next, we review modern approaches of deriving concentration inequalities. These approaches start with the geometric isoperimetric property, which leads to concentration inequalities for the Gaussian distribution in Section 1.2.1.2. Then in Section 1.2.1.3, we exhibit Talagrand's inequality, which is an important development to generalize Gaussian concentration results to other product distribution. Information inequalities are useful tools to derive concentration inequalities. The entropy method (Section 1.2.1.4) and the transportation cost method (Section 1.2.1.5) are two major approaches based on information inequalities. Finally, we review a recent approach based on Stein's method in Section 1.2.1.6.

1.2.1.1 Classical Concentration Inequalities

In classical probability we have multiple concentration inequalities for a sum of independent random variables. Classical moment bounds include the Khintchine inequality [118, 134, 170], which controls the L_p norm of a Rademacher sum in both directions:

$$A_p \left(\sum_{i=1}^n a_i^2\right)^{1/2} \leq \left(\mathbb{E}\left|\sum_{i=1}^n a_i \varepsilon_i\right|^p\right)^{1/p} \leq B_p \left(\sum_{i=1}^n a_i^2\right)^{1/2}, \tag{1.2.3}$$

where $a_1, \ldots, a_n \in R$ are deterministic and $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables. The optimal constants A_p, B_p depend on the value of p. When $0 < p \leq 2$, $B_p = 1$ and when $2 \leq p < \infty$, $A_p = 1$. Szarek [207] established that $A_1 = 1/\sqrt{2}$. Young [247] obtained the optimal B_p for $p \geq 3$. Haagerup [88] established the values of A_p, B_p for the remaining cases.

Another moment bound is the Rosenthal's inequality [191], which controls the L_p norm of a sum independent mean-zero random variables. Suppose $\{X_i\}$ are independent and mean-zero, then the Rosenthal inequality says

$$\mathbb{E}\left|\sum_{i=1}^n X_i\right|^p \leq C_p \cdot \max\left(\sum_{i=1}^n \mathbb{E}\left|X_i\right|^p, \left(\sum_{i=1}^n \mathbb{E}\, X_i^2\right)^{p/2}\right). \tag{1.2.4}$$

Many researchers [175, 75, 160, 105, 106] work to obtain the optimal constant C_p. Later, Burkholder et. al. [38, 39] generalized the Rosenthal's inequality to the martingales and the result is called the Burkholder–Davis–Gundy inequality.

Classical large deviation inequalities control the large derivation probability of the sum from the mean value and they are developed by various authors including Bennet [15], Bernstein [16], Hoeffding [96] etc. One example is the Hoeffding's inequality, which assumes that $X_1, \ldots, X_n \in \mathbb{R}$ are independent random variables and each X_i is bounded by the interval $[a_i, b_i]$ almost surely. The Hoeffding's inequality describes the probability of the deviation of the sum $Z = X_1 + \cdots + X_n$ from the expectation as

$$\mathbb{P}\left\{|Z - \mathbb{E}\, Z| \geq t\right\} \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

This group of concentration results rely on a common procedure. The first step is the

Laplace transform method, which controls the deviation probability with the moment generating function of the sum. The second step is to decouple the random variables from the moment generating function of the sum using their independence. Convenient as it is, the decoupling step also prevents the Laplace transform method to derive concentration inequalities for more general functions of random variables because a linear decoupling does not always exist. However, the following line of research that started from the geometric isoperimetric properties provides an alternative route and produces an abundant set of scalar concentration inequalities.

1.2.1.2 Geometric Isoperimetry and Gaussian Concentration Inequalities

Intuitively, the isoperimetric property says that in a high-dimension real space, the Euclidean ball has the smallest surface among all compact sets with the same volume. This idea, attributed to Levy [129] and Schmidt [198], leads to an important concentration of measure result on high-dimensional real unit spheres. Assume that the n-dimensional unit ball S^n is endowed with a uniform measure μ with total measure $\mu(S^n) = 1$. Suppose $D \in S^n$ covers more than half of the surface of S^n, that is, $\mu(D) \geq 1/2$. Expand D outwards from its boundary by a distance of t to arrive at the fattened set

$$D_t = \{\boldsymbol{y} \in S^n : \exists \boldsymbol{x} \in D \quad s.t. \, \|\boldsymbol{x} - \boldsymbol{y}\| \leq t\}.$$

Then Levy and Schmidt show that the measure of the complement of D_t decays very rapidly as we increase t:

$$1 - \mu(D_t) \leq \mathrm{e}^{-nt^2/2}. \tag{1.2.5}$$

Intuitively this result says that fattening D by a distance t quickly absorbs the remaining measure on the unit sphere. It also implies that the majority of measure concentrates around the boundary of any half sphere. Milman applied this result extensively in his proof of the Dvoretsky theorem [157]. Later, Gromov [85] defined observable diameter, which provides a visual description of the concentration of measure phenomenon. Levy [130] considered continuous functions on high dimensional spheres and proved that the functions concentrate around the median value of the function on the sphere.

A key step of generalizing the geometric approach is to consider Lipschitz functions taking variables from abstract metric spaces. One of the early isoperimetric results on the real space is established for the Gaussian distribution by Borell [25] and Tsirelson et. al. [237]. Assume that $\boldsymbol{x} := (X_1, \ldots, X_n)$ is a vector of independent standard normal random variables and the function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L. Then Borell and Tsirelson establish that the value of $f(\boldsymbol{x})$ concentrates around the mean and the Lipschitz constant controls the Gaussian-type deviation probability:

$$\mathbb{P}\{|f(\boldsymbol{x}) - \mathbb{E}\, f(\boldsymbol{x})| > t\} \leq 2\exp\left(-\frac{t^2}{2L^2}\right), \quad \text{for all } t > 0.$$

1.2.1.3 Talagrand's Inequality and Empirical Process

The next development occurred to generalize the concentration results for Gaussian distributions to other product distributions. Talagrand made important contributions in this direction and the paper [214] summarizes the new methods that Talagrand developed to study concentration of measure phenomenon on product spaces. One important result, called Talagrand's inequality, appeared in [210] and was later refined and extended in [212]. It considers a product space $\Omega = \Omega_1 \times \cdots \times \Omega_n$ associated with a product probability $\mathbb{P} = \mu_1 \otimes \cdots \otimes \mu_n$ and says that for any nonempty measurable subset $A \subset \Omega$, the convex distance $d_c(x, A)$ between any point $x \in \Omega$ and the set A is related by the reciprocal of the probability of the set A:

$$\int_\Omega \mathrm{e}^{d_c(x,A)^2/4}\, \mathrm{d}\mathbb{P}(x) \leq \frac{1}{\mathbb{P}\{A\}}. \tag{1.2.6}$$

Talagrand's inequality essentially extends the isoperimetric intuition of (1.2.5) to the product measure $\mathbb{P}$ because it implies a similar rapid probability decay of the complement of the fattened set:

$$1 - \mathbb{P}\{A_t\} \leq \frac{1}{\mathbb{P}\{A\}}\mathrm{e}^{-t^2/4},$$

where A_t is the fattened set defined as $A_t = \{x \in \Omega : d_c(x, A) \leq t\}$. Talagrand's inequality also leads to the following concentration result for convex Lipschitz functions on product real space. Suppose f is a convex Lipschitz function with Lipschitz constant equal to 1 and $\mathbb{P} = \mu_1 \otimes \cdots \otimes \mu_n$ is a product of probability measures on $[0, 1]$. The median of f in the measure $\mathbb{P}$ is M. Then

$$\mathbb{P}\{f \geq M + t\} \leq 2\mathrm{e}^{-t^2/4}. \tag{1.2.7}$$

As a special case, based on (1.2.7) Talagrand proved the scalar Khintchine's inequality in [210]. We mentioned that Maurey [156] provided an alternative proof of Talagrand's inequality (1.2.6).

Talagrand also developed a method to control the tail probability of a supremum of Gaussian processes in [211]. He isolated the main contribution of the tail probability to a finite number of points and controlled the deviations of these points using Gaussian concentration results. An important application of this method relates to a problem in the empirical processes which was previously studied by Keifer [119], Massart [151], etc. Suppose $X_1, \dots, X_n$ are independently samples from the same probability distribution on $\mathbb{R}$ and $\mathcal{F}$ is a countable set of real functions. The goal of studying the suprema of the empirical processes is to quantify the large deviation probability of the following quantity

$$Z = \sup_{f \in \mathcal{F}} \left| \sum\nolimits_{i=1}^{n} f(X_i) - n\,\mathbb{E}\, f(X_1) \right|. \tag{1.2.8}$$

This quantity characterizes the worst speed of convergence to the expectation $\mathbb{E}\, f(X_1)$ among all the functions in the set $\mathcal{F}$. Bounding the large deviation probability of (1.2.8) describes the nonasymptotic behavior of (1.2.8) that evolves with the sample size n, which is an important question in the nonasymptotic theory of model selection [154] in statistics. Talagrand's approach [213] provided an improved upper bound for the deviation probability when the function set $\mathcal{F}$ satisfies certain conditions and motivated researchers to use concentration of measure inequalities to study model selection nonasymptotically. As pointed out by Talagrand [213], these conditions are unwieldy and that the ideal goal is to derive results based on easier conditions such as an upper bound on the supremum of the variance in the function set $\mathcal{F}$:

$$\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}}(\mathbb{E}(f - \mathbb{E}\, f))^2. \tag{1.2.9}$$

Talagrand later provided a result in his work [214] and provided a Bennett-type bound for the deviation inequality of Z based on the variance supremum $\sigma^2(\mathcal{F})$. Noticing the power of concentration inequalities to study many application problems, many researchers develop other methods for developing general-purpose and easy-to-use concentration inequalities. A particularly fruitful approach is to take advantage of information inequalities which we illustrate in the next two sections.

1.2.1.4 The Entropy Method

The connection between concentration results with information inequalities started to appear in the works of Marton [150], Dembo [63], Ledoux [123], and Bobkov & Ledoux [24]. Generally speaking, we can classify their results into two classes. The first class such as the work of Ledoux and Bobkov & Ledoux is the entropy method and focuses on deriving concentration results with functional inequalities such as the logarithmic-Sobolev inequality. These inequalities provide tools and methods to overcome the limit of the Laplace transform method when the function does not admit linear decoupling. Ledoux's book [125] is a good reference for this method and in Section 1.2.3, we review the main arguments of the entropy method. The second class such as Marton's work depends on another type of functional inequality, the transportation cost inequality, which is the subject of the next section.

L. Gross [86] introduced the logarithmic-Sobolev inequality to study Markov semigroups and he established that a Markov semigroup is hypercontractive if and only if the corresponding invariant measure of the semigroup satisfies a logarithmic-Sobolev inequality. The logarithmic-Sobolev inequality controls the entropy of a function under a probabilistic distribution with a Dirichelet form, which relates to the derivative of the function. A related inequality is the Poincaré inequality, which controls the variance of the function by its derivative. In [123], Ledoux developed an inductive method to tensorize one-dimensional logarithmic-Sobolev inequality to a product measure. The tensorization property is essentially the subadditivity property of the entropy functional. Ledoux obtained a slightly different version of Talagrand's Gaussian-type concentration result (1.2.7) for convex Lipschitz functions, where the median is replaced by the function's mean value. In the same paper, Ledoux also provided a simplified proof to for the deviation probability of the maxima of empirical processes established by Talagrand [214].

After Ledoux's work, Massart noticed that the constants in Talagrand's and Ledoux's deviation probabilities for the maixima of empirical processes were not clearly derived. Specifically, when reduced to the one-dimensional case, their deviation probabilities do not recover the best constants. With this discovery in mind, Massart worked to obtain the missing constants in his work [152], although still not optimal. Massart also adopted the argument using the logarithmic-Sobolev inequality. Massart's approach reframed the proof of the tensorization property of the logarithmic-Sobolev inequality that exists in Ledoux's work [123] using a probabilistic approach based on the variational characterizations of the entropy functional. This streamlined presentation is easier to adapt for different applications and Massart is credited with refining the entropy method such that it becomes widely adopted. Based on Massart's version of the entropy method., Boucheron et. al. [28] derived multiple concentration inequalities considered as exponential versions of the Efron–Stein inequality, which controls the variance of general functions. In their work, the authors also demonstrate that these results apply to graph theory and other statistical estimation problems.

Another important implication of the entropy method relates to the previously mentioned subject of empirical processes. The optimal deviation probability for the suprema of empirical processes is provided by Bousquet in [31] based on Massart's refinement of the entropy method. We exhibit Bousquet's deviation probability [31, Theorem 2.3] for the suprema (1.2.8) with n i.i.d. samples:

$$\mathbb{P}\left\{Z - \mathbb{E}\, Z \geq t\right\} \leq \exp\left(-vh(t/v)\right),$$

where $v = n \cdot \sigma^2(\mathcal{F}) + 2\,\mathbb{E}\, Z$ is a function of the variance suprema (1.2.9) and $h(x) = (1+x)\log(1+x) - x$.

An significant extension of the entropy method occurs in the work of Latała & Oleszkiewicz [122]. The authors discovered that the subadditivity property is not unique to the Shannon entropy generated from the logarithmic function. They extended the concept of entropy to a large class of φ-entropy functional and constructed the correspondingly φ-Sobolev inequality. Specifically, a subclass of power functions also generates a set of φ-Sobolev inequalities that are tensorizable in a product space. The authors interpreted these φ-Sobolev inequalities as interpolations between the logarithmic-Sobolev and the Poincaré inequalities. In addition, the authors also established the corresponding deviation probabilities for probability measures that satisfy the φ-Sobolev inequalities.

Based on the tensorization property of the generalized φ-entropy, Boucheron et. al [26] derived multiple moment bounds for functions of independent random variables. In particularly, their method recovers classical Rosenthal and Khintchine-type moment bounds for sums of independent random variables. The authors also established the moment bounds of the suprema of empirical processes that complement the results of Bousquet.

Finally, we mentioned Maurer's work [155] which provides a slightly different formulation of the entropy method based on thermodynamics. Among the results, Maurer derived a tighter version of the bounded difference inequality and presented several directions of extending his method.

1.2.1.5 Transportation Cost Method

A different approach is the transportation cost method, for which Talagrand [216], Marton [146, 150, 147], Dembo [63], Bobkov & Götze [23] are major contributors. This method accommodates nonproduct measures especially for Markov chains. We briefly described the development of the transportation cost method in this section.

The starting point of the transportation cost method is the Pinsker's inequality [176], which controls the total variation distance between two measures with their relative entropy. Based on the Pinsker's inequality, Marton [146] developed an elegant argument to produce isoperimetric results similar to (1.2.5). Expanding this technique to study contracting Markov chains in [150] and [147], Marton produced a refined argument for the transportation cost method, which relies on bounding two types of measure distances, the L_1 Wasserstein and L_2 Wasserstein distances, by the relative entropy. The L_2 Wasserstein distance has the convenient property that it is dimensional free and tensorizes to Euclidean product spaces, which suits the requirement for developing concentration results for functions of multiple random variables. Later, Dembo [63] applied Marton's argument and reproduced Talagrand's concentration results in [214]. Other recent works include [148], where Marton extended this method to study strong mixing Markov chains. In [149] Marton studied concentration inequalities with the transportation cost method when a logarithmic-Sobolev inequality cannot be easily proved.

We mention that Gozlan & Léonard's survey [84] is a good source for reference. Ledoux's lecture notes [126] provides an in-depth discussion about the connection between the transport cost method and the logarithmic-Sobolev inequality. Finally, Villani's book [240] also contains insightful studies on the transportation cost method.

1.2.1.6 Stein's Method

Another recent approach of deriving concentration inequalities for scalar random variables is Chatterjee's method of exchangeable pairs [51]. The idea has roots in the classical Stein's method [205, 10] which was developed by Charles Stein and his student Louis Chen and is originally conceived to measure the difference between two probability distributions and prove the convergence to standard distributions such as Gaussian and Poisson. Chatterjee [51] constructed the method of exchangeable pairs based on the Poisson approximation framework of the Stein's method.

The key of Chatterjee's exchangeable pairs method is to construct an efficient exchangeable pair that contains two random variables that are very 'close' to each other. A good exchangeable pair admits a locally randomized characterization of the variance and relates the moment generating function or moment bounds with the variance characterization. This method leads to many fruitful concentration results. Chatterjee demonstrated the exchangeable pair constructed from a sum of independent random variables and established variants of multiple classical concentration results such as the Hoeffding's inequality [51, Theorem 3.3] and the Bernstein's inequality [51, Theorem 3.13]. He also derived a exchangeable pair version of Burkholder–Davis–Gundy moment bounds [51, Theorem 3.14]. The exchangeable pairs method also applied to multiple application settings such as the Curie–Weiss model [51, Section 3.3] of ferromagnetic interactions, Spearman's footrule [51, Section 3.7], and the Sherrington–

Kirkpatrick model [51, Section 3.10] of spin glasses.

Chatterjee [51, Chapter 4] also designed a general approach to construct exchangeable pairs based on the Poisson equation and a Markov chain coupling method. This construction leads to Gaussian-type concentration inequalities [51, Theorem 4.3] for functions of weakly dependent random variables. The dependence relationship is characterized by the Dobrushin's independence matrix. This concentration result improves earlier results such as that of Stroock & Zegarlinksi [206] who used the logarithmic-Sobolev inequalities to produce concentration inequalities from the Dobrushin mixing condition but their results did not come with explicit constants. Chatterjee also exemplify this general approach with applications in graph theory [51, Section 4.4], study concentrations on the Haar measure [51, Section 4.5], and make connections with free probability [51, Section 4.6].

The exchangeable pairs method shares many technical details with the entropy method and the transportation cost method. For example, some of the variance characterizations in the exchangeable pair method share common intuitions with the variance quantities defined in [27, 26]. Exploring this connection to gain a better understanding of the exchangeable pair method is an ongoing area of research. For example, in [127] Ledoux et. al. explore the connections between Stein's method, the logarithmic-Sobolev inequality, and the transportation cost inequalities.

Inspired by Chatterjee's work, the authors of [138] and [173] extended the method of exchangeable pairs to random matrices. In Chapter 2, we will discuss the technical ingredients and intuition of the matrix version of the exchangeable pairs method, which are very similar to the original argument of Chatterjee's for scalar random variables.

The matrix concentration results developed in this thesis are strongly tied to the approach of developing classical concentration inequalities via the scalar Laplace transform method and the scalar entropy method. So in the following two sections, we focus our attention and summarize the major technical ingredients of these two methods.

1.2.2 Classical Concentration Inequalities via the Laplace Transform Method

The derivation of classical concentration inequalities, such as the Hoeffding's inequality, contains two major steps. The first step (Section 1.2.2.1) is the Laplace transform method, which controls the deviation probability with the moment generating function of the independent sum. The second step (Section 1.2.2.2) decouples the moment generating function of the sum and bounds the individual moment generating function. In the following, we instantiate this method by proving the scalar Hoeffding's inequality.

1.2.2.1 Scalar Laplace Transform Method

The Laplace transform method, also called the Cramer–Chernoff method [154], converts the problem of bounding the large deviation probability of a sum of independent random variables into a problem of controlling the moment generating function of each random variable. Assume that $\theta > 0$ and a scalar random variable Z, then Markov's inequality bounds the upper deviation probability of Z with the moment generating function

$$\mathbb{P}\left\{Z - \mathbb{E}\, Z \geq t\right\} = \mathbb{P}\left\{\mathrm{e}^{\theta(Z-\mathbb{E}\, Z)} \geq \mathrm{e}^{\theta t}\right\} \leq \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{e}^{\theta(Z-\mathbb{E}\, Z)}.$$

The inequality does not depend on the specific value of θ, which behaves as a tuning parameter for the probability bound. We can obtain the best upper bound by taking the infimum over all positive θ:

$$\mathbb{P}\left\{Z - \mathbb{E}\, Z \geq t\right\} \leq \inf_{\theta>0} \left\{\mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{e}^{\theta(Z-\mathbb{E}\, Z)}\right\}. \tag{1.2.10}$$

Equation (1.2.10) is the essence of the Laplace transform method. Note that the moment generating function $\mathbb{E}\,\mathrm{e}^{\theta(Z-\mathbb{E}\, Z)}$ is also the Laplace transform of $Z - \mathbb{E}\, Z$, thus giving the name of this method.

1.2.2.2 Linear Decoupling of the Moment Generating Function

The moment generating function decouples naturally for a sum of independent random variables $Z = X_1 + \cdots + X_n$ where $\{X_i\}$ are independent:

$$\mathbb{E}\,\mathrm{e}^{\theta(Z-\mathbb{E}\, Z)} = \prod_{i=1}^{n} \mathbb{E}\,\mathrm{e}^{\theta(X_i-\mathbb{E}\, X_i)}. \tag{1.2.11}$$

Thus, we can instead control the deviation probability with the individual moment generating function $\mathbb{E}\,\mathrm{e}^{\theta(X_i - \mathbb{E}\,X_i)}$ by directing breaking down the independent sum on the right-hand side of (1.2.10)

$$\mathbb{P}\left\{Z - \mathbb{E}\,Z \geq t\right\} \leq \inf_{\theta>0} \left\{\mathrm{e}^{-\theta t} \cdot \prod_{i=1}^{n} \mathbb{E}\,\mathrm{e}^{\theta(X_i - \mathbb{E}\,X_i)}\right\}. \tag{1.2.12}$$

This step of linear decoupling is the key to deriving concentration results for a sum of independent random variables such as Hoeffding's inequality, Bernstein's inequality, etc. The remaining step is to control the individual moment generating function $\mathbb{E}\,\mathrm{e}^{\theta(X_i - \mathbb{E}\,X_i)}$ based on either the boundedness conditions of X_i or assumptions on the moments of X_i, substitute these upper bounds on the moment generating functions back into (1.2.12), and find the infimum over the parameter θ. In the case of the scalar Hoeffding's inequality, the associated assumption is that the random variables $\{X_i\}$ are bounded:

$$X_i \in [a_i, b_i] \quad \text{almost surely for all } i.$$

This boundedness assumption leads to the following bound of the individual moment generating function

$$\mathbb{E}\,\mathrm{e}^{\theta(X_i - \mathbb{E}\,X_i)} \leq \exp\left(\frac{\theta^2 (b_i - a_i)^2}{8}\right). \tag{1.2.13}$$

Substituting (1.2.13) into the right-hand side of (1.2.12) gives us a probability deviation bound that depends on the non-negative parameter θ only:

$$\mathbb{P}\left\{Z - \mathbb{E}\,Z \geq t\right\} \leq \inf_{\theta>0} \exp\left(-\theta t + \frac{\theta^2 \sum_{i=1}^{n}(b_i - a_i)^2}{8}\right).$$

Optimize and choose $\theta = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$. We arrive at the upper deviation probability of the Hoeffding's inequality:

$$\mathbb{P}\left\{Z - \mathbb{E}\,Z \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Applying the same argument to $-(Z - \mathbb{E}\,Z)$ gets the lower deviation.

1.2.3 The Entropy Method

In this section, we summarize the main ideas of the scalar entropy method based on the logarithmic-Sobolev inequality. As we mention before, the entropy method overcomes the limit of the Laplace transform method and leads to concentration inequalities for general functions of independent random variables. In Section 1.2.3.1, we first review the definition of entropy for a scalar random variables and an argument by Herbst that bounds the deviation probability of a random variable in terms of entropy. Then in Section 1.2.3.2, we derive the concentration inequality for product distributions that satisfy the logarithmic-Sobolev inequality. In Section 1.2.3.3, we demonstrate that a modified logarithmic-Sobolev inequality based on the subadditivity property of entropy leads to a Gaussian-type deviation probability for certain functions of independent random variables. Finally, we exhibit the generalization of the entropy to φ-entropy and related concentration results in Section 1.2.3.4.

1.2.3.1 Scalar Entropy, Deviation Probability Bound, and Herbst's Argument

For each nonnegative, real random variable Z, the entropy functional is defined as

$$H(Z) := \mathbb{E}(Z \log Z) - (\mathbb{E}\, Z) \log(\mathbb{E}\, Z). \tag{1.2.14}$$

The entropy method contains two steps. The first step is to control the deviation probability with the entropy of the random variable's moment generating function. Suppose $Y = Y(\boldsymbol{x})$ is a function that depends on a random vector $\boldsymbol{x} = (X_1, \dots, X_n) \in R^n$ and $\mathbb{E}\, Y = 0$. Just as in the Laplace transform method, we can use Markov's inequality and control the deviation probabilities of Y with the cumulant $\log \mathbb{E}\, \mathrm{e}^{\theta Y}$ such that for all $t > 0$,

$$\mathbb{P}\left\{Y \geq t\right\} \leq \inf_{\theta > 0} \exp\left(-\theta t + \log \mathbb{E}\, \mathrm{e}^{\theta Y}\right), \tag{1.2.15}$$

$$\mathbb{P}\left\{Y \leq -t\right\} \leq \inf_{\theta < 0} \exp\left(\theta t + \log \mathbb{E}\, \mathrm{e}^{\theta Y}\right). \tag{1.2.16}$$

Herbst [60] derived the following important relation that expresses the cumulant generating function of Y as in terms of the entropy of $\mathrm{e}^{\theta Y}$:

$$\log \mathbb{E}\,\mathrm{e}^{\theta Y} = \theta \int_0^\theta \frac{H(\mathrm{e}^{\beta Y})}{\mathbb{E}\,\mathrm{e}^{\beta Y}} \cdot \frac{\mathrm{d}\beta}{\beta^2}. \tag{1.2.17}$$

The major steps of the entropy method are deriving tight estimates for the entropy functional $H(\mathrm{e}^{\beta Y})$ usually in relation to the moment generating function $\mathbb{E}\,\mathrm{e}^{\beta Y}$, substituting the estimates into (1.2.17) and then into (3.3.12), and finally choosing the optimal tuning parameter θ to obtain probabilistic deviation inequalities of Y.

1.2.3.2 Gaussian Concentration From the Logarithmic-Sobolev Inequalities

There are two approaches to control the entropy function $H(\mathrm{e}^{\beta Y})$ with the moment generating function. The first approach is to apply the logarithmic-Sobolev inequalities for certain joint probability distributions. A probability distribution μ in $\mathbb{R}^n$ satisfies the logarithmc-Sobolev inequality [86] if $\boldsymbol{x}$ is a random vector distributed as μ and all continuously differentiable functions $f : \mathbb{R}^n \mapsto \mathbb{R}$ with the integrability condition $\mathbb{E}\left[f(\boldsymbol{x})^2 \log f(\boldsymbol{x})\right] < \infty$ satisfy the following inequality:

$$\mathbb{E}(f(\boldsymbol{x})^2 \log f(\boldsymbol{x})^2) - \mathbb{E}\,f(\boldsymbol{x})^2 \log \mathbb{E}\,f(\boldsymbol{x})^2 \leq c \cdot \mathbb{E}(\|\nabla f(\boldsymbol{x})\|^2), \tag{1.2.18}$$

where c is a positive constant. The left-hand side of inequality (1.2.18) can be expressed in terms of the entropy:

$$H\big(f^2(\boldsymbol{x})\big) \leq c \cdot \mathbb{E}\,\|\nabla f(\boldsymbol{x})\|^2\,. \tag{1.2.19}$$

If the distribution of the random vector $\boldsymbol{x}$ satisfies the logarithmic-Sobolev inequality with constant c and $Y(\boldsymbol{x})$ is Lipschitz continuous with constant L, then substitute $f(\boldsymbol{x}) = \mathrm{e}^{\theta Y(\boldsymbol{x})/2}$ into the right-hand side of (1.2.19) such that

$$\mathbb{E}\,\|\nabla f(\boldsymbol{x})\|^2 = \frac{\theta^2}{4} \cdot \mathbb{E}\left[\,\|\nabla Y\|^2 \cdot \mathrm{e}^{\theta Y}\right] \leq \frac{L^2\theta^2}{4} \cdot \mathbb{E}\,\mathrm{e}^{\theta Y}.$$

And we obtain the following bound that controls the entropy of $\mathrm{e}^{\theta Y(\boldsymbol{x})}$ by the moment generating function of $Y(\boldsymbol{x})$:

$$H\left(\mathrm{e}^{\theta Y}\right) \leq \frac{cL^2\theta^2}{4} \cdot \mathbb{E}\,\mathrm{e}^{\theta Y}.$$

Substitute this result into (1.2.17) and we control the cumulant

$$\log \mathbb{E}\,\mathrm{e}^{\theta Y} \leq \frac{cL^2\theta^2}{4}.$$

Then substitute into (3.3.12), set $\theta = 2t/cL^2$, and we arrive at the desired devation bound:

$$\mathbb{P}\left\{Y \geq t\right\} \leq \mathrm{e}^{-\frac{t^2}{cL^2}}.$$

The remaining question is to determine the set of distributions in $\mathbb{R}^n$ that satisfy the logarithmic-Sobolev inequality and quantify the corresponding constant c. Proving whether a multivariate distribution satisfies the logarithmic-Sobolev inequality is challenging. A sufficient condition is that the distribution exhibits the hypercontractivity property. Distributions that are known to satisfy the logarithmic-Sobolev inequality include the multivariate Gaussian distribution, whose logarithmic-Sobolev constant is $c = 2$, and the symmetric Bernoulli distribution. For an in-depth distribution of the logarithmic-Sobolev inequality, one can refer to the lecture notes of Ledoux [125] and Guionnet and Zegarlinski [87].

1.2.3.3 Bounded Differences Inequality from Modified Logarithmic-Sobolev Inequalities

To overcome the difficulty of establishing the logarithmic-Sobolev inequality for an arbitrary product distribution, another approach takes a divide-and-conquer method by taking advantage of the subadditivity property of the entropy functional and constructs a modified logarithmic-Sobolev inequality. This approach depends on two duality characterizations of the entropy functional with a supremum and infimum, which we exhibit in Section 1.2.3.3.1. We demonstrate that the supremum representation leads to the subadditivity property of entropy in Section 1.2.3.3.2, while the infimum representation bounds the conditional entropy and leads to a modified logarithmic-

Sobolev inequality in Section 1.2.3.3.3. Finally, we illustrate how the modified-Sobolev inequality leads to Gaussian-type deviation probabilities of Boucheron et. al. [26] in Section 1.2.3.3.4.

1.2.3.3.1 Two Duality Representations The first characterization is due to the convexity of the entropy functional and approximates $H(Z)$ from below with a set of linear functions of Z, each of which is characterized by a non-negative random variable U:

$$H(Z) = \sup_{U>0} \big\{ \mathbb{E}[(\log U - \log \mathbb{E}\, U)(Z - U)] + H(U) \big\}. \tag{1.2.20}$$

The second characterization approximates the entropy of Z from above with the infimum of functions that depend on both Z and another positive random variable U:

$$H(Z) = \inf_{U>0} \mathbb{E}\left[Z(\log Z - \log U) - (Z - U)\right]. \tag{1.2.21}$$

The supremum of (1.2.20) and the infimum of (1.2.21) are both attained when $U = Z$.

1.2.3.3.2 The Supremum Representation Establishes the Subadditivity Property of Entropy As we assume previously, $Z = \mathrm{e}^{\theta Y(X_1,\ldots,X_n)}$ is a function of independent random variables $X_1, \ldots, X_n$. The conditional entropy functional is defined with a conditional expectation instead:

$$H_i(Z) := \mathbb{E}_i(Z \log Z) - (\mathbb{E}_i\, Z) \cdot \log(\mathbb{E}_i\, Z),$$

where $\mathbb{E}_i = \mathbb{E}(\cdot|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ denotes the expectation taken with respect to X_i, while keeping all other $X_j (j \neq i)$ fixed. The first duality relation (1.2.20) leads to the subadditivity propert of the entropy, which is the following result that controls the total entropy $H(Z)$ by a sum of the individual conditional entropies:

$$H(Z) \leq \sum\nolimits_{i=1}^{n} \mathbb{E}[H_i(Z)]. \tag{1.2.22}$$

Controlling the total entropy $H(Z)$ for distributions that do not satisfy the logarithmic-Sobolev inequality (1.2.18) is difficult. The expectation is that one might be able to control the conditional entropy $H_i(Z)$ based on the properties of marginal distribution for each X_i. The first step of establishing (1.2.22) is to apply the duality relation to obtain the following Jensen-type inequality, which captures the convexity of the entropy functional:

$$\begin{aligned} H(\mathbb{E}_1 Z) &= \sup_{U>0}\{\mathbb{E}[(\log U - \log \mathbb{E}\, U)(\mathbb{E}_1 Z - U)] + H(U)\} \\ &\le \mathbb{E}_1 \sup_{U>0}\{\mathbb{E}[(\log U - \log \mathbb{E}\, U)(Z - U)] + H(U)\} \\ &= \mathbb{E}_1 H(Z). \end{aligned} \tag{1.2.23}$$

The first and third relations are both the duality relation (1.2.20). The second relation is because taking the supremum over a linear function is convex.

The second step is to use the convexity property (1.2.23) of entropy to break down the total entropy. The argument is iterative and the following steps of extracting the first conditional entropy in the sum of (1.2.22) contain the main ideas:

$$\begin{aligned} H(Z) &= \mathbb{E}[\varphi(Z) - \varphi(\mathbb{E}_1 Z) + \varphi(\mathbb{E}_1 Z) - \varphi(\mathbb{E}\, Z)] \\ &= \mathbb{E}[\mathbb{E}_1 \varphi(Z) - \varphi(\mathbb{E}_1 Z)] + \mathbb{E}[\varphi(\mathbb{E}_1 Z) - \varphi(\mathbb{E}\,\mathbb{E}_1 Z)] \\ &= \mathbb{E}\, H_1(Z) + H(\mathbb{E}_1 Z) \\ &\le \mathbb{E}\, H_1(Z) + \mathbb{E}_1 H(Z), \end{aligned}$$

where the last relation is due to the convexity property (1.2.23). Apply the same argument repeatedly to establish the subadditivity inequality:

$$\begin{aligned} H(Z) &\le \mathbb{E}\, H_1(Z) + \mathbb{E}_1 H(Z) \\ &\le \mathbb{E}\, H_1(Z) + \mathbb{E}\, H_2(Z) + \mathbb{E}_2 H(Z) \\ &\le \cdots \\ &\le \textstyle\sum_{i=1}^n \mathbb{E}[H_i(Z)]. \end{aligned}$$

1.2.3.3.3 The Infimum Representation Establishes a Modified Logarithmic-Sobolev Inequality With the integration representation (1.2.17) of the cumulant and the subadditivity property (1.2.22), we can use the upper bounds on the conditional entropy $H_i(Z)$ to obtain concentration inequalities for functions of independent random variables from a larger set of distributions. One common approach is to control each of the conditional entropies on the right-hand side of (1.2.22) by applying the second duality representation (1.2.21) conditionally, such that

$$H_i(Z) \leq \mathbb{E}_i\left[Z(\log Z - \log Z^{(i)}) - (Z - Z^{(i)})\right], \tag{1.2.24}$$

where

$$Z^{(i)} := \mathrm{e}^{\theta Y^{(i)}} \quad \text{and} \quad Y^{(i)} = Y(X_1, \dots, X_i', \dots, X_n),$$

and X_i' is an independent copy of X_i. Intuitively, $Y^{(i)}$ as previously defined is a local perturbation of Y by swapping one of the random variable X_i with an independent copy X_i'.

Conditioned on $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$, the joint distribution of $(Z, Z^{(i)})$ is the same as that of $(Z^{(i)}, Z)$, so we can apply the following symmetry argument to (1.2.24):

$$\begin{aligned} H_i(Z) &= \frac{1}{2} \cdot \Big[\mathbb{E}_i\left[Z(\log Z - \log Z^{(i)}) - (Z - Z^{(i)})\right] + \mathbb{E}_i\left[Z^{(i)}(\log Z^{(i)} - \log Z) - (Z^{(i)} - Z)\right]\Big] \\ &= \frac{1}{2}\,\mathbb{E}_i\left[(Z - Z^{(i)}) \cdot (\psi(Z) - \psi(Z^{(i)})\right], \end{aligned} \tag{1.2.25}$$

where $\psi(x) = 1 + \log x$. We substitute (1.2.25) into the right-hand side of the subadditivity inequality (1.2.22) and obtain the following modified logarithmic-Sobolev inequality:

$$H(Z) \leq \frac{1}{2} \sum\nolimits_{i=1}^{n} \mathbb{E}\left[(Z - Z^{(i)})(\psi(Z) - \psi(Z^{(i)})\right]. \tag{1.2.26}$$

1.2.3.3.4 Boucheron's Gaussian-Type Concentration Inequalities The inequality (1.2.26) allows us to use second order statistics that characterize local-variations of Y to control its entropy. We introduce the approach of Boucheron et. al. [28, 29]

which depends on a variant of (1.2.26) that we derive in the following steps:

$$\begin{aligned} H(Z) &\leq \frac{1}{2}\sum\nolimits_{i=1}^{n}\Big[\mathbb{E}\left[(Z - Z^{(i)})_+(\psi(Z) - \psi(Z^{(i)})\right] + \mathbb{E}\left[(Z - Z^{(i)})_-(\psi(Z) - \psi(Z^{(i)})\right]\Big] \\ &= \frac{1}{2}\sum\nolimits_{i=1}^{n}\Big[\mathbb{E}\left[(Z - Z^{(i)})_+(\psi(Z) - \psi(Z^{(i)})\right] + \mathbb{E}\left[(Z^{(i)} - Z)_-(\psi(Z^{(i)}) - \psi(Z)\right]\Big] \\ &= \sum\nolimits_{i=1}^{n}\mathbb{E}\left[(Z - Z^{(i)})_+(\psi(Z) - \psi(Z^{(i)})\right], \qquad (1.2.27) \end{aligned}$$

where $X_+ = \max\{0, X\}$ and $X_- = \min\{0, X\}$ are the positive and negative parts of X respectively. In the first relation, we break down the summand into two parts depending on the sign of $(Z - Z^{(i)})$. The second relation relies on the fact that the joint distribution of $(Z, Z^{(i)})$ is the same as that of $(Z^{(i)}, Z)$ such that we can swap Z with $Z^{(i)}$ inside the second expectation. We combine the two expectations in the last relation by identifying $(Z^{(i)} - Z)_-$ with $-(Z - Z^{(i)})_+$.

Boucheron et. al. also define the following two non-negative random variables which are functions of $\boldsymbol{x} = (X_1, \ldots, X_n)$ and characterize the up-side and down-side second-order local variations of the random variable Y:

$$V_+(\boldsymbol{x}) = \mathbb{E}\left[\sum\nolimits_{i=1}^{n}(Y - Y^{(i)})^2 \cdot 1_{Y \geq Y^{(i)}} \middle| \boldsymbol{x}\right], \qquad (1.2.28)$$

$$V_-(\boldsymbol{x}) = \mathbb{E}\left[\sum\nolimits_{i=1}^{n}(Y - Y^{(i)})^2 \cdot 1_{Y < Y^{(i)}} \middle| \boldsymbol{x}\right]. \qquad (1.2.29)$$

The random variables V_+ and V_- can be interpreted as one-sided conditional variance. The first V_+ implicitly controls the upper deviation probability via the conditional entropy while V_- controls the lower deviation because (1.2.27) leads to

$$H(\mathrm{e}^{\theta Y}) \leq \theta^2\,\mathbb{E}\left[V_+ \cdot \mathrm{e}^{\theta Y}\right], \quad \text{when } \theta > 0, \qquad (1.2.30)$$

$$H(\mathrm{e}^{\theta Y}) \leq \theta^2\,\mathbb{E}\left[V_- \cdot \mathrm{e}^{\theta Y}\right], \quad \text{when } \theta < 0. \qquad (1.2.31)$$

If we can bound V_+ and V_- by constants C_+ and C_- uniformly, we control the entropy by substituting (1.2.30) and (1.2.31) into the subadditivity inequality(1.2.22):

$$H\left(\mathrm{e}^{\theta Y}\right) \leq C_+\theta^2 \cdot \mathbb{E}(\mathrm{e}^{\theta Y}) \quad \text{when } \theta > 0, \qquad (1.2.32)$$

$$H\left(\mathrm{e}^{\theta Y}\right) \leq C_-\theta^2 \cdot \mathbb{E}(\mathrm{e}^{\theta Y}) \quad \text{when } \theta < 0. \qquad (1.2.33)$$

The final step is to substitute (1.2.32) and (1.2.33) into (1.2.17) and then (3.3.12) to arrive at the concentration bounds:

$$\mathbb{P}\left\{Y > t\right\} \leq \mathrm{e}^{-t^2/4C_+} \quad \text{and} \quad \mathbb{P}\left\{Y < -t\right\} \leq \mathrm{e}^{-t^2/4C_-} \quad \text{for all } t > 0.$$

In their paper, Boucheron et. al. also show that one can obtain more refined concentration bounds by correspondingly tightening the upper bound on V_+ and V_-.

1.2.3.4 Φ-Entropy and Moment Bounds

It turns out that in addition to the logarithmic entropy (1.2.14), the supremum representation (1.2.20) and the subadditivity property (1.2.22) exist for a larger set of entropy functionals. Let $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}$ be a convex function. The φ-entropy functional [122], which is a general class of entropy, is defined as

$$H_\varphi(Z) := \mathbb{E}\,\varphi(Z) - \varphi(\mathbb{E}\,Z),$$

while with the conditional φ-entropy is defined as

$$H_{\varphi,i} := \mathbb{E}_i\,\varphi(Z) - \varphi(\mathbb{E}_i\,Z)$$

for a product probability distribution. The subadditivity property of the φ-entropy is the following inequality:

$$H_\varphi(Z) \leq \sum\nolimits_{i=1}^{n} \mathbb{E}[H_{\varphi,i}(Z)].$$

The logarithmic entropy (1.2.14) is a special case with $\varphi : t \mapsto t \log t$. Researchers including Latała & Oleszkiewicz [122], Chafaï [47, 48], and Boucheron et al. [26] established the conditions that the function φ shall satisfy such that the φ-entropy functional is subadditive and as a result extended the scalar entropy method. In particular, the power functions $\varphi : t \mapsto t^p$ with $p \in [1, 2]$ belong to this function class. Based on the power functions, Latała & Oleszkiewiczï [122] delineate the set of distributions that satisfy a new set of φ-Sobolev inequality, which interpolates between the logarithmic-Sobolev inequality that controls the entropy of a function and the

Poincaré inequality that controls the variance of a function. The authors also established that the subadditivity of the φ-entropy functional leads to the tensorization of the φ-Sobolev inequalities to product measure and the φ-Sobolev inequalities produce probabilities tail bounds between Gaussian-type and exponential decay.

The φ-entropy functional with the corresponding φ-Sobolev inequality expands the set of concentration inequalities obtained via the entropy method. In particular, the power functions $\varphi : t \mapsto t^p$ with $p \in [1, 2]$ lead to polynomial concentration inequalities that control the moment growth of functions of independent random variables. For example, Boucheron et. al. [26] proved that if there exists a constant C such that controls the two variation quantities $V_+ \leq C$ and $V_- \leq C$ almost surely, then the following moment inequality holds for Y:

$$\big(\mathbb{E}\, Y^q\big)^{1/q} \leq 2^{1/q} \cdot \sqrt{\frac{qC}{\mathrm{e} - \sqrt{\mathrm{e}}}}. \tag{1.2.34}$$

This result says that higher-order moments of the random variable Y can be controlled by second-order statistics.

Chapter 2

Context and Impact

In this chapter, we first summarize the field of matrix concentration inequalities in Section 2.1. We focus on the development of the theories. In particular, we exhibit two important approaches of deriving concentration inequalities for random matrices. The first approach is based on the powerful Lieb's Theorem and parallels the scalar argument of developing classical concentration inequalities. The second approach is the method of exchangeable pairs. Our goal is to provide the theoretical background based on which the major results of the thesis develop.

Next, we illustrate the first two main results of the thesis in Section 2.2, the analysis of the masked sample covariance estimator and the work on deriving matrix concentration inequalities with matrix φ-entropy that appear in Chapter 3 and Chapter 4 respectively.

2.1 Matrix Concentration Inequalities

We begin this section by discussing the fruitful development of matrix concentration inequalities in the recent decades in Section 2.1.1. Afterwards, we recap the major technical ingredients for the matrix Laplace transform method in Section 2.1.2 and the matrix exchangeable pairs method in Section 2.1.3.

2.1.1 History of Matrix Concentration Inequalities

As we mentioned in Chapter 1, there are two major types of problems in studying the concentration inequalities of matrix functions. The first is deriving large deviation probabilities that control the fluctuations of a random matrix around its mean

$$\mathbb{P}\left\{\|\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}\| \geq t\right\},$$

where the deviation is measured in the matrix spectral norm. The second type of problem is deriving upper bounds for the matrix moments

$$\mathbb{E}\,\|\boldsymbol{X}\|_p^p \quad \text{for } p \in \mathbb{Z}_+.$$

Then the Shatten p-norm [99] for a $d \times d$ matrix $\boldsymbol{A}$ is defined as

$$\|\boldsymbol{A}\|_p := \left\| \sum\nolimits_{i=1}^{n} s_i^p(\boldsymbol{A}) \right\|^{1/p},$$

where $s_1(\boldsymbol{A}) \geq \cdots \geq s_d(\boldsymbol{A}) \geq 0$ are the singular values of $\boldsymbol{A}$. These two problems are dual approaches to obtain matrix concentration results because integrating the large deviation probabilities leads to matrix moment bounds, while matrix moment bounds also controls the large deviation probabilities via the Markov inequality.

Matrix concentration inequalities lie in the setting of developing non-commutative concentration inequalities. We first identify the different classes of non-commutativity in Section 2.1.1.1. Next, we review the line of research that develops non-commutative moment bounds in Section 2.1.1.2. We summarize the development of non-commutative large deviation inequalities in Section refsection:noncommutative-prob. We mention some other related results in 2.1.1.4. Finally, we lay out the notations for this Chapter in Section 2.1.1.5.

2.1.1.1 Classification of Non-Commutative Situations

As Junge & Zeng [110] mentioned, random matrices are semi-commutative, in the sense that random matrices retain randomness from classical probability, which is commutative. This semi-commutativity lies between the commutative scalar algebra and a full non-commutative algebra space.

There are two types of full non-commutative algebra space. The first type, called tracial non-commutativity [178], is defined on a semi-finite von Neumann algebra $\mathcal{M}$ with a normal faithful semi-finite trace τ which is a function from $\mathcal{M}$ to the complex set $\mathbb{C}$ and is invariant under cyclic permutation. As a result, there is a unit element $\infty \in \mathcal{M}$ such that $\tau(1) = 1$ and for any two elements $x, y \in \mathcal{M}$, $\tau(xy) = \tau(yx)$. The von Neumann $\mathcal{M}$ algebra together with the trace τ generates a non-commutative L^p space. The L^p norm is defined as

$$\|x\|_p = \big(\tau(|x|^p)\big)^{1/p}, \quad \text{for any } x \in L^p(\mathcal{M}, \tau).$$

One can observe that this tracial category contains the semi-commutative case of random matrices. The non-commutative matrix algebra $\mathbb{R}^{d\times d}$ together with the normalized matrix trace $\bar{\text{tr}}$, which is the division of the sum of the matrix diagonal entries and the matrix dimension, generates the L^p space for $d \times d$ matrices.

A second type of full non-commutativity is non-tracial [110], such that a non-commutative L^p space is associated with a von Neumann algebra equipped with a faithful normal state. In this setting, there does not exist a trace function with the cyclical invariant property. The non-tracial setting of non-commutativity is more general than the tracial non-commutativity, such that any non-tracial non-commutative moment bound also holds in the tracial setting.

2.1.1.2 Non-Commutative Moment Inequalities

Before the appearance of matrix Laplace transform bounds, researchers focused on deriving noncommutative moment inequalities. We review the non-commutative Khintchine inequality in 2.1.1.2.1 and the Burkholder/Rosenthal inequality in 2.1.1.2.2.

2.1.1.2.1 Non-Commutative Khintchine Inequality Deriving non-commutative moment inequalities starts with the classical Khintchine inequality (1.2.3), which as we recall bounds the moments of a Rademacher sum or Gaussian sum. Most research considers the tracial setting of non-commutative Khintchine inequality, which translates into bounding the L_p norm of a matrix Rademacher or Gaussian sum

$$\left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i\right\|_{L^p} = \left(\mathbb{E}\left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i\right\|_p^p\right)^{1/p} \tag{2.1.1}$$

in the non-commutative matrix algebra, where $\{\varepsilon_i\}$ is a Rademacher or Gaussian series and without loss of generality, $\{\boldsymbol{X}_i\}$ is a sequence of deterministic self-adjoint matrices of the same dimension. As in the scalar setting, the behavior of (2.1.1) varies with different values of p.

Tomczak–Jaegermann's work [227] was a first attempt to generalize the Khintchine inequality to the noncommutative matrix scenario. In her paper, she considers bounding (2.1.1) when $\{\varepsilon_i\}$ is a Rademacher series using the following term:

$$\left(\sum\nolimits_{i=1}^{n} \|\boldsymbol{X}_i\|_{2p}^2\right)^{1/2} \tag{2.1.2}$$

for $p \in [1, \infty)$. She established that there exists constants C_p, C_p' such that

$$\begin{aligned} \left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i\right\|_{L^p} &\le C_p \left(\sum\nolimits_{i=1}^{n} \|\boldsymbol{X}_i\|_p^2\right)^{1/2}, \quad \text{for } p \ge 2, \\ \left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i\right\|_{L^p} &\ge C_p' \left(\sum\nolimits_{i=1}^{n} \|\boldsymbol{X}_i\|_p^2\right)^{1/2}, \quad \text{for } 1 \le p \le 2. \end{aligned}$$

After Tomczak–Jaegermann, Lust–Piquard [136] and Lust–Piquard & Piser [137] made an important improvement and established a qualitatively sharper version of noncommutative Khintchine's inequality for the regime $1 \le p < \infty$. They demonstrate that if $2 \le p \le \infty$

$$C_1(p) \cdot \left\|\left(\sum\nolimits_{i=1}^{n} \boldsymbol{X}_i^2\right)^{1/2}\right\|_p \le \left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i\right\|_{L^p} \le C_2(p) \cdot \left\|\left(\sum\nolimits_{i=1}^{n} \boldsymbol{X}_i^2\right)^{1/2}\right\|_p, \tag{2.1.3}$$

where $C_1(p), C_2(p)$ are functions of p. When $1 \le p \le 2$, the non-commutative Khintchine inequality takes another form:

$$C_1'(p) \cdot \inf_{\boldsymbol{X}_i = \boldsymbol{A}_i + \boldsymbol{B}_i} \left\{ \left\|\left(\sum\nolimits_i \boldsymbol{A}_i^* \boldsymbol{A}_i\right)^{1/2}\right\|_p + \left\|\left(\sum\nolimits_i \boldsymbol{B}_i^* \boldsymbol{B}_i\right)^{1/2}\right\|_p \right\} \le \left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i\right\|_{L^p} \tag{2.1.4}$$

$$\le C_2'(p) \cdot \inf_{\boldsymbol{X}_i = \boldsymbol{A}_i + \boldsymbol{B}_i} \left\{ \left\|\left(\sum\nolimits_i \boldsymbol{A}_i^* \boldsymbol{A}_i\right)^{1/2}\right\|_p + \left\|\left(\sum\nolimits_i \boldsymbol{B}_i^* \boldsymbol{B}_i\right)^{1/2}\right\|_p \right\}. \tag{2.1.5}$$

Lust–Piquard and Pisier showed that the same bounds hold for a matrix Gaussian sum. When $p \ge 2$, the lower bound is given with $C_1(p) = 1$ and when $p \le 2$, the upper bound is give with $C_2'(p) = 1$. Proving other side of these two inequalities and obtaining the optimal constants are more difficult. Lust–Piquard and Pisier [137] established that when $q \to \infty$, $C_2(p) = \mathcal{O}(\sqrt{q})$. Buchholz [34] demonstrated that when $p > 2$ is an even integer, the optimal value for $C_2(p)$ is the same as in the scalar Khintchine inequality. The case $1 < q < 2$ is proved using a duality argument and Haagerup & Musat [89] established the optimal constant $C_1'(p) = \sqrt{2}$ for $p = 1$. A recent work by Pisier & Ricard [177] established the non-commutative Khintchine inequalities in the L_p space for all $0 < p < 1$. Their result takes the form of (2.1.5). The works of Lust–Piquard and other authors rely heavily on noncommutative probability, and thus the study of non-commutative Khintchine inequality not only benefits the random matrix theory but also advances other fields such as non-commutative algebra.

We mention that using the noncommutative Khintchine inequality directly in applications can be cumbersome and requires delicate matrix norm estimation. Rudelson [193] translates the noncommutative Khintchine inequality into a moment bound for a Radamacher sum of rank-1 matrices $\{\boldsymbol{x}_i\boldsymbol{x}_i^*\}$, which is easier to use in application. We illustrate the following version that appears in [230]:

$$\left(\mathbb{E}\left\|\sum\nolimits_{i=1}^{n}\varepsilon_i\boldsymbol{x}_i\boldsymbol{x}_i^*\right\|^p\right)^{1/p} \leq C\sqrt{p}\cdot\max_i\|\boldsymbol{x}_i\|\,\|\boldsymbol{X}\|\,, \quad p \geq 2\log n, \tag{2.1.6}$$

where $\{\varepsilon_i\}$ is a sequence of Rademacher random variables and $\{\boldsymbol{x}_i\}$ form the columns of the matrix $\boldsymbol{X}$.

2.1.1.2.2 Non-Commutative Inequality for Martingales Besides the non-commutative Khintchine inequality, researchers have extended other classical moment inequalities to the non-commutative algebra, for example, the Rosenthal inequality (1.2.4) for the L_p norm of the sum of independent mean-zero random variables and its martingale counterpart, the Burkholder–Gundy inequality. We summarize the development in this direction.

Pisier & Xu [178] established the Burkholder–Gundy inequality for non-commutative martingales in the tracial setting, which implies a non-commutative version of the Rosenthal inequality. They show that the Burkholder–Gundy inequality also exhibits two different forms for $p \geq 2$ and $1 < p \leq 2$. The order of constants in the tracial setting are characterized in [183]. Junge & Xu [110] extended the non-commutative Burkholder–Gundy/Rosenthal inequality to the non-tracial setting. Junge & Xu [111] and Nandrianantoanina [184] determined the the optimal order of constants.

2.1.1.3 Non-Commutative Large Deviation Inequalities

The recent development of large deviation inequalities in the matrix algebra has emerged as a user-friendly approach to study matrix concentration inequalities. Specialized to the semi-commutative matrix algebra, we summarize the development of an approach using matrix Laplace transform method and Lieb's theorem in Section 2.1.1.3.1 and the method of exchangeable pair in Section 2.1.1.3.2. Finally, we mention several recent results that extend the matrix large deviation inequalities to the full non-commutative setting in Section 2.1.1.3.3.

2.1.1.3.1 Matrix Laplace Transform Method and Lieb's Theorem In developing matrix large deviation probabilities, researchers constantly draw inspirations from existing methods for developing scalar concentration results. One important approach stems from the scalar Laplace transform method that we present in Chapter 1 and eventually crystalizes as the matrix Laplace transform method. The scalar argument of deriving classical concentration inequalities based on the Laplace transform method and linear decoupling of the moment generating function is especially suited to control deviation probabilities for sums of independent random variables. In order to develop a matrix version of this argument such that one can control the deviation probabilities of sums of independent random matrices, we need two ingredients. The first ingredient is to construct a matrix version of the Laplace transform bounds, which control the deviation probability of a random matrix. The second ingredient is an appropriate decoupling method, such that the matrix moment generating function of a sum matrix can be controlled by the moment generating function of the summands. The second ingredient will be more involved than the corresponding step for scalar random variables due to the noncommutative property of matrices.

The development of the first ingredient goes back to Ahlswede and Winter [1]. Their paper contains the first appearances of what would later be called the matrix moment generating function. The authors developed the following argument, called the Bernstein trick, which bounds the following comparison probability of two Hermitian random matrix $\boldsymbol{Y}$ and $\boldsymbol{B}$

$$\mathbb{P}\{\boldsymbol{Y} \not\preceq \boldsymbol{B}\} \leq \mathbb{E}\,\mathrm{tr}\exp(\boldsymbol{TYT}^{*} - \boldsymbol{TBT}^{*}), \tag{2.1.7}$$

where $\boldsymbol{T}$ is a matrix of compatible dimension such that $\boldsymbol{T}^*\boldsymbol{T} \succcurlyeq \mathbf{0}$. The bound (2.1.7) contains the spirit of the scalar Laplace transform method and based on (2.1.7), the authors developed a deviation probability bound for a sum of independent random matrices, where they relied on the following Golden–Thompson inequality to decouple the matrix moment generating function of the sum matrix:

$$\operatorname{tr}(\mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}}) \leq \operatorname{tr}(\mathrm{e}^{\boldsymbol{A}}\,\mathrm{e}^{\boldsymbol{B}}), \tag{2.1.8}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are Hermitian matrices of the same dimension. However, the authors mentioned that this approach of decoupling is not optimal and they presented several conjectures for improved decoupling.

After the work of Ahlswede and Winter, Oliveira [164] refined the matrix Laplace bounds. Oliveira's version, which we display in Proposition 2.1.2, has become the standard interface connecting the matrix large deviation probabilities with the matrix moment generating functions, based on which later results of matrix large deviation probability start to flourish. In addition, his work provided a simpler proof of Rudelson's Khintchine lemma. Other authors apply the approach of Ahlswede and Winter to obtain multiple matrix concentration inequalities. For example, Christofides and Markström [57] derive a Hoeffding-type inequality for a sum of bounded independent random matrices.

A major breakthrough appears in the work of Tropp [231, 233]. Working with the matrix cumulant functions instead, Tropp developed an improved decoupling argument based on a deep theorem due to Lieb [131], which we exhibit as Theorem 2.1.3 in Section 2.1.2.3, and produced a plethora of large deviation probabilities for sums of independent random matrices that are easily applicable to different application scenarios. These results include matrix Hoeffding inequality, matrix Bernstein inequality, matrix Chernoff inequality, etc., and they are direct matrix extensions from their classical scalar counterparts. These matrix deviation probabilities are considerably sharper than previous results.

In addition to concentration results for independent matrix sums, Tropp's decoupling approach also applies to weakly dependent sequences such as matrix martingales. He obtained a matrix version of the Azuma's inequality, which leads to a bounded difference inequality for matrix function of independent random variables. In a separate paper [231], Tropp established the matrix Freedman's inequality, which is a Bernstein-type bound for matrix martingales.

In Tropp's concentration results, the deviation probability bounds depend directly on the ambient dimension of the random matrices. Authors include Hsu et. al. [103] and Minsker [158] derive improved concentration probabilities that are tighter for degenerate matrix distributions that concentration in a low-dimensional subspace. For a complete coverage of the method of deriving matrix concentration inequalities based on the matrix Laplace transform method and the Lieb's theorem, we refer readers to Tropp's monograph [235]. We also provide a detailed discussion in Section 2.1.2

2.1.1.3.2 Method of Exchangeable Pairs Another approach of deriving matrix concentration inequalities extends from the method of exchangeable pairs [51] for developing scalar concentration inequalities that we mentioned in Chapter 1. It turns out that this approach is not limited to scalar random variables. The paper of Mackey et. al. [138] developed the matrix version of the exchangeable pairs method. The matrix exchangeable pairs method also depends on the matrix Laplace transform bounds to relate the large deviation probabilities to the matrix moment generating function, but the exchangeable pairs provide a different approach to control the matrix moment generating function compared with that of [233]. The major results in [138] include matrix large deviation probabilities such as matrix Hoeffding, matrix Bernstein, etc. In addition, the authors establish a matrix version of Burkholder–Davis–Gundy inequality [37] which leads to a very simple proof of the non-commutative Khintchine inequality. Another result is the matrix Rosenthal inequality that controls the moments of a sum of random matrices with second order bounds of the individual matrix. Tropp's recent manuscript [234] isolates this matrix moment bound in a streamlined presentation.

In [173], Paulin et. al. extended the concept of matrix Stein pair to the Kernel Stein pair. The authors show that it is feasible to extend the method of Markov chain decoupling, due to Chatterjee [51], to generate Kernel Stein pairs for a larger set of random matrices. They establish multiple matrix moment bounds that are matrix versions of classical Efron–Stein inequalities. Their results also include an improved version of the bounded difference inequality for random matrices. We discuss the technical details of the method of exchangeable pairs in Section 2.1.3.

2.1.1.3.3 Large Deviation Inequalities in the Full Non-Commutative Setting Matrix large deviation inequalities can also be extended to the full non-commutative setting. Junge and Zeng [115] extended the Bernstein, Bennett inequalities for a sum of independent random variables to the tracial non-commutative setting. Similar to the random matrix case, their proof also relies on the Laplace transform method applied to the non-commutative setting. The authors pointed out that Lieb's theorem only applies to the random matrix case which has commutative randomness, and instead the authors depended on the Golden–Thompson inequality (2.1.8) generalized to the tracial non-commutative setting to develop the non-commutative counterparts of Bernstein, Bennett inequalities. In [114], Junge and Zeng obtained a martingale version of non-commutative Bernstein inequality. In addition, they also derived a non-commutative Poincaré inequality and a non-commutative transportation cost inequality. These new results extended their scalar counterparts to the tracial non-commutative setting. Sadeghi and Sal Moslehian [196] also developed Azuma-type inequalities for tracial non-commutative martingales, which implies a non-commutative Hoeffding's inequality and a non-commutative McDiarmid inequality. Their approach also relies on the Golden–Thompson inequality.

2.1.1.4 Other Results

We also mention that Tropp's recent manuscript [236] contains an improvement of the Khintchine inequality for random matrices. For a Hermitian Gaussian matrix sum $\boldsymbol{X} = \sum_{i=1}^{n} \gamma_i \boldsymbol{H}_i$ where $\{\gamma_i\}$ are independent standard normal variables, the author goes beyond the matrix variance and develops a matrix alignment parameter $w_q(\boldsymbol{X})$ which measures the degree of commutativity for a series of deterministic matrices:

$$w_q(\boldsymbol{X}) := \max_{\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3} \left\| \left| \sum\nolimits_{i,j=1}^{n} \boldsymbol{H}_i \boldsymbol{Q}_1 \boldsymbol{H}_j \boldsymbol{Q}_2 \boldsymbol{H}_i \boldsymbol{Q}_3 \boldsymbol{H}_j \right| \right\|_q, \quad q \geq 1,$$

where $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3$ are unitary matrices. The idea is that the higher the degree of commutativity in the matrix series, the more similar the matrix Gaussian sum behaves to their scalar counterparts. The second-order matrix Khintchine inequality developed in [236] controls the high-order matrix moments of the Hermitian Gaussian matrix sum $\boldsymbol{X}$ with both the traditional matrix variance and the matrix alignment parameter:

$$\left(\mathbb{E}\,\|\boldsymbol{X}\|_{2p}^{2p}\right)^{1/(2p)} \leq C_1 \cdot p^{1/4} \cdot \sigma_{2p}(\boldsymbol{X}) + C_2\sqrt{p} \cdot w_{2p}(\boldsymbol{X}), \quad p \geq 3, \tag{2.1.9}$$

where

$$\sigma_q(\boldsymbol{X}) = \left\| \left(\sum\nolimits_{i=1}^{n} \boldsymbol{H}_i^2\right)^{1/2} \right\|_q$$

is the Shatten q-norm of the matrix variance of $\boldsymbol{X}$. The author demonstrated that this new Khintchine's inequality is never a significantly worse bound than the matrix Khintchine inequality in [138] and is sharper for more commutative matrix series, such as the GOE matrices.

In the remainder of this section, we restrict our attention to matrix non-commutativity and provide detailed summaries for both the method based on Lieb's Theorem (Section 2.1.2) and the method of exchangeable pairs (Section 2.1.3) within the relevant scope of the thesis.

2.1.1.5 General Notations

We lay out the notations in this chapter. Capitalized bold letters such as $\boldsymbol{X}$ denote matrices. The set $\mathbb{H}^d$ is the linear space of self-adjoint $d \times d$ matrices. The maximum and minimum eigenvalues of $\boldsymbol{X}$ are denoted as $\lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X})$. The normalized trace of $\boldsymbol{X} \in \mathbb{H}^d$ is

$$\bar{\operatorname{tr}}(\boldsymbol{X}) = \frac{1}{d} \cdot \sum\nolimits_{i=1}^{d} x_{ii}.$$

Curly inequalities such as $\succcurlyeq, \preccurlyeq$ denote matrix comparisons in the positive-semidefinite order. Finally, we point out that any scalar function $f : \mathbb{R} \mapsto \mathbb{R}$ can generate a standard matrix function $f : \mathbb{H}^d \mapsto \mathbb{H}^d$ for any dimension d by operating on the spectrum of the self-adjoint matrix.

2.1.2 Matrix Concentration Inequalities via Lieb's Theorem

In this section, we summarize the method of deriving concentration inequalities based on the Lieb's theorem. Similar to the scalar Laplace transform method, there are two major steps to derive large deviation probability bounds for sums of independent random matrices using this method. The first step is bounding the large deviation probability with the matrix moment generating function, as we describe in Section 2.1.2.2. The second step is to decouple the matrix moment generating function of the sum matrix into the individual moment generating functions, which can be controlled more easily with assumptions on the individual matrix. The second step differs from the scalar approach as it depends on a deep concavity result, Lieb's theorem. We explain the details in Section 2.1.2.3. Next, Section 2.1.2.4 presents a master probability tail bound for sums of independent random matrices by combining the matrix probability bound and the decoupling step. Finally, in Section 2.1.2.5, we summarize the major concentration inequalities obtained via this approach. To start with, we review the definition of the matrix moment generating function and matrix cumulant in Section 2.1.2.1.

2.1.2.1 Matrix Moment Generating Function

The following restates the definition of the matrix moment generating function and the matrix cumulant [233, 138]. In comparison to their scalar counterparts, we have an additional trace function that aggregates the diagonals of the matrix exponential.

Definition 2.1.1 (Matrix Moment Generating Function and Matrix Cumulant)**.** *Let $\boldsymbol{X}$ be a self-adjoint random matrix. The normalized trace moment generating function of $\boldsymbol{X}$ is defined as*

$$m(\theta) := m_{\boldsymbol{X}}(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\mathrm{e}^{\theta \boldsymbol{X}} \quad \textit{for } \theta \in \mathbb{R}. \tag{2.1.10}$$

The matrix cumulant function is

$$c(\theta) := c_{\boldsymbol{X}}(\theta) := \log \mathbb{E}\,\bar{\mathrm{tr}}\mathrm{e}^{\theta \boldsymbol{X}}, \quad \textit{for } \theta \in \mathbb{R}. \tag{2.1.11}$$

2.1.2.2 Matrix Laplace Bound

The following proposition appears in [138, Proposition 3.3] and it encapsulates the essence of the matrix Laplace transform method, which provides a variational upper bound on the large deviation probability of random matrices using the matrix moment generating function.

Proposition 2.1.2. *Let $\boldsymbol{X}$ be a self-adjoint random matrix in $\mathbb{H}^d$. For all $t \in R$,*

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}) \geq t\right\} \leq d \cdot \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot m_{\boldsymbol{X}}(\theta), \tag{2.1.12}$$

$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{X}) \leq t\right\} \leq d \cdot \inf_{\theta<0} \mathrm{e}^{-\theta t} \cdot m_{\boldsymbol{X}}(\theta). \tag{2.1.13}$$

Proposition 2.1.2 translates the problem of estimating the deviation probabilities of a random matrix into that of bounding its matrix moment generating function, which opens up many new approaches of bounding large deviation probabilities. Due to its importance, we provide a complete proof of the proposition.

Proof. The main step of the proof is also the application of the Markov inequality to control the probability tail bound, which appears in the following first inequality. Compared with the scalar Laplace transform method, the proof of Proposition 2.1.2 relies on some additional matrix algebra. In the second equality of the following derivation, standard functions operate on the eigenvalues of self-adjoint matrices, thus the exponent of the maximum eigenvalue is equal to the maximum eigenvalue of the exponent. In the last relation, we bound the maximum eigenvalue of a positive matrix with its trace.

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}) \geq t\right\} &= \mathbb{P}\left\{\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})} \geq \mathrm{e}^{\theta t}\right\} \leq \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})} \\ &= \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\lambda_{\max}(\mathrm{e}^{\theta\boldsymbol{X}}) \leq \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta\boldsymbol{X}} = d \cdot \mathrm{e}^{-\theta t} \cdot m_{\boldsymbol{X}}(\theta).\end{aligned} \tag{2.1.14}$$

Taking the infimum over θ in the last relation of (2.1.14) establishes (2.1.12). The same argument establishes the lower deviation. □

2.1.2.3 Lieb's Theorem and the Subadditivity of Matrix Cumulant Generating Function

Unlike in the scalar case where the moment generating function of sums of independent random variables decouples naturally, the noncommutativity of matrices poses significant challenges to decouple the moment generating function. Ahlswede & Winter [1] achieved decoupling using the Golden–Thompson inequality (2.1.8). Tropp [233] took a different approach based on the following concavity theorem due to Lieb [131].

Theorem 2.1.3 (Lieb's Theorem)**.** *Suppose $\boldsymbol{H} \in \mathbb{H}^d$ is a fixed deterministic self-adjoint matrix. The function*

$$\boldsymbol{A} \mapsto \bar{\operatorname{tr}} \exp(\boldsymbol{H} + \log \boldsymbol{A})$$

is a concave map on the set $\mathbb{H}_+^d$ of $d \times d$ positive-definite matrices.

Lieb established this result in the context of solving the Wigner–Yanase–Dyson conjecture. Tropp [232] provided an alternative proof based on the joint convexity property of the quantum relative entropy. With the help of Jensen's inequality, Lieb's theorem translates into the following inequality:

$$\mathbb{E}\, \bar{\operatorname{tr}} \exp(\boldsymbol{H} + \boldsymbol{X}) \leq \bar{\operatorname{tr}} \exp(\boldsymbol{H} + \log \mathbb{E}\, \mathrm{e}^{\boldsymbol{X}}), \tag{2.1.15}$$

where $\boldsymbol{H}$ is a deterministic self-adjoint matrix and $\boldsymbol{X}$ is self-adjoint and random.

The inequality (2.1.15) leads to the following decoupling lemma by Tropp [233, Lemma 3.4], which achieves a sharper decoupling of the moment generating function of an independent sum using the matrix cumulant function compared with using the Golden–Thompson inequality.

Lemma 2.1.4. *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices of the same dimension. Then the matrix mgr of the matrix sum decouples as follows*

$$\mathbb{E}\, \bar{\operatorname{tr}} \exp\left(\sum\nolimits_k \theta \boldsymbol{X}_k\right) \leq \bar{\operatorname{tr}} \exp\left(\sum\nolimits_k \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k}\right) \quad \textit{for } \theta \in \mathbb{R}. \tag{2.1.16}$$

2.1.2.4 Probability Bounds for Sums of Independent Random Matrices

Substituting (2.1.16) of the subadditivity lemma into the matrix Laplace bound, Proposition 2.1.2, we arrive at the following theorem, which is the master probability bounds in [233, Theorem 3.6]. This theorem controls the matrix large deviation probabilities with the matrix moment generating functions of the individual matrices $\{\boldsymbol{X}_k\}$ and becomes the standard interface for various scenarios of sums of independent random matrices.

Theorem 2.1.5. *Suppose $\{\boldsymbol{X}_k\}$ is a finite sequence of independent, random matrices in $\mathbb{H}^d$.*

Then for all $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} \leq d \cdot \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \bar{\operatorname{tr}} \exp\left(\sum\nolimits_k \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k}\right), \tag{2.1.17}$$

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} \leq d \cdot \inf_{\theta<0} \mathrm{e}^{-\theta t} \cdot \bar{\operatorname{tr}} \exp\left(\sum\nolimits_k \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k}\right). \tag{2.1.18}$$

2.1.2.5 Matrix Concentration Inequalities via Lieb's Theorem

In this section, we summarize the major matrix concentration inequalities obtained based on Lieb's theorem. The first set of results contains the large deviation probabilities for a sum of independent self-adjoint random matrices under various assumptions. In Section 2.1.2.5.1, we display the deviation probabilities for matrix Gaussian and Rademacher Series. Section 2.1.2.5.2 contains the case where the random matrices are bounded and is a matrix version of Hoeffding's inequality. In addition to the boundedness assumptions of the random matrices, one can also use the matrix variance to deliver more refined concentration probabilities, such as the matrix Bernstein inequality [233, Theorem 6.1 and Theorem 6.2]. For a sum of positive-semidefinite random matrices, one can derive a matrix version of the Chernoff bound [233, Theorem 1.1]. These concentration results are direct matrix extensions of corresponding scalar concentration inequalities and they all depend on the master probability bounds of Theorem 2.1.5. The difference is the argument to control the individual matrix cumulant which depends on the varied assumptions of the matrix distribution. We refer to the original paper [233] for the complete proofs.

The second set of results [233, 231] relaxes the independence requirement for the random matrices in the sum and instead considers matrix martingales. Section 2.1.2.5.3, we first display the matrix Azuma inequality for an adapted sequence of random matrices. The matrix Azuma inequality leads to the matrix version of bounded differences inequality, which we display in Section 2.1.2.5.3 as well.

2.1.2.5.1 Matrix Gaussian and Rademacher Series We first present the following large deviation probability for matrix Gaussian and Rademacher Series [233, Theorem 1.2]. This result juxtaposes with the noncommutative Khintchine's inequality that bounds the matrix moments of Gaussian and Rademacher series.

Theorem 2.1.6. *Suppose $\boldsymbol{A}_k$ is a finite sequence of fixed self-adjoint matrices in $\mathbb{H}^d$. Let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Denote the variance parameter*

$$\sigma^2 := \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\| .$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \gamma_k \boldsymbol{A}_k \right) \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}. \tag{2.1.19}$$

The same probability bound holds if $\{\gamma_k\}$ is a series of independent Rademacher random variables.

Proof Sketch of Theorem 2.1.6. The main technical piece to establish the deviation bound (2.1.19) is to bound the individual matrix cumulant and apply the master probability bounds, Theorem 2.1.5. The good structure of the standard normal and the Rademacher distributions ensures the following cumulant bound:

$$\log \mathbb{E}\, \mathrm{e}^{\gamma_k \boldsymbol{A}_k} \preccurlyeq \theta^2 \boldsymbol{A}_k^2 / 2. \tag{2.1.20}$$

Substitute (2.1.20) into (2.1.17) and choose the optimal $\theta = t/\sigma^2$ to arrive at (2.1.19).

$$\begin{aligned}
\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \gamma_k \boldsymbol{A}_k \right) \right\} &\leq d \cdot \mathrm{e}^{\theta t} \cdot \bar{\operatorname{tr}} \exp\left(\frac{\theta^2}{2} \cdot \sum\nolimits_k \boldsymbol{A}_k^2 \right) \\
&\leq d \cdot \mathrm{e}^{-\theta t} \cdot \exp(\theta^2 \sigma^2 / 2) \\
&= d \cdot \mathrm{e}^{-t^2/2\sigma^2}.
\end{aligned}$$

□

2.1.2.5.2 Matrix Hoeffding Inequality We display a matrix Hoeffding inequality [233, Theorem 1.3] as follows, which clearly resembles the scalar Hoeffding's inequality that we demonstrate in Chapter 1.

Theorem 2.1.7 (Matrix Hoeffding Inequality)**.** *Suppose $\{\boldsymbol{Y}_k\}$ is a finite sequence of independent, random matrices in $\mathbb{H}^d$. Let $\{\boldsymbol{A}_k\} \subset \mathbb{H}^d$ be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies*

$$\mathbb{E}\, \boldsymbol{Y}_k = \boldsymbol{0} \quad \textit{and} \quad \boldsymbol{Y}_k^2 \preccurlyeq \boldsymbol{A}_k^2 \quad \textit{almost surely.} \tag{2.1.21}$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{Y}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2} \quad \textit{where} \quad \sigma^2 := \left\|\sum\nolimits_k \boldsymbol{A}_k^2\right\|.$$

2.1.2.5.3 Matrix Azuma and Bounded Differences Inequality The following matrix Azuma inequality [233, Theorem 7.1] extends the matrix Hoeffding's inequality, Theorem 2.1.7, to a sum of matrices that are not necessarily independent. Specifically, the matrix Azuma inequality considers an adapted sequence of random matrices $\{\boldsymbol{X}_k\}$. In this adapted sequence, each matrix $\boldsymbol{X}_k$ is measurable to the probability space $(\Omega, \mathcal{F}_k, \mathbb{P})$, which is a subspace of $(\Omega, \mathcal{F}, \mathbb{P})$. In addition, the sequence of signal algebras $\{\mathcal{F}_k\}$ forms a filtration such that

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_\infty \subset \mathcal{F},$$

where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the trivial sigma algebra. We follow the presentation of [233] and abbreviate the conditional expectation with respect to one signal algebra in the filtration as $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_k]$.

Theorem 2.1.8 (Matrix Azuma Inequality)**.** *Suppose $\{\boldsymbol{X}_k\}$ is a finite adapted sequence of self-adjoint matrices in $\mathbb{H}^d$ and $\{\boldsymbol{A}_k\}$ is a fixed sequence of self-adjoint matrices in $\mathbb{H}^d$. They satisfy*

$$\mathbb{E}_{k-1}\,\boldsymbol{X}_k = \boldsymbol{0} \quad \textit{and} \quad \boldsymbol{X}_k^2 \preccurlyeq \boldsymbol{A}_k^2 \quad \textit{almost surely.} \tag{2.1.22}$$

Denote the variance parameter

$$\sigma^2 := \left\|\sum\nolimits_k \boldsymbol{A}_k^2\right\|.$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2}.$$

The key to establishing the matrix Azuma inequality is to derive a new version of the subadditivity lemma, Lemma 2.1.4, by replacing the total expectation in the proof with the conditional expectations $\{\mathbb{E}_k\}$, which leads to a modified version of the master probability bounds, Theorem 2.1.5. The matrix Bernstein's inequality Theorem [233, Theorem 6.1] can also be extended to an adapted sequence in a similar fashion. The resulting concentration inequality is called matrix Freedman's inequality, first proposed by Oliveira [163] and later refined by Tropp [231].

An important application of the matrix Azuma inequality is the following matrix version of bounded differences inequality [233, Corollary 7.5], which controls the deviation probability of a general matrix function that depends on a sequence of independent random variables.

Theorem 2.1.9 (Matrix Bounded Differences Inequality). *Suppose $\{Z_k : k = 1, 2, ..., n\}$ is sequence of independent random variables. Let $\boldsymbol{H}$ be a function that maps $\{Z_k\}$ to a self-adjoint matrix of dimension d. Suppose $\{\boldsymbol{A}_k\}$ is a sequence of fixed self-adjoint matrices that satisfy the following boundedness condition*

$$(\boldsymbol{H}(z_1, \ldots, z_k, \ldots, z_n) - \boldsymbol{H}(z_1, \ldots, z_k', \ldots, z_n))^2 \preccurlyeq \boldsymbol{A}_k^2,$$

where z_i and z_i' range over all possible values of Z_i for all index i. Denote the variance parameter

$$\sigma^2 := \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\|.$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}(\boldsymbol{H}(\boldsymbol{z}) - \mathbb{E}\, \boldsymbol{H}(\boldsymbol{z})) \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2},$$

where $\boldsymbol{z} = (Z_1, \ldots, Z_N)$.

2.1.2.6 Extension

There are several directions of continuing work that follow the mains steps of this method based on Lieb's theorem, that is, the matrix Laplace bounds (Proposition 2.1.2) and decoupling using the subadditivity lemma (Lemma 2.1.4). The key is to modify the matrix Laplace bounds and we briefly summarize the main intuition in this section.

The first direction is to derive the large deviation probabilities of the interior spectrum in addition to those for the extreme eigenvalues of a sum of independent random matrices. In [80], Gittens and Tropp applied the Courant–Fisher theorem, which expresses the interior eigenvalues of a self-adjoint matrix as the maximum eigenvalue of the matrix projected into an appropriate subspace. In the first step, they modified the Laplace probability bounds and controlled the deviation probabilities of the interior eigenvalues of a random matrix $\boldsymbol{X} \in \mathbb{H}^d$ using the following variational bound [80, Theorem 3.1]:

$$\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \geq t\right\} \leq \inf_{\theta>0} \min_{\boldsymbol{V}\in\mathbb{V}^d_{d-k-1}} \left\{\mathrm{e}^{-\theta t} \cdot \mathbb{E}\operatorname{tr}\mathrm{e}^{\theta\boldsymbol{V}^*\boldsymbol{X}\boldsymbol{V}}\right\} \quad \text{for all } t \geq 0,$$

where $\mathbb{V}^d_k = \{\boldsymbol{V} \in \mathbb{C}^{d\times k} : \boldsymbol{V}\boldsymbol{V}^* = \mathbf{I}\}$ is the collection of orthonormal bases for the k-dimensional subspaces of $\mathbb{C}^d$.

In the second step, the authors proceed to establish a different subadditivity lemma that relies on an extension of Lieb's theorem. With these two modifications, the authors deliver Chernoff and Bernstein-type deviation inequalities for the interior eigenvalues. They apply these results to application problems such as column sampling and covariance matrix estimation.

Another direction is to derive more refined deviation probabilities for random matrices and capture the situation when the distribution concentrates in a lower-dimensional subspace. This line of work started with the paper of Hsu et al [103], and in a later work [158] Minsker modified the arguments and provided an improved version of the matrix Bernstein's inequality. Tropp streamlined the ideas and extended the proofs to other matrix concentration inequalities in the monograph [235]. Take the example of the matrix Bernstein's inequality. The idea of Minsker's approach is to use the spectrum structure of the variance parameter $\boldsymbol{\Sigma} = \sum_k \mathbb{E}\,\boldsymbol{X}_k^2$ that controls the deviation probabilities to improve the dimensional dependency. Instead of depending on the ambient dimension d of the random matrix, Minster's version of the matrix Bernstein's inequality depends on the following quantity, which is named by Tropp as the intrinsic dimension of the matrix variance [235, Definition 7.1.1]

$$\operatorname{intdim}(\boldsymbol{\Sigma}) = \frac{\operatorname{tr}\boldsymbol{\Sigma}}{\|\boldsymbol{\Sigma}\|}.$$

The intrinsic dimension can be much smaller than the ambient dimension when the trailing eigenvalues of the variance matrix $\boldsymbol{\Sigma}$ decays very rapidly. This improved dimensional dependency tightens the deviation probabilities for many cases of random matrices.

The key ingredient to the improved deviation probabilities is to generalize the

Laplace transform bounds of a self-adjoint matrix $\boldsymbol{X}$ to the following bound [235, Proposition 7.4.1]

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}) \geq t\right\} \leq \frac{1}{\psi(t)} \cdot \mathbb{E}\operatorname{tr}\psi(\boldsymbol{X}) \quad \text{for all } t \geq 0,$$

where the function $\psi : \mathbb{R} \mapsto \mathbb{R}_+$ is nonnegative and nondecreasing on the internal $[0, \infty)$. In the proof of the improved versions of matrix Chernoff or matrix Bernstein inequalities, the function ψ is set to be the exponential function, as in the original Laplace transform bounds, augmented with a linear function. So the later step of decoupling the matrix moment generating function proceeds with some technical modifications.

2.1.3 Method of Exchangeable Pairs

In this section, we summarize the framework of the method of exchangeable pairs and the associated concentration inequalities. First, in Section 2.1.3.1, we review the main concepts and definitions related to the method of exchangeable pairs and provide their intuition. In Section 2.1.3.2, we instantiate the definitions in Section 2.1.3.1 by constructing a matrix Stein pair from a sum of independent random matrices. This construction is the key connection to applications with a sum of independent random matrices. Next, in Section 2.1.3.3 we illustrate how to control the matrix moment generating function and Section 2.1.3.4 presents the main theorem that derives large deviation probabilities from properties of a matrix Stein pair. We illustrate the major concentration inequalities obtained with the method of exchangeable pairs in Section 2.1.3.5. Finally, we summarize the extension of the method of exchangeable pairs to more general applications in Section 2.1.3.6.

2.1.3.1 Matrix Stein Pairs

In this section, we review the definitions of exchangeable pairs, matrix Stein pairs, and the conditional variance as appeared in [138, Section 2]. These concepts are the backbones for the method of exchangeable pairs. We also provide the intuition of using the exchangeable method to derive concentration inequalities at the end.

Definition 2.1.10 (Exchangeable Pair)**.** *Let Z and Z' be random variables taking values* *in a Polish space $\mathcal{Z}$. We call (Z, Z') an exchangeable pair if it has the same distribution as (Z', Z).*

One constructs a matrix Stein pair based on the exchangeable pair (Z, Z') by applying a matrix-valued function. The exchangeable pair (Z, Z') captures the full scope of randomness that generates the Stein pair. We impose a regularity assumption that $\mathbb{E}\,\|\boldsymbol{X}\|^2 < \infty$.

Definition 2.1.11 (Matrix Stein Pair)**.** *Suppose (Z, Z) is an exchangeable pair of random variables taking values in a Polish space $\mathcal{Z}$. Let $\Phi : \mathcal{Z} \mapsto \mathbb{H}^d$ be a measurable function. Define the random Hermitian matrices*

$$\boldsymbol{X} := \Phi(Z) \quad \textit{and} \quad \boldsymbol{X}' := \Phi(Z').$$

We call $(\boldsymbol{X}, \boldsymbol{X}')$ a matrix Stein pair if it satisfies the linear reproducing property, that is, there exists a constant $\alpha \in (0, 1]$ for which

$$\mathbb{E}[\boldsymbol{X} - \boldsymbol{X}'|Z] = \alpha \boldsymbol{X} \quad \textit{almost surely.} \tag{2.1.23}$$

The constant α is called the scale factor of the pair.

As stated in [138], a matrix Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$ is also exchangeable. In addition, the linear reproducing property (2.1.23) implies that the matrix $\boldsymbol{X}$ is centered:

$$\mathbb{E}\,\boldsymbol{X} = \frac{1}{\alpha} \cdot \mathbb{E}\left[\mathbb{E}[\boldsymbol{X} - \boldsymbol{X}'|Z]\right] = \frac{1}{\alpha} \cdot (\mathbb{E}\,\boldsymbol{X} - \boldsymbol{X}') = \mathbf{0}.$$

The linear reproducing property implicitly captures the difference between $\boldsymbol{X}$ and $\boldsymbol{X}'$. The larger the scale factor α is, the more different the two matrices are. When $\boldsymbol{X}$ and $\boldsymbol{X}'$ are completely independent, we have $\alpha = 1$. Taking advantage of the structure of $\boldsymbol{X}$, we can construct the matrix $\boldsymbol{X}'$ that differs from $\boldsymbol{X}$ slightly, thus reducing the value of α.

For a matrix Stein pair, we associate a random matrix called the conditional variance. The conditional variance is a second-order characterization on the difference of the two matrices in the Stein pair.

Definition 2.1.12 (Conditional Variance)**.** *Suppose that $(\boldsymbol{X}, \boldsymbol{X}')$ is a matrix Stein pair, constructed from an exchangeable pair (Z, Z). The conditional variance is the random matrix*

$$\boldsymbol{\Delta}_{\boldsymbol{X}} := \boldsymbol{\Delta}_{\boldsymbol{X}}(Z) := \frac{1}{2\alpha}\,\mathbb{E}[(\boldsymbol{X} - \boldsymbol{X}')^2 | Z], \tag{2.1.24}$$

where α is the scale factor of the pair.

The conditional variance is a random perturbation of the variance, because the expectation of the conditional variance is equal to the matrix variance: $\mathbb{E}[\boldsymbol{\Delta}_{\boldsymbol{X}}] = \mathbb{E}\,\boldsymbol{X}^2$. It plays an important part in establishing concentration inequalities for random matrices. A uniform bound for the conditional variance also controls the matrix variance uniformly, which facilitates deriving concentration bounds such as the large deviation probabilities. The smaller the uniform bound for the conditional variance is, the better the resulting deviation probabilities will be. Thus, the goal of the method of exchangeable pairs is to explore the structure of the random matrix $\boldsymbol{X}$ and construct an appropriate matrix Stein pair. We want the Stein pair to have these properties: the corresponding conditional variance $\boldsymbol{\Delta}_{\boldsymbol{X}}$ is easy to calculate, and in addition the conditional variance is a small random perturbation of the matrix variance such that the uniform bound for $\boldsymbol{\Delta}_{\boldsymbol{X}}$ is as close to the variance of $\boldsymbol{X}$ as possible. The existence of a matrix Stein pair for an application scenario is a prerequisite for applying the method of exchangeable pairs to derive concentration inequalities.

2.1.3.2 Constructing Exchangeable Pairs from Sums of Independent Random Matrices

In this section, we demonstrate a Stein pair from a sum of independent self-adjoint random matrices. This construction is the key to applying the method of exchangeable pairs to obtain concentration inequalities for sums of independent random matrices, as we demonstrate in the proof of the Hoeffding's inequality in Section 2.1.3.5.1. This construction appears in [235, Section 2.4] and is the matrix extension of the scalar exchangeable pairs that Chatterjee [51] created for a scalar independent sum.

Suppose

$$\boldsymbol{X} = \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_n \quad \text{and} \quad Z = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n),$$

where $\{\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_n\}$ are independent self-adjoint random matrices of the same dimension and $\mathbb{E}\,\boldsymbol{Y}_i = \mathbf{0}$ for all index i. Suppose I is a random index drawn uniformly from $\{1, \ldots, n\}$. We construct Z' by replacing $\boldsymbol{Y}_I$ with an independent copy $\boldsymbol{Y}_I'$ such that

$$Z' := (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{I-1}, \boldsymbol{Y}_I', \boldsymbol{Y}_{I+1}, \ldots, \boldsymbol{Y}_n).$$

Correspondingly, we augment $\boldsymbol{X}$ with

$$\boldsymbol{X}' := \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_{I-1} + \boldsymbol{Y}_I' + \boldsymbol{Y}_{I+1} + \cdots + \boldsymbol{Y}_n$$

to arrive at the Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$. We can see that $\boldsymbol{X}'$ is a very local perturbation of $\boldsymbol{X}$ as $\boldsymbol{X}'$ is constructed by choosing uniformly one summand $\boldsymbol{Y}_I$ and swap it with an independent realization $\boldsymbol{Y}_I'$. The closeness between $\boldsymbol{X}$ and $\boldsymbol{X}'$ also manifests itself by the small value of the corresponding scale factor α:

$$\mathbb{E}[\boldsymbol{X} - \boldsymbol{X}'|Z] = \frac{1}{n}\boldsymbol{X}.$$

The conditional variance for the Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$ is

$$\boldsymbol{\Delta}_{\boldsymbol{X}} = \frac{1}{2}\sum\nolimits_{i=1}^{n}(\boldsymbol{Y}_i^2 + \mathbb{E}\,\boldsymbol{Y}_i^2).$$

2.1.3.3 Control Matrix Moment Generating Function with Method of Exchangeable Pairs

In this section, we demonstrate the connection between the matrix Stein pair and the matrix moment generating function with Lemma 2.1.14. The latter is the input to the matrix Laplace bound that controls large deviation probabilities for random matrices. The following technical lemma [138, Lemma 2.4] contains the key intuition of the exchangeable pair method.

Lemma 2.1.13. *Suppose that $(\boldsymbol{X}, \boldsymbol{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ is a matrix Stein pair with scale factor α. Let $\boldsymbol{F} : \mathbb{H}^d \mapsto \mathbb{H}^d$ be a measurable function that satisfies the regularity condition*

$$\mathbb{E}\left\|(\boldsymbol{X} - \boldsymbol{X}') \cdot \boldsymbol{F}(\boldsymbol{X})\right\| \leq \infty. \tag{2.1.25}$$

Then

$$\mathbb{E}[\boldsymbol{X} \cdot \boldsymbol{F}(\boldsymbol{X})] = \frac{1}{2\alpha}\,\mathbb{E}[(\boldsymbol{X} - \boldsymbol{X}')(\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{F}(\boldsymbol{X}'))]. \tag{2.1.26}$$

The goal of of Lemma 2.1.13 is to take advantage of the structure of the matrix $\boldsymbol{X}$ to provide an alternative approach of bounding quantities that can be represented in the form of $\mathbb{E}[\boldsymbol{X} \cdot \boldsymbol{F}(\boldsymbol{X})]$. As Lemma 2.1.14 shows, quantities of such format include the derivative of the moment generating function. When the structure of the random matrix $\boldsymbol{X}$ leads to a good matrix Stein pair such that $\boldsymbol{X}$ and $\boldsymbol{X}'$ only deviates locally from each other, the hope is that $\boldsymbol{F}(\boldsymbol{X})$ is very similar to $\boldsymbol{F}(\boldsymbol{X}')$ as well and as a result the right-hand side of (2.1.26) is relatively easier to control. In many applications, we can relate the right-hand side of (2.1.26) with the conditional variance $\boldsymbol{\Delta}_{\boldsymbol{X}}$ of the matrix Stein pair, which as we mentioned earlier, is a random perturbation of the matrix variance.

Lemma 2.1.14 substantiates this intuition and controls the derivative of the matrix moment generating function with the conditional variance. The bounds in Lemma 2.1.14 lead to Theorem 2.1.15, which produces matrix large deviation probabilities based on the property of the conditional variance. Finally, we show that Theorem 2.1.15 applies to the matrix Stein pair in Section 2.1.3.2 and produces an improved version of matrix Hoeffding's inequality as in Theorem 2.1.16.

Lemma 2.1.14. *Suppose that $(\boldsymbol{X}, \boldsymbol{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ is a matrix Stein pair, and assume that $\boldsymbol{X}$ is almost surely bounded in norm. Recall the trace moment generating function $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\mathrm{e}^{\theta\boldsymbol{X}}$. Then*

$$m'(\theta) \le \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\boldsymbol{\Delta}_{\boldsymbol{X}} \cdot \mathrm{e}^{\theta\boldsymbol{X}}\big] \quad \textit{when } \theta \ge 0 \tag{2.1.27}$$

$$m'(\theta) \ge \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\boldsymbol{\Delta}_{\boldsymbol{X}} \cdot \mathrm{e}^{\theta\boldsymbol{X}}\big] \quad \textit{when } \theta \le 0, \tag{2.1.28}$$

where $\boldsymbol{\Delta}_{\boldsymbol{X}}$ is the conditional variance.

Once we obtain the uniform upper bounds for the conditional variance $\boldsymbol{\Delta}_{\boldsymbol{X}}$, we control the derivative of the matrix moment generating function with Lemma 2.1.14, which in turn bounds the matrix moment generating function itself.

2.1.3.4 Probability Bounds for Matrix Stein Pairs

We present the following theorem that produces large deviation probabilities for a random matrix based on the boundedness property of the conditional variance of the matrix Stein pair. This theorem appears in [138, Theorem 4.1] and is the matrix extension of Chatterjee's deviation probability bound for scalar random variables [50, 51].

Theorem 2.1.15. *Consider a matrix Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$ of $d \times d$ self-adjoint matrices. Suppose there exist nonnegative constants c, v for which the conditional variance of the pair satisfies*

$$\boldsymbol{\Delta}_{\boldsymbol{X}} \preccurlyeq c \cdot \boldsymbol{X} + v \cdot \mathbf{I} \quad \textit{almost surely.} \tag{2.1.29}$$

Then for all $t \geq 0$,

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\min}(\boldsymbol{X}) \leq -t\right\} &\leq d \cdot \exp\left(-\frac{t^2}{2v}\right), \\ \mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}) \geq t\right\} &\leq d \cdot \exp\left(\frac{-t^2}{2v + 2ct}\right). \end{aligned}$$

Theorem 2.1.15 derives from Lemma 2.1.14 and the proof consists of three steps. First, combine the bounds in Lemma 2.1.14 with the assumption on the conditional variance (2.1.29) to arrive at two differential inequalities of the moment generating function. Second, integrate the differential inequalities which bound the moment generating function. Finally, substitute into the matrix Laplace bound to arrive at the deviation probabilities.

2.1.3.5 Matrix Concentration Inequalities via the Method of Exchangeable Pairs

In this section, we summarize the major concentration inequalities obtained by the method of exchangeable pairs. Based on the matrix Stein pair in Section 2.1.3.2, the authors of [138] obtained an improved version of matrix Hoeffding's inequality (Section 2.1.3.5.1), a version of matrix Bernstein's inequality [138, Corollary 5.2], and non-commutative Khintchine's inequality (Section 2.1.17).

Inspired by the work of Chatterjee [50], the authors of [138] also construct a appropriate matrix Stein pair for a random combinatorial sum of deterministic matrices and establish a Bernstein-type deviation inequality for the distribution of the combinatorial sum. As an application, they obtain deviation probabilities for a sum of matrices sampled without replacement from a set of deterministic matrices. The authors also produce a version of matrix bounded difference inequality for random matrices that satisfy a self-producing property.

2.1.3.5.1 Matrix Hoeffding's Inequality The method of exchangeable pairs produces the following Hoeffding's inequality [138, Corollary 4.2], which improves over that of Theorem 2.1.7 with a smaller constant and a more refined version of the variance parameter σ^2.

Theorem 2.1.16 (Matrix Hoeffding Inequality)**.** *Under the same assumptions of Theorem 2.1.7, we have*

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{Y}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}, \quad \textit{for all } t \geq 0,$$

where the variance parameter is

$$\sigma^2 := \frac{1}{2} \cdot \left\|\sum\nolimits_k (\boldsymbol{A}_k^2 + \mathbb{E}\,\boldsymbol{Y}_k^2)\right\|.$$

The following proof concludes a complete picture of deriving concentration inequalities with the method of exchangeable pairs.

Proof. We use the construction of the matrix Stein pair for a sum of random matrices in Section 2.1.3.2. The conditional variance satisfies

$$\boldsymbol{\Delta}_{\boldsymbol{X}} = \frac{1}{2}\sum\nolimits_k (\boldsymbol{Y}_k^2 + \mathbb{E}\,\boldsymbol{Y}_k^2) \preccurlyeq \sigma^2 \mathbf{I},$$

where the last relation is due to the assumption that $\boldsymbol{Y}_k^2 \preccurlyeq \boldsymbol{A}_k^2$. Finally, we choose $c = 0$, $v = \sigma^2$ and apply Theorem 2.1.15 to arrive at the Hoeffding's bound. □

2.1.3.5.2 Non-Commutative Matrix Moment Inequalities The method of exchangeable pairs also leads to non-commutative moment inequalities for random matrices [138, Section 7]. Specifically, Theorem 2.1.17 restates [138, Theorem 7.1] and is the matrix Burkholder–Davis–Gundy (BDG) inequality that controls the matrix moments with the moments of the conditional variance of the Stein pair.

Theorem 2.1.17 (Matrix BDG Inequality)**.** *Let $p = 1$ or $p \geq 1.5$. Suppose that $(\boldsymbol{X}, \boldsymbol{X}')$ is a matrix Stein pair where $\mathbb{E}\,\|\boldsymbol{X}\|_{2p}^{2p} < \infty$. Then*

$$\big(\mathbb{E}\,\|\boldsymbol{X}\|_{2p}^{2p}\big)^{1/(2p)} \leq \sqrt{2p-1} \cdot \big(\mathbb{E}\,\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{p}^{p}\big)^{1/(2p)},$$

where $\|\boldsymbol{X}\|_p = (\operatorname{tr}|X|^p)^{1/p}$ is the matrix Shatten norm.

The following theorem [138, Corollary 7.3] is a corollary of Theorem 2.1.17 and contains a version of non-commutative Khintchine's inequality. It is a result of applying Theorem 2.1.17 to the matrix Stein pair for a sum of independent random matrices presented in Section 2.1.3.2.

Theorem 2.1.18 (Non-commutative Khintchine's Inequality)**.** *Suppose that $p = 1$ or $p \geq 1.5$. Suppose the independent Hermitian matrices $(\boldsymbol{Y}_k)_{k\geq 1}$ and the deterministic sequence $(\boldsymbol{A}_k)_{k\geq 1}$ satisfy the same assumptions as those of matrix Hoeffding's inequality, that is*

$$\mathbb{E}\,\boldsymbol{Y}_k = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{Y}_k^2 \preccurlyeq \boldsymbol{A}_k^2 \quad \text{almost surely for each index } k.$$

Then

$$\left(\mathbb{E}\left\|\sum\nolimits_k \boldsymbol{Y}_k\right\|_{2p}^{2p}\right)^{1/(2p)} \leq \sqrt{p-0.5} \cdot \left\|\left(\sum\nolimits_k (\boldsymbol{A}_k^2 + \mathbb{E}\,\boldsymbol{Y}_k^2)\right)^{1/2}\right\|_{2p}.$$

In particular, when $(\epsilon_k)_{k\geq 1}$ is an independent sequence of Rademacher random variables,

$$\left(\mathbb{E}\left\|\sum\nolimits_k \epsilon_k \boldsymbol{A}_k\right\|_{2p}^{2p}\right)^{1/(2p)} \leq \sqrt{2p-1} \cdot \left\|\left(\sum\nolimits_k \boldsymbol{A}_k^2\right)^{1/2}\right\|_{2p}.$$

2.1.3.6 Extension

The authors of [173] extend the method of exchangeable pairs in order to develop concentration results for general matrix functions. The main idea is to extend the concept of the matrix Stein pair in Definition 2.1.11 to a more generalized definition of Kernel Stein pair. Following the notational setups of Definition 2.1.11, the definition for the Kernel Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$ modifies the linear reproducing property (2.1.23) with a kernel producing property [173, Definition 7.2]:

$$\mathbb{E}[\boldsymbol{K}(Z, Z')|Z] = \boldsymbol{X},$$

where (Z, Z') is the exchangeable pair that generates the Kernel Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$ and $\boldsymbol{K}$ is a anti-symmetric matrix kernel function such that $\boldsymbol{K}(z, z') = -\boldsymbol{K}(z', z)$. In additional to the conditional variance (Definition 2.1.12), the authors augmented the variance characterization with a new concept of kernel conditional variance [173, Definition 8.5],

$$\boldsymbol{V}^{\boldsymbol{K}} = \frac{1}{2}\,\mathbb{E}[\boldsymbol{K}(Z, Z')^2|Z],$$

which together with the conditional variance bounds the matrix variance of $\boldsymbol{X}$ in the positive-semidefinite order. Other key lemmas such as Lemma 2.1.13 are extended by taking into account the kernel reproducing property and the main arguments to establish concentration results follow similar steps as we described in previous sections.

The Kernel Stein pair accommodates a broader set of applications for which a matrix Stein pair with a linear reproducing property might not exist. An important approach of constructing a kernel matrix from a random matrix is via Markov chain coupling, which as explained by the authors of [173] is originally developed for the scalar exchangeable methods of Chatterjee [51]. The coupling speed of the implicit Markov chain corresponding to the Kernel Stein pair implies the size of the kernel conditional variance. Using the Markov chain coupling method, the authors derived multiple Efron–Stein type concentration inequalities that control the matrix moments. In addition, they obtain a more generalized version of bounded difference inequality for dependent random matrices whose dependency property is characterized by a Dobrushin interdependence matrix.

2.2 Overview of Main Results of the Thesis

In this section, we summarize the main results of the thesis. In Section 2.2.1, we provide the context and the contributions of our work in the analysis of the masked covariance estimator [53]. In Section 2.2.2, we exhibit the connection between the random matrix entropy that we propose with the existing entropy concepts, illustrate the main contributions of our work as published in [54], and discuss implications and recent development.

2.2.1 Analysis of the Masked Sample Covariance Estimator via the Matrix Laplace Transform Method

In Chapter 3, we apply the matrix Laplace transform method to study the masked sample covariance estimator. In Section 2.2.1.1, we briefly review the important problem of covariance matrix estimation. Then in Section 2.2.1.2, we introduce the masked sample covariance estimator, provide its intuition, and connect with a conjecture by Levina and Vershynin that was the starting point of our work. Finally, we summarize our contributions in Section 2.2.1.3.

2.2.1.1 Covariance Matrix Estimation

Covariance matrix estimation from independent samples of a distribution is a fundamental problem. It arises in theoretical statistical problems such as regression analysis [76] and principal component analysis [109]. An accurate covariance matrix is also key to an array of applications, such as the study of genetic correlation in biostatistics [94] and the capital asset pricing model [70] in modern portfolio management theory.

In the setting of classical statistics, the number of samples exceeds the number of variables and the sample covariance matrix is the standard estimator [108, 159, 145]. In the context of high dimensional statistics, the dimension of the covariance matrix can be much larger than the number of samples and it is necessary to leverage additional assumptions to estimate the covariance structure. As summarized by Rothman et. al. [192], the study of high-dimensional covariance matrix estimation can be divided into two categories. The first category includes works by Bickel & Levina [19, 20], Furer & Bengtsson [77] and the random variables exhibit a natural ordering or distance. Examples include time series data, spatial data, etc. In the second category, the random variables do not have a natural ordering. Examples include data that exhibit a graph structure. The methods developed in this category are invariant to variable permutation and a common approach is to apply sparsity regularization. See El Karoui [117], Bickel & Levina [21].

2.2.1.2 Masked Sample Covariance Estimator

In Chapter 3, we restrict our attention to the estimation of a high-dimensional covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ where a natural ordering among the random variables does not exist. We assume a deterministic mask matrix $\boldsymbol{M}$ that contains sparsity information is available. Based on the independent samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and the mask matrix $\boldsymbol{X}$, the masked sample covariance estimator is

$$\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \cdot \sum\nolimits_{k=1}^{n} \boldsymbol{M} \odot (\boldsymbol{x}_k \boldsymbol{x}_k^*). \tag{2.2.1}$$

We point out that the masked sample covariance estimator is a sum of independent random matrices. We study the required number of samples $n(\epsilon)$ that ensures the the discrepancy as measured in matrix spectral norm between the masked sample covariance and the masked covariance estimator

$$\left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\|$$

is less than ϵ with high probability. The symbol $\odot$ denotes the componentwise Hadamard product. The mask matrix $\boldsymbol{M}$ acts as a filter and we can reduce the influence of entries that either we cannot estimate reliably or we do not want to estimate due to certain a priori assumptions by setting the corresponding entries in $\boldsymbol{M}$ to zero. Intuitively, the required number of samples n increases with the problem's dimension p. We want to study the dependency of n on p and whether it is feasible to estimate the masked covariance matrix $\boldsymbol{M} \odot \boldsymbol{\Sigma}$ accurately when $n \ll p$.

The masked sample covariance estimator was introduced by Levina and Vershynin [128]. They demonstrated that when the distribution is Gaussian, it is feasible to estimate the masked covariance matrix in the $n \ll p$ case because the required number of samples grows logarithmically as the dimension p increases. Levina and Vershynin hypothesized that such results can extend to more general distributions with appropriate moment growth. The challenge is that their argument relies on successfully decoupling a second-order Gaussian chaos by applying the rotational invariance property that is unique to the Gaussian distributions.

The goal of Chapter 3 is to validate the hypothesis of Levina and Vershynin. We develop a completely different approach based on the Laplace transform method. The probability tail bounds of Theorem 2.1.5 applies directly to the masked covariance estimator as it is a sum of independent random matrices. Our rationale for this new approach is that bounding the moment generating function is a more general problem than decoupling a second-order chaos, such that we do not need to rely on the individual structure of the distribution and we can incorporate the moment growth properties more easily into the argument. Indeed, the main challenge in our proof is to derive a tight upper bound of the moment generating function of the individual matrix $\boldsymbol{M} \odot \boldsymbol{x}_k \boldsymbol{x}_k^*$ in the masked sample covariance matrix (2.2.1). Our argument takes a truncation approach and the key is a detailed bound for the second-order moments of the summand.

2.2.1.3 Contributions

Our first contribution is the following theorem, where we improve upon Levina and Vershynin's result in the Gaussian setting by obtaining a tighter upper bound of the estimation error. Our theorem also implies a tighter lower bound of the required number of samples for a specified level of estimation accuracy.

Theorem 2.2.1 (Masked Covariance Estimation for Gaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a Gaussian random vector in $\mathbb{R}^p$ with mean zero. The expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq 8\left[\left(\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \cdot \log(6p)}{n}\right)^{1/2} + \frac{\|\boldsymbol{M}\| \cdot \log^2(6np)}{n}\right]\|\boldsymbol{\Sigma}\|.$$

We argue that our result is near optimal. Second, we validate Levina and Vershynin's hypothesis by showing that estimating the masked sample covariance matrix in the $n \ll p$ setting is feasible for the more general class of subgaussian distributions, as summarized in the following theorem

Theorem 2.2.2 (Masked Covariance Estimation for Subgaussian Distribution)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a subgaussian random vector in $\mathbb{R}^p$ with mean* *zero. Then the expected estimation error satisfies:*

$$\mathbb{E}\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\leq\left[\frac{16\kappa^2\nu^2\left\|\boldsymbol{M}\right\|_{1\to 2}^2\log(2\mathrm{e}p)}{n}\right]^{1/2}+\frac{4\kappa^2\left\|\boldsymbol{M}\right\|\log^2(2\mathrm{e}np)}{n},$$

where κ and ν are constants depend on the distribution of $\boldsymbol{x}$.

Finally, compared with the previous covering argument of Vershynin and Levina [128], our new approach via the matrix Laplace transform method is more transparent and involves fewer technical complexities.

Our analysis of the masked sample covariance estimator is based on the assumption that a mask matrix $\boldsymbol{M}$ exists. This assumption decouples the problem of estimating a sparse covariance matrix into two parts. The first part is to obtain mask matrix that is consistent with the sparsity structure of the covariance matrix, such that the masked covariance matrix captures the major content of the original covariance matrix. Our work focuses on the second part of measuring the deviation of the masked sample covariance matrix from the masked covariance matrix. One can obtain a reliable mask matrix when the covariance matrix exhibits structures that translate directly into an informative mask matrix. For example, in the case of time-series data, the corresponding covariance matrix has significant values along the diagonal while entries far from the diagonal tend to have insignificant magnitude. In a general situation, estimating a good mask matrix can be difficult and our work does not provide guidance on overcoming this challenge.

2.2.2 Subadditivity of Matrix φ-Entropy and Matrix Concentration Inequalities

In Chapter 4, our goal is to construct an approach for random matrices that is similar to the scalar entropy method. We define entropy functionals on finite-dimensional random matrices, establish the subadditivity properties of entropy functionals, and prove several matrix concentration inequalities.

In Section 2.2.2.1, we first go over our thought process of defining the matrix entropy. Then in Section 2.2.2.2, we illustrate the connections between the matrix entropy and matrix concentration results, and explain how to obtain matrix concentration inequalities via the subadditivity property of the matrix entropy. Then in Section 2.2.2.3, we exhibit the more general concept of matrix φ-entropy and display our matrix concentration results in Section 2.2.2.4. We also show a generalized subadditivity property of the matrix entropy in the $*$-algebra setting in Section 2.2.2.5. Finally, we provide a thorough discussion of our work in Section 2.2.2.6.

2.2.2.1 Matrix Entropy

The first step is to construct an entropy for random matrices that captures the randomness of the matrix entries. The von Neumann quantum entropy provides guidance for us to define entropies for random matrices. The quantum entropy is defined on density matrices that characterize quantum systems. A density matrix $\rho \in \mathbb{H}^d$ is deterministic, positive semidefinite, and satisfies the unit trace condition $\operatorname{tr} \rho = 1$. The positive-semidefiniteness of the density matrix and the unit trace condition impliy that the eigenvalues of a density matrix are non-negative and sum up to 1, which forms a probability distribution. The eigenvectors specifies the quantum states that a system can have and the corresponding eigenvalues are the probability that the system is in each state. Thus, a rank-1 density matrix corresponds to a pure-state system and higher-rank density matrices characterize mixed-state systems. The von Neumann quantum entropy $S(\rho)$ measures the uncertainty of a quantum system as specified by the density matrix, and is defined as

$$S(\rho) := -\operatorname{tr}(\rho \log \rho).$$

Denote the eigenvalues of ρ as $\{\lambda_k\}_{k=1}^n$. Then the von Neumann entropy can be equivalently represented as

$$S(\rho) = -\sum\nolimits_{k=1}^n \lambda_k \log \lambda_k,$$

which is the Shannon entropy on the discrete probability distribution of the eigenvalues.

In practice, we find the following trace-normalized negative entropy, which is a convex function on the density matrix ρ, to be more convenient:

$$\hat{S}(\rho) = \bar{\operatorname{tr}}(\rho \log \rho) = \frac{1}{d} \cdot \operatorname{tr}(\rho \log \rho).$$

One can generalize the negative von Neumann entropy to any deterministic positive-semidefinite matrix $\boldsymbol{P} \in \mathbb{H}^d$ with a normalization by $\bar{\operatorname{tr}}(\boldsymbol{P})$:

$$\hat{S}(\boldsymbol{P}) := \bar{\operatorname{tr}}(\boldsymbol{P} \log \boldsymbol{P}) - \bar{\operatorname{tr}}(\boldsymbol{P}) \cdot \log \bar{\operatorname{tr}}(\boldsymbol{P}). \tag{2.2.2}$$

We interprete (2.2.2) as the entropy encoded by the eigen structure of a deterministic matrix. We also note that the normalized trace behaves like an expectation over the eigenvalues and the function $x \mapsto x \log x$ is convex. So by Jensen's inequality,

$$\bar{\operatorname{tr}}(\boldsymbol{P} \log \boldsymbol{P}) \geq \bar{\operatorname{tr}}(\boldsymbol{P}) \log \bar{\operatorname{tr}}(\boldsymbol{P}) \tag{2.2.3}$$

and $\hat{S}(\boldsymbol{P})$ equals the gap between the left-hand side and right-hand side of (2.2.3).

When we define entropy for a positive-semidefinite random matrix $\boldsymbol{Z}$, we want to capture the randomness due to the distribution of the matrix entries only, instead of the matrix structure. We start with the logarithmic function and the following definition becomes our natural choice

$$H(\boldsymbol{Z}) := \mathbb{E} \, \bar{\operatorname{tr}}(\boldsymbol{Z} \log \boldsymbol{Z}) - \bar{\operatorname{tr}}(\mathbb{E} \, \boldsymbol{Z} \log \mathbb{E} \, \boldsymbol{Z}). \tag{2.2.4}$$

The function $\boldsymbol{X} \mapsto \bar{\operatorname{tr}}(\boldsymbol{X} \log \boldsymbol{X})$ is convex. So

$$\mathbb{E} \, \bar{\operatorname{tr}}(\boldsymbol{Z} \log \boldsymbol{Z}) \geq \bar{\operatorname{tr}}(\mathbb{E} \, \boldsymbol{Z} \log \mathbb{E} \, \boldsymbol{Z})$$

and $H(\boldsymbol{Z})$ as defined in (2.2.4) is non-negative and measures the function gap due to the randomness of the matrix entries only.

Similar to the scalar entropy, we can also define a conditional matrix entropy for a product probability distribution. Suppose $\boldsymbol{Z}$ is a function of $\boldsymbol{x} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ where the X_i's are independent. We denote $\boldsymbol{x}_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ and abbreviate $\mathbb{E}_i = \mathbb{E}(\cdot|\boldsymbol{x}_{-i})$. Then the conditional matrix entropy is

$$H(\boldsymbol{Z}|\boldsymbol{x}_{-i}) := \mathbb{E}_i \,\bar{\operatorname{tr}}(\boldsymbol{Z} \log \boldsymbol{Z}) - \bar{\operatorname{tr}}(\mathbb{E}_i \,\boldsymbol{Z} \log \mathbb{E}_i \,\boldsymbol{Z}). \tag{2.2.5}$$

2.2.2.2 Subadditivity Property and Matrix Concentration

In order to derive matrix concentration inequalities for a self-adjoint matrix $\boldsymbol{Y} \in \mathbb{H}^d$ from our matrix entropy (2.2.4), we need several ingredients. The argument is very similar to the steps of establishing the subadditivity property of the scalar entropy. Section 2.2.2.2.1 provides a matrix Laplace bound based on the matrix cumulant and exhibits a matrix Herbst argument. We compare them with their scalar counterparts which appear in Section 1.2.3.1. Then we exhibit a supremum representation of the matrix entropy and the resulting subadditivity property in Section 2.2.2.2.2. They are matrix generalizations of the scalar results that appear in Section 1.2.3.3.

2.2.2.2.1 Matrix Laplace Bound and Matrix Herbst Argument The first step is the matrix Laplace transform method, which connects the matrix large deviation probability with the matrix trace cumulant $\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta \boldsymbol{Y}}$ with the following bound:

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \geq t\right\} \leq \inf_{\theta > 0} d \cdot \exp\left(-\theta t + \log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta \boldsymbol{Y}}\right). \tag{2.2.6}$$

Second, we use the matrix version of the Herbst argument to bound the matrix trace cumulant with an integral function of the matrix entropy.

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta \boldsymbol{Y}} = \theta \cdot \int_0^\theta \frac{H(\mathrm{e}^{\beta \boldsymbol{Y}})}{\mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\beta \boldsymbol{Y}}} \cdot \frac{\mathrm{d}\beta}{\beta^2}. \tag{2.2.7}$$

Note the resemblance between (2.2.7) and the orignal Herbst argument (1.2.17) in the scalar entropy method.

2.2.2.2.2 Supremum Representation of the Matrix Entropy and the Subadditivity Property The next step is to derive bounds for the entropy $H(\mathrm{e}^{\beta \boldsymbol{Y}})$ in (2.2.7). As in the scalar entropy method, one approach is to develop matrix versions of logarithmic-Sobolev inequality for certain distributions to control the entropy directly. Alternatively, we can develop the subadditivity property of the matrix entropy, which has led to a modified logarithmic-Sobolev inequality in the scalar case. In our work, we take the second approach.

The key step of establishing the subadditivity property of the matrix φ-entropy is the following supremum characterization that resembles the supremum representation of the scalar entropy (1.2.20):

$$H(\boldsymbol{Z}) = \sup_{\boldsymbol{T}\succcurlyeq\boldsymbol{0}} \mathbb{E}\,\bar{\mathrm{tr}}\big[(\log(\boldsymbol{T}) - \log(\mathbb{E}\,\boldsymbol{T}))(\boldsymbol{Z}-\boldsymbol{T}) + H(\boldsymbol{T})\big]. \tag{2.2.8}$$

As in the scalar case, the supremum representation of the matrix entropy is also due to the convexity of the matrix entropy. The proof of (2.2.8) relies on some convexity arguments in the operator theory and follow similar steps in the proof of the subadditivity property of the scalar φ-entropies in Boucheron et al. [26].

The subadditivity property of the matrix entropy has a similar structure to the scalar version. We show that the matrix entropy functional as defined in (2.2.4) exhibits the subadditivity property

$$H(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^{n} \mathbb{E}\left[H(\boldsymbol{Z}|\boldsymbol{x}_{-i})\right], \tag{2.2.9}$$

where $H(\boldsymbol{Z}|\boldsymbol{x}_{-i})$ is the conditional matrix entropy (2.2.5).

2.2.2.2.3 Infimum Representation of the Matrix Entropy and Matrix Concentration Inequalities The derivation of a modified logarithmic-Sobolev inequality depends on the following infimum representation of the matrix entropy:

$$H(\boldsymbol{Z}) = \inf_{\boldsymbol{T}\succcurlyeq\boldsymbol{0}} \mathbb{E}\,\bar{\mathrm{tr}}[\boldsymbol{Z}(\log\boldsymbol{Z} - \log\boldsymbol{T}) - (\boldsymbol{Z}-\boldsymbol{T})], \tag{2.2.10}$$

which is clearly the matrix counterpart of the infimum representation (1.2.21) of the scalar entropy. Apply (2.2.10) to the conditional matrix entropies on the right-hand side of (2.2.9) to obtain the modified logarithimic-Sobolev inequality:

$$H(\boldsymbol{Z}) \le \frac{1}{2}\sum\nolimits_{k=1}^{n} \mathbb{E}\,\bar{\mathrm{tr}}\left[(\boldsymbol{Z}-\boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i')\right], \tag{2.2.11}$$

where $\psi(x) = 1 + \log x$ and

$$\boldsymbol{Z}_i' = \boldsymbol{Z}(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n),$$

with X_i' an independent copy of X_i. We see that (2.2.11) is the matrix version of the scalar modified logarithmic-Sobolev inequality (1.2.26).

Based on (2.2.11), we derive a large deviation probability for matrix functions that are invariant under signed permutation and have bounded difference. We exhibit this result in Section 2.2.2.4. We summarize the main steps to derive matrix concentration inequalities using the matrix entropy. They are very similar to main steps of the scalar entropy method. First, set $\boldsymbol{Z} = \mathrm{e}^{\theta \boldsymbol{Y}}$ and control the sum on the right-hand side of the modified logarithmic-Sobolev inequality (2.2.11). Second, substitute the entropy bound into the matrix Herbst inequality (2.2.7) to arrive at an upper bound on the matrix trace cumulant. Finally, substitute into the matrix Laplace bound (2.2.6), and minimize over the parameter θ to arrive at the matrix deviation probability.

2.2.2.3 Matrix φ-Entropy

As in the scalar entropy method, we want to go beyond the logarithmic entropy and explore all convex functions with which we can define an entropy functional that exhibits the subadditivity property. The matrix φ-entropy functional we define for a positive-semidefinite random matrix $\boldsymbol{Z}$ is

$$H_\varphi(\boldsymbol{Z}) := \mathbb{E}\,\bar{\operatorname{tr}}\,\varphi(\boldsymbol{Z}) - \bar{\operatorname{tr}}\,\varphi(\mathbb{E}\,\boldsymbol{Z}),$$

and the corresponding conditional matrix φ-entropy for a product probability distribution is

$$H_\varphi(\boldsymbol{Z}|\boldsymbol{x}_{-i}) := \mathbb{E}_i\,\bar{\operatorname{tr}}\varphi(\boldsymbol{Z}) - \bar{\operatorname{tr}}\varphi(\mathbb{E}_i\,\boldsymbol{Z}).$$

The subadditivity property of the matrix φ-entropy is

$$H_\varphi(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^n \mathbb{E}\left[H_\varphi(\boldsymbol{Z}|\boldsymbol{x}_{-i})\right]. \tag{2.2.12}$$

In order for the subadditivity property to hold for the matrix φ-entropy, the function $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}$ needs to satisfy certain conditions. In Chapter 4, we delineate the sufficient conditions that the function φ should satisfy to ensure the subadditivity of the corresponding matrix φ-entropy. We call this group of functions the Φ function class. And we verify these conditions for the functions $\varphi : x \mapsto x \log x$ and $\varphi : x \mapsto x^p$ where $p \in [1, 2]$, the first of which generates the matrix entropy (2.2.4).

Based on applying the subadditivity property (2.2.12) of the matrix φ-entropy to the power function $\varphi : x \mapsto x^p$ with $p \in [1, 2]$, we obtain a matrix moment bound for matrix functions that are invariant under signed permutation and have bounded difference. We exhibit this result in Section 2.2.2.4.

2.2.2.4 Concentration Inequalities from the Subadditivity Properties of Matrix Entropy

With the subadditivity property of the matrix φ-entropy, we obtain several matrix concentration inequalities for random matrices that are invariant under sign permutations. We have explained the main steps of deriving probabilistic concentration results in Section section:supreumum-subadditivity for the φ-entropy with the choice of $\varphi(x) = x \log x$. The result is the following version of the bounded difference inequalities.

Theorem 2.2.3 (Bounded Differences)**.** *Let $\boldsymbol{x} := (X_1, \dots, X_n)$ be a vector of independent random variables, and let $\boldsymbol{x}' := (X_1', \dots, X_n')$ be an independent copy of $\boldsymbol{x}$. Consider the following self-adjoint random matrices in $\mathbb{R}^d$*

$$\begin{aligned} \boldsymbol{Y} &:= \boldsymbol{Y}(X_1, \dots, X_i, \dots, X_n) \quad \textit{and} \\ \boldsymbol{Y}_i' &:= \boldsymbol{Y}(X_1, \dots, X_i', \dots, X_n) \quad \textit{for } i = 1, \dots, n. \end{aligned}$$

Assume that $\boldsymbol{Y}$ is invariant under signed permutation and that $\|\boldsymbol{Y}\|$ is bounded almost surely. Define the variance measure

$$V_{\boldsymbol{Y}} := \sup \left\| \mathbb{E} \left[\sum\nolimits_{i=1}^{n} (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \middle| \boldsymbol{x} \right] \right\|,$$

where the supremum occurs over all possible values of $\boldsymbol{x}$*. For each* $t \geq 0$*,*

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}, \quad \textit{and}$$
$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \leq -t\right\} \leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}.$$

Similar to the case of the scalar entropy method (Section 1.2.3.4), the power function $\varphi(x) = x^p$ for $p \in [1, 2]$ leads to matrix moment bounds. We exhibit our matrix moment bound in the following theorem. Again, we require that the distribution of the random matrices are invariant under sign permutations.

Theorem 2.2.4 (Matrix Moment Bound)**.** *Fix a number* $q \in \{2, 3, 4, \dots\}$*. Let* $\boldsymbol{x} := (X_1, \dots, X_n)$ *be a vector of independent random variables, and let* $\boldsymbol{x}' := (X_1', \dots, X_n')$ *be an independent copy of* $\boldsymbol{x}$*. Consider the following self-adjoint positive-semidefinite random matrices in* $\mathbb{R}^d$

$$\boldsymbol{Y} := \boldsymbol{Y}(X_1, \dots, X_i, \dots, X_n) \quad \textit{and}$$
$$\boldsymbol{Y}_i' := \boldsymbol{Y}(X_1, \dots, X_i', \dots, X_n) \quad \textit{for } i = 1, \dots, n.$$

Assume that $\boldsymbol{Y}$ *is invariance under signed permutation and* $\mathbb{E}(\|\boldsymbol{Y}\|^q) < \infty$*. Suppose that there is a constant* $c \geq 0$ *with the property*

$$\boldsymbol{V}_{\boldsymbol{Y}} := \mathbb{E}\left[\sum\nolimits_{i=1}^{n} (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \middle| \boldsymbol{x}\right] \preccurlyeq c\boldsymbol{Y}.$$

Then the random matrix $\boldsymbol{Y}$ *satisfies the moment inequality*

$$\left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q)\right]^{1/q} \leq \mathbb{E}\,\bar{\mathrm{tr}}\boldsymbol{Y} + \frac{q-1}{2} \cdot c. \tag{2.2.13}$$

Compare our matrix moment inequality (2.2.13) with Boucheron's result (1.2.34).

2.2.2.5 Generalized Subadditivity Properties

Another result of our work in Chapter 4 is that we further establish the generalized subadditivity properties of the matrix φ-entropy defined on the $*$-algebra, which is the tracial full non-commutativity. A $*$-subalgebra $\mathfrak{A}$ is a subspace of self-adjoint matrices with fixed dimension which contains the matrix identity and is close under matrix multiplication and conjugation. As explained by Carlen [45], orthogonal projection onto a $*$-subalgebra resembles taking conditional expectation in probability theory. We use functions in the same Φ function class to define a generalized concept of φ-entropy conditional on a $*$-subalgebra:

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}) := \bar{\operatorname{tr}}[\varphi(\boldsymbol{A}) - \varphi(\mathbb{E}_{\mathfrak{A}}\, \boldsymbol{A})] \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d,$$

where $\mathbb{E}_{\mathfrak{A}} : \mathbb{H}^d \mapsto \mathfrak{A}$ is the projection onto the $*$-subalgebra $\mathfrak{A}$.

Two $*$-subalgebras $\mathfrak{A}_1$, $\mathfrak{A}_2$ commute when the result of sequentially projecting a matrix onto one and the other does not depend on the ordering:

$$(\mathbb{E}_{\mathfrak{A}_1}\, \mathbb{E}_{\mathfrak{A}_2})\boldsymbol{M} = (\mathbb{E}_{\mathfrak{A}_2}\, \mathbb{E}_{\mathfrak{A}_1})\boldsymbol{M} \quad \text{for all } \boldsymbol{M} \in \mathbb{H}^d.$$

The φ-entropy can be defined on commuting $*$-subalgebras. For example, when we have two commuting $*$-subalgebras, the corresponding φ-entropy is

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_1, \mathfrak{A}_2) := \bar{\operatorname{tr}}[\varphi(\boldsymbol{A}) - \varphi(\mathbb{E}_{\mathfrak{A}_1}\, \mathbb{E}_{\mathfrak{A}_2}\, \boldsymbol{A})] \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d.$$

A series of commuting $*$-subalgebras $\{\mathfrak{A}_1, \ldots, \mathfrak{A}_n\}$ corresponds to the conditional expectations that appear in the subadditivity property of the matrix φ-entropy and we establish the following subadditivity property of the matrix φ-entropy:

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_1, \ldots, \mathfrak{A}_n) \le \sum\nolimits_{i=1}^n H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_i) \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d. \tag{2.2.14}$$

The argument is very similar to the steps of establishing the subadditivity property of the φ-entropy in the usual expectation setting.

2.2.2.6 Discussions

In this section, we discuss the impacts of our work. We first evaluate our concentration results in Section 2.2.2.6.1. Then we summarize related works in Section 2.2.2.6.3.

2.2.2.6.1 Evaluating Our Results We first comment on the concentration inequalities obtained in our work. Our results, Theorems 2.2.3 and 2.2.4, are matrix extensions to the scalar concentration results in [26]. However, the requirement that the random matrices are invariant under sign permutation limits them from being applied widely. The most comparable result is Theorem 2.1.9, which contains bounded differences inequalities without assumptions on the matrix distribution. The main challenge of extending our results to more general situations arises when we bound the conditional entropy on the right-hand side of the modified logarithmic-Sobolev inequality (2.2.11). The sign permutation invariance assumption allows us to eliminate a tricky term when we control the conditional entropy.

Next, we evaluate our results in the framework of a potential matrix version of the entropy method. As we demonstrated in Section 1.2.3, in the scalar entropy method, we can either use a logarithmic-Sobolev inequality or the subadditivity property of the scalar entropy to bound the entropy of the random variable's moment generating function. In our work, we define the matrix φ-entropy and establish the corresponding subadditivity property. However, we fall short of establishing matrix versions of concentration results such as logarithmic-Sobolev inequalities or Poincaré inequalities, whose scalar counterparts have played an important role in establishing scalar concentration results for various applications. We acknowledge that Hansen independently developed similar subadditivity results in his work [92] .

2.2.2.6.2 Comparison with Other Methods of Deriving Matrix Concentration Inequalities As we summarized in Section 1.2.1.4, the scalar entropy method connects information theoretic inequalities and is a powerful approach to produce scalar concentration inequalities for general functions of independent random variables. This approach does not depend on the specific structure of the function of random variables. In the matrix setting, we do not have many general approaches of deriving concentration inequalities. For example, the Lieb's Theorem applies to sums of random matrices by decoupling their matrix moment generating functions, but is limited beyond this class of structured random matrices. The method of exchangeable pairs produces new concentration inequalities but at the moment it is not clear what class of matrices from which one can create an appropriate exchangeable pair.

Our goal of developing a matrix version of the entropy method is to find a general approach of deriving matrix concentration inequalities that does not depend on the specific structure of random matrices in interest. Similarly, the matrix entropy method relies on some information inequalities in the matrix setting. However, in the matrix case, we do not have a well-established version of the entropy theory. So, by defining the matrix φ-entropy, we are laying out some first steps for an entropy theory of random matrices. We also develop matrix information inequalities that are required for the derivation of the subadditivity property of the matrix φ-entropy and the associated concentration inequalities. In addition, we show the subadditivity property also holds in the tracial non-commutative $*$-algebra setting. Due to the non-commutativity of matrices, the tools for developing matrix concentration inequalities are quite limited. Our work relies on advanced results from the operator theory, which will potentially lead to new techniques for random matrices.

Besides deriving concentration inequalities, the entropy method has connections with other fields as well. For example, the scalar entropy method has deep connections with results from Markov semigroups. Previously we discuss the Kernel Stein pairs method where the authors applied matrix coupling to construct Kernel Stein pairs. A complete matrix entropy theory will potentially lead a detailed understanding of the connections. The definition of matrix entropy draws inspirations from the von Neumann entropy. In return, a thorough study of matrix entropy might lead to new results in quantum information theory as well. We discuss these connections based on the research after our work is published in the next section.

2.2.2.6.3 Related Work After our work is published, various authors continue to explore and study the properties of the matrix φ-entropy. In [179], Pitrik and Virosztek show that the a scalar function satisfies the conditions to generate a matrix φ-entropy if and only if the corresponding matrix f-Bregman divergence from this scalar function is jointly convex, thus establishing the equivalence of the matrix φ-entropy with the joint convexity of the Bregman divergence. The authors establish an improved inequality for the Tsallis entropy of a tripartite state which generalizes the strong subadditivity property of the von Neumann entropy. In [93], Hansen and Zhang provide alternative and transparent characterizations of the matrix φ-entropy. In [248], Zhang show that a larger class of power functions satisfy the conditions for the Φ function class. The recent works [55, 56] establish important results based on the matrix φ-entropy. We detail their contributions next.

2.2.2.6.3.1 New Characterization of Matrix Φ-Entropy, Matrix Poincaré and Sobolev Inequalities, and the Holevo Quantity The work of Cheng and Hsieh [55] extends significantly our results on the matrix φ-entropy. First, Cheng and Hsieh provide an augmented characterization of the matrix φ-entropy and show that the matrix φ-entropy satisfies all known equivalent characterizations for the scalar φ-entropy. The authors pointed out that the additional characterizations for the matrix φ-entropy are useful in many instances and they establish clear connections with other related topics, as exemplified in [179] mentioned earlier.

Second, based on the subadditivity property (2.2.12) of the matrix φ-entropy proved in our work, Cheng and Hsieh extend a important set of classical concentration inequalities to the matrix setting. The first result is a new proof of the matrix Efron–Stein inequality which controls the trace variance of a matrix function $\boldsymbol{Z} := \mathcal{L}(\boldsymbol{X}) := \mathcal{L}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ of independent random variables $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ defines as

$$\mathrm{Var}(\boldsymbol{Z}) := \bar{\mathrm{tr}}\big[\, \mathbb{E}(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})^2 \big] \tag{2.2.15}$$

with the expected local perturbation of $\boldsymbol{Z}$:

$$\mathrm{Var}(\boldsymbol{Z}) \leq \sum\nolimits_{i=1}^{n} \bar{\mathrm{tr}}\, \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')_+^2 \right], \tag{2.2.16}$$

where $\boldsymbol{Z}_i' := \mathcal{L}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_i', \boldsymbol{X}_{i+1}, \dots, \boldsymbol{X}_n)$ differs from $\boldsymbol{Z}$ with an independent instantiation X_i' of X_i. The right-hand side of (2.2.16) captures the conditional variance of the function $\mathcal{L}(\boldsymbol{X})$ denoted as $\mathcal{E}(\boldsymbol{Z})$:

$$\begin{aligned} \mathcal{E}(\boldsymbol{Z}) := \mathcal{E}(\mathcal{L}(\boldsymbol{X})) &:= \sum\nolimits_{i=1}^{n} \bar{\mathrm{tr}}\, \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')_+^2 \right] \\ &= \frac{1}{2} \sum\nolimits_{i=1}^{n} \bar{\mathrm{tr}}\, \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')^2 \right]. \end{aligned} \tag{2.2.17}$$

Note that $\mathcal{E}(\boldsymbol{Z})$ resembles the quantities (1.2.30) and (1.2.31) that appear in the scalar entropy method (Section 1.2.3.3.4).

Matrix Efron–Stein inequality first appeared in [173], which produced a more general version based on the matrix exchangeable pairs method. The constant of (2.2.16) in the special case of [55] has a better constant. The overlapped result implies the theoretical connections between the matrix exchangeable pairs method and the matrix entropy method. Based on (2.2.16), the authors establish matrix versions of the classical Poincaré inequality and the classical φ-Sobolev inequality, which are fundamental results behind the classical entropy method for scalar random variables.. The matrix Poincaré inequality bounds the trace variance of the matrix function $\mathcal{L}(\boldsymbol{X})$ that is separately convex by the expected norm of the Fréchet derivatives taken with respect to individual random matrices $\{\boldsymbol{X}_i\}$:

$$\mathrm{Var}(\mathcal{L}(\boldsymbol{X})) \leq \sum\nolimits_{i=1}^{n} \mathbb{E}\left[\|\mathrm{D}_{\boldsymbol{X}_i}\mathcal{L}[\boldsymbol{X}]\|_2^2\right]. \tag{2.2.18}$$

The matrix φ-Sobolev inequality of [55] controls the matrix φ-entropy of a non-negative matrix value function $\boldsymbol{F} : \mathbb{R}^n \mapsto \mathbb{H}_+^d$ taking n independent random variables $(X_1, \ldots, X_n)$ as input variables when φ is a power function. The exact form of the matrix φ-Sobolev inequality depends on the distribution of $\{X_i\}$. When $\{X_i\}$ are independent Bernoulli taking the values of 0 and 1 with equal probability, the following φ-Sobolev holds for all $p \in (1, 2)$

$$H_\varphi(\boldsymbol{F}^p) \leq (2-p)\mathcal{E}(\boldsymbol{F}) \cdot d^{1-2/p} + \bar{\mathrm{tr}}\, \mathbb{E}[\boldsymbol{F}^2] \cdot (1 - d^{1-2/p}), \tag{2.2.19}$$

with $\varphi : x \mapsto x^{2/p}$ and $\mathcal{E}(\boldsymbol{F})$ is the conditional variation (2.2.17). When $\{X_i\}$ are independent standard Gaussian random variables, the φ-Sobolev inequality takes a different form:

$$H_\varphi(\boldsymbol{F}^p) \leq (2-p) \sum\nolimits_{i=1}^{n} \mathbb{E}\left[\, \|\mathrm{D}_{X_i}\boldsymbol{F}(\boldsymbol{X})\|_2^2 \,\right] \cdot d^{1-2/p} + \bar{\mathrm{tr}}\, \mathbb{E}[\boldsymbol{F}^2] \cdot (1 - d^{1-2/p}), \tag{2.2.20}$$

with $p \in (1, 2)$ and $\varphi : x \mapsto x^{2/p}$. The authors show that the φ-Sobolev inequalities (2.2.19) and (2.2.20) lead to the logarithmic-Sobolev inequalities for both distributions, which control the matrix entropy, that is the matrix φ-entropy with $\varphi : x \mapsto x \log x$. The logarithmic-Sobolev inequality for symmetric Bernoulli random variables is

$$H(\boldsymbol{F}^2) \leq 2\mathcal{E}(\boldsymbol{F}) + \log(d) \cdot \bar{\mathrm{tr}}\, \mathbb{E}\left[\boldsymbol{F}^2\right].$$

The logarithmic-Sobolev inequality for independent Gaussian random variables is

$$H(\boldsymbol{F}^2) \leq 2\sum\nolimits_{i=1}^{n} \mathbb{E}\left[\, \|\mathrm{D}_{X_i}\boldsymbol{F}(\boldsymbol{X})\|_2^2 \,\right] + \log(d) \cdot \bar{\mathrm{tr}}\, \mathbb{E}\left[\boldsymbol{F}^2\right].$$

These inequalities are very important in constructing a potential complete framework of the entropy method for random matrices.

In the last contribution, the authors of [55] show an interesting connection of the matrix φ-entropy to the quantum information theory. The demonstrate that the matrix φ-entropy conincides with the Holevo quantity. They prove an upper bound of the Holevo quantity for quantum ensembles with Markov evolution. The Holevo quantity [97] is an important quantity in quantum information theory as it measures the quantity of information that a quantum communication channel can transmit. The authors of [55] point out that their upper bound for the Holevo quantity is a stronger form of the strong data processing inequality [180, 181], which is key in classical information theory.

2.2.2.6.3.2 Exponential Decay of Matrix Φ-Entropies on Markov Semigroup and Applications Based on the matrix φ-entropy and the associated matrix functional inequalities, matrix Poincaré and φ-Sobolev inequalities, developed in [55], Cheng et. al. [56] extend the study of the Markov semigroup to the matrix setting. Considering the matrix function $\boldsymbol{F} : \Omega \mapsto \mathbb{C}^{d\times d}$, the authors use a Markov semigroup evolves the matrix function $\boldsymbol{F}$ according to

$$\mathrm{P}_t\boldsymbol{F}(x) = \int_{y\in\Omega} \mathrm{T}_t(x, \mathrm{d}y) \odot \boldsymbol{F}(y),$$

where the evolution of the matrix dynamical system is governed by a completely positive map $\mathrm{T}_t(x, \mathrm{d}y) : \mathbb{C}^{d\times d} \mapsto \mathbb{C}^{d\times d}$ such that $\int_{y \int \Omega} \mathrm{T}_t(x, \mathrm{d}y)$ is unital.

In the scalar setting, under proper conditions the variance and φ-entropy decreases exponentially under the influence of a Markov semigroup [47]. The authors of [56] also establish the necessary and sufficient conditions for the exponential decay of matrix trace variance and matrix φ-entropy under Markov semigroup. The arguments of [56] to establish such results parallel those in the scalar setting. The first component is the infinitesimal generator for a Markov semigroup $\{\mathrm{P}_t\}_{t\geq 0}$:

$$\mathrm{L}(\boldsymbol{F}) := \lim_{t\to 0_+} \frac{1}{t} \cdot (\mathrm{P}_t\boldsymbol{F} - \boldsymbol{F}), \tag{2.2.21}$$

which characterizes the infinitesimal 'gradient' due to the Markov semigroup. The set of matrix-valued functions $\boldsymbol{F}$ such that the above limit (2.2.21) exists is called the Dirichelet domain $\mathcal{D}(\mathrm{L})$ of L. The second component is the following carré du champ operator $\boldsymbol{\Gamma} : \mathcal{D}(\mathrm{L}) \times \mathcal{D}(\mathrm{L}) \mapsto \mathcal{D}(\mathrm{L})$:

$$\boldsymbol{\Gamma}(\boldsymbol{F}, \boldsymbol{F}) := \frac{1}{2} \cdot \big(\mathrm{L}(\boldsymbol{F}^2) - \boldsymbol{F}\mathrm{L}(\boldsymbol{F}) - \mathrm{L}(\boldsymbol{F})\boldsymbol{F}\big).$$

The carré du champ operator has a symmetric and bilinear extension:

$$\boldsymbol{\Gamma}(\boldsymbol{F}, \boldsymbol{G}) = \boldsymbol{\Gamma}(\boldsymbol{G}, \boldsymbol{F}) := \frac{1}{2} \cdot \big(\boldsymbol{\Gamma}(\boldsymbol{F}+\boldsymbol{G}, \boldsymbol{F}+\boldsymbol{G}) - \boldsymbol{\Gamma}(\boldsymbol{F}, \boldsymbol{F}) - \boldsymbol{\Gamma}(\boldsymbol{G}, \boldsymbol{G})\big),$$

which reduces to the following simple form when the two matrix functions $\boldsymbol{F}$ and $\boldsymbol{G}$ commute:

$$\boldsymbol{\Gamma}(\boldsymbol{F}, \boldsymbol{G}) = \frac{1}{2} \cdot \big(\mathrm{L}(\boldsymbol{F}\boldsymbol{G}) - \boldsymbol{F}\mathrm{L}(\boldsymbol{G}) - \boldsymbol{G}\mathrm{L}(\boldsymbol{F})\big).$$

The third component is the invariant measure for the semigroup $\{\mathrm{P}_t\}_{t\geq 0}$

$$\int \mathrm{P}_t\boldsymbol{F}(x)\mu(\mathrm{d}x) = \int \boldsymbol{F}(x)\mu(\mathrm{d}x), \quad \text{for all } t \in \mathbb{R}_+.$$

The fourth component is a symmetric bilinear Dirichlet form, which integrates the carré du champ operator $\boldsymbol{\Gamma}$ with respect to the invariant measure μ:

$$\boldsymbol{\mathcal{E}}(\boldsymbol{F}, \boldsymbol{G}) := \int \boldsymbol{\Gamma}(\boldsymbol{F}, \boldsymbol{G})\mathrm{d}\mu.$$

The authors prove that the necessary and sufficient condition for the matrix φ-entropy to decay exponentially under the Markov semigroup $\{\mathrm{P}_t\}_{t\geq 0}$

$$H_\varphi(\mathrm{P}_t\boldsymbol{F}) \leq \mathrm{e}^{-t/C} \cdot H_\varphi(\boldsymbol{F}), \quad \text{for all } t \geq 0$$

is that the Markov triple $(\Omega, \boldsymbol{\Gamma}, \mu)$ satisfies the following φ-Sobolev inequality

$$H_{\varphi}(\boldsymbol{F}) \leq -C \operatorname{tr} \mathbb{E}_{\mu} \left[\varphi'(\boldsymbol{F}) \mathrm{L}(\boldsymbol{F})\right],$$

with a positive constant C. As a corollary, the authors establish that the necessary and sufficient condition for the trace variance (2.2.15) to decay exponentially

$$\operatorname{Var}(\mathrm{P}_t \boldsymbol{F}) \leq \mathrm{e}^{-2t/C} \cdot \operatorname{Var}(\boldsymbol{F})$$

is that the Markov triple $(\Omega, \boldsymbol{\Gamma}, \mu)$ satisfies the following spectral gap inequality with a constant $C > 0$ for all matrix functions $\boldsymbol{F}$:

$$\operatorname{Var}(\boldsymbol{F}) \leq C \operatorname{tr}[\boldsymbol{\mathcal{E}}(\boldsymbol{F})].$$

The authors of [56] also show that the exponential decay of the matrix φ-entropy under the Markov semigroup has immediate implications in quantum information theory. In particular, since the matrix φ-entropy coincides with the Holevo quantity, the authors show that the Holevo quantity of a quantum ensemble decays exponentially through a Markov dynamical evolution that does not depend on the history. In addition, based on the subadditivity property (2.2.12) of the matrix φ-entropy, the authors characterize the convergence rate of a Markov jump process defined on the Boolean hypercude. A final result in the paper studies random walks on the quantum random graph and bounds the mixing time of the quantum ensembles.

Chapter 3

The Masked Sample Covariance Estimator: An Analysis via the Matrix Laplace Transform Method

Preface

This chapter is adopted from the technical report [53]. Another version where we analyze the masked sample covariance estimator using matrix moment concentration inequalities is published as [52] in the journal of Information and Inference. They are collaboratively produced by the candidate, the candidate's advisor Joel A. Tropp, and Alex Gittens, who was a senior graduate student at the time of the work.

3.1 Introduction

In this section, we provide an overview of masked covariance estimation and its relationship with classical covariance estimation. In Section 3.1.6, we present a simplified result for the behavior of the masked sample covariance estimator applied to a Gaussian distribution, and we offer a concrete comparison with the results of Levina and Vershynin [128, Thm. 2.1]. More detailed results appear in Section 3.3.

3.1.1 Classical Covariance Estimation

Consider a random vector

$$\boldsymbol{x} = (X_1, X_2, \dots, X_p)^* \in \mathbb{R}^p.$$

Let $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ be independent random vectors that follow the same distribution as $\boldsymbol{x}$. For simplicity, we assume that the distribution is known to have zero mean: $\mathbb{E}\,\boldsymbol{x} = \mathbf{0}$. The *covariance matrix* $\mathbf{\Sigma}$ is a $p \times p$ matrix that tabulates the second-order statistics of the distribution:

$$\mathbf{\Sigma} := \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*), \tag{3.1.1}$$

where * denotes the transpose operation. The classical estimator for the covariance matrix is the *sample covariance matrix*, which is obtained from (3.1.1) by the plug-in principle:

$$\widehat{\mathbf{\Sigma}}_n := \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^*. \tag{3.1.2}$$

The sample covariance matrix is an unbiased estimator of the covariance matrix.

Given a tolerance $\varepsilon \in (0,1)$, we can study how many samples n are typically required to provide an estimate with relative error ε in the spectral norm:

$$\mathbb{E}\left\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\right\| \le \varepsilon \left\|\mathbf{\Sigma}\right\|. \tag{3.1.3}$$

This type of spectral-norm error bound is quite powerful. It limits the magnitude of the estimation error for each entry of the covariance matrix; it provides information about the variance of each marginal of the distribution of $\boldsymbol{x}$; it even controls the error in estimating the eigenvalues of the covariance using the eigenvalues of the sample covariance.

Unfortunately, an error bound of the form (3.1.3) demands a lot of samples. Suppose that the covariance matrix has full rank. Then the number of samples must be at least as large as the number of variables to obtain a nontrivial guarantee. Indeed, when $n < p$, the sample covariance does not even have full rank, so the spectral norm error is bounded away from zero!

Typical positive results on covariance estimation state that we can obtain an accurate estimate for the covariance matrix when the number of samples is proportional to the number of variables, provided that the distribution decays fast enough. For example, assuming that $\boldsymbol{x}$ follows a normal distribution,

$$n \geq \mathrm{C}\,\varepsilon^{-2} p \quad \Longrightarrow \quad \left\| \widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma} \right\| \leq \varepsilon \left\| \mathbf{\Sigma} \right\| \quad \text{with high probability.} \tag{3.1.4}$$

We use the analyst's convention that C denotes an absolute constant whose value may change from appearance to appearance. See [239, Thm. 57 et seq.] for details of obtaining the bound (3.1.4). The work of Srivastava and Vershynin [204] contains the most recent news on the classical covariance estimation problem.

3.1.2 Motivation for Masked Covariance Estimation

In the regime $n \ll p$, where we have very few samples, we can never hope to achieve the estimate (3.1.3). So we must lower our standards. The following example provides some insight on how to proceed.

Example 3.1.1 (Simultaneous Variance Estimation). Let us how many realizations of a Gaussian random vector we need to accurately estimate the variance of each component.

First, suppose that Z is a zero-mean normal variable with variance v. Given independent copies $Z_1, \dots, Z_n$ of the random variable Z, we can compute the sample variance

$$\widehat{v} := \frac{1}{n} \sum\nolimits_{i=1}^{n} Z_i^2.$$

The estimator $\widehat{v}$ is unbiased, and it follows a chi-square distribution, so the probability of error satisfies

$$\mathbb{P}\left\{ |\widehat{v} - v| \geq tv \right\} \leq 2\,\mathrm{e}^{-nt^2/4} \quad \text{for } t \geq 0. \tag{3.1.5}$$

For a clean proof of this inequality, see [11, Lect. 1].

Next, suppose that the random vector $\boldsymbol{x}$ follows a zero-mean normal distribution with arbitrary covariance $\mathbf{\Sigma}$, and write σ_{ij} for the (i, j) entry of this matrix. When we use the sample covariance to estimate each of the p diagonal entries of $\mathbf{\Sigma}$, the bound (3.1.5) implies that

$$\mathbb{P}\left\{ \max\nolimits_i |(\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma})_{ii}| \geq (\max\nolimits_i \sigma_{ii}) \cdot t \right\} \leq 2p\,\mathrm{e}^{-nt^2/4}.$$

We conclude that

$$n \geq \mathrm{C}\,\varepsilon^{-2} \log p \quad \Longrightarrow \quad \max_i |(\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma})_{ii}| \leq \varepsilon \max_i \sigma_{ii} \quad \text{with high probability.} \tag{3.1.6}$$

Since $\max_i \sigma_{ii} \leq \|\mathbf{\Sigma}\|$, the error obtained in (3.1.6) is smaller than the spectral-norm error in (3.1.4).

When the covariance $\mathbf{\Sigma} = \mathbf{I}$, it can be shown that at least $\log p$ samples are *required* to achieve the bound (3.1.6).

> Example 3.1.1 suggests an intriguing possibility. Although we need at least p samples to estimate the entire covariance matrix, roughly $\log p$ samples suffice to estimate the diagonal. It turns out that this phenomenon is generic: *If we estimate only a small portion of the covariance matrix, then we can reduce the number of samples dramatically.* This observation is widely applicable because there are many problems where we do not need to know all of the second-order statistics.

Partitioning Variables Suppose that we divide the stock market into disjoint sectors, and we would like to study the interactions among the monthly returns for stocks within each sector. The list of returns for all the stocks can be treated as a random vector. We block the covariance matrix of this random vector to conform with the market sectors, and we estimate only the entries in the diagonal blocks.

Spatial or Temporal Localization A simple random model for grayscale images treats the intensity of each pixel as a random variable. Nearby pixels tend to be bright or dark together, while distant pixels are usually uncorrelated. Thus, we might limit our attention to the interactions between a pixel and the pixels directly adjacent to it. This model suggests that we estimate the entries of the covariance that lie within a (generalized) band about the diagonal.

Graph Structures Consider a stochastic model for the spread of an epidemic through a social network. At each time instant, we label an individual with a random variable that measures how sick he is. Since transmission only occurs along links in the network, neighbors are likely to be sick or well together. As a result, we might want to focus on estimating the covariance for individuals separated by one degree. In this case, the adjacency matrix of the graph determines which pairs to estimate.

3.1.3 The Mask Matrix

We can treat all the examples from Section 3.1.2 using a formalism that was introduced by Levina and Vershynin [128]. Let $\boldsymbol{M}$ be a fixed $p \times p$ symmetric matrix with real entries, which we call the *mask matrix*. The basic idea is to construct a mask that guides our attention to specific parts of the covariance matrix.

In the simplest case, the mask has 0–1 values that indicate which entries of the covariance we must attend to. The presence of a unit entry $m_{ij} = 1$ tells us to estimate the interaction between the ith and jth variable; a zero entry $m_{ij} = 0$ means that we abdicate from making any estimate of this interaction. In Example 3.1.1, we are only interested in the diagonal entries of the covariance, so we are using the mask $\boldsymbol{M}_{\text{diag}} = \mathbf{I}$. Here are some other basic examples:

$$\boldsymbol{M}_{\text{group}} := \begin{bmatrix} 1 & 1 & & & \\ 1 & 1 & & & \\ & & 1 & 1 & \\ & & 1 & 1 & \\ & & & & 1 \end{bmatrix}; \quad \boldsymbol{M}_{\text{band}} := \begin{bmatrix} 1 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 1 & 1 & \\ & & 1 & 1 & 1 \\ & & & 1 & 1 \end{bmatrix}; \quad \boldsymbol{M}_{\text{graph}} := \begin{bmatrix} 1 & & 1 & 1 & \\ & 1 & & & 1 \\ 1 & & 1 & & 1 \\ 1 & & & 1 & \\ & 1 & 1 & & 1 \end{bmatrix}.$$

The matrix $\boldsymbol{M}_{\text{group}}$ corresponds to the case where we partition variables into three subgroups, and we make estimates only within subgroups. Masks such as $\boldsymbol{M}_{\text{band}}$ arise from banded covariance estimation, which occurs for spatially localized random fields. The mask $\boldsymbol{M}_{\text{graph}}$ might occur when the variables exhibit a graphical dependency structure.

In more complicated situations, we can allow the mask to take arbitrary nonnegative values and then interpret the magnitude of each entry as a requirement on the precision of the estimate. When m_{ij} is large, we must study the interaction between the ith and jth variable carefully. When m_{ij} is small, we are less vigilant about how well we estimate the (i, j) entry of the covariance matrix. An example of a mask with general entries is the Kac matrix

$$\boldsymbol{M}_{\text{Kac}} := \begin{bmatrix} 1 & \varphi & \varphi^2 & \varphi^3 & \varphi^4 \\ \varphi & 1 & \varphi & \varphi^2 & \varphi^3 \\ \varphi^2 & \varphi & 1 & \varphi & \varphi^2 \\ \varphi^3 & \varphi^2 & \varphi & 1 & \varphi \\ \varphi^4 & \varphi^3 & \varphi^2 & \varphi & 1 \end{bmatrix} \quad \text{where } \varphi \in (0,1).$$

The mask $\boldsymbol{M}_{\text{Kac}}$ tapers the covariances exponentially depending on the distance $|i-j|$ between the variables. This type of example might be relevant for the study of spatially localized processes.

Most of the regularization techniques for sparse covariance estimation studied in the literature, such as [21, 77, 40], can be described using mask matrices. The initial works focus on specific cases, such as banded masks and tapered masks, whereas we have followed Levina and Vershynin [128] by allowing an arbitrary symmetric matrix $\boldsymbol{M}$. We refer to the papers cited in this paragraph for further background and references.

Remark 3.1.2. Let us emphasize that the entries of the mask can take both positive and negative values, but it is harder to find a clear interpretation of a mask that has negative entries.

3.1.4 The Masked Sample Covariance Estimator

Suppose that we have specified a symmetric $p \times p$ mask $\boldsymbol{M}$ with real entries. The *masked covariance* and the *masked sample covariance estimator* are the two matrices

$$\boldsymbol{M} \odot \boldsymbol{\Sigma} \quad \text{and} \quad \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n,$$

where the symbol $\odot$ denotes the componentwise (i.e., Schur or Hadamard) product. The goal of this work is to study the error incurred when we estimate the masked covariance matrix using the masked sample covariance:

$$\left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\|. \tag{3.1.7}$$

As noted by Levina and Vershynin [128, Sec. 1], control on the error (3.1.7) also delivers information about how well we estimate the full covariance because

$$\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\right\| \leq \left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| + \left\|\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\right\|. \tag{3.1.8}$$

The first term in (3.1.8) reflects the variance of the estimator about its mean value, while the second term represents the bias in the estimate owing to the presence of the mask. It is important to select a mask $\boldsymbol{M}$ that simultaneously controls both the variance and the bias. Understanding the variance term requires an excursion into random matrix theory, and it comprises the main subject of this work. Studying the bias term involves only a deterministic analysis, which should be undertaken with a specific application in mind.

When the error (3.1.7) is small, the masked sample covariance yields accurate estimates for each component of the covariance where the corresponding entry of $\boldsymbol{M}$ is large, as well as the variance of some specially chosen marginals. When the error (3.1.8) is also small, the masked sample covariance provides additional information about the variance of all marginals of the distribution of $\boldsymbol{x}$, as well as estimates for the eigenvalues of the covariance.

3.1.5 The Complexity of a Mask

The number of samples we need to control (3.1.7) depends on "how much" of the covariance matrix we are attempting to estimate. We quantify the complexity of the mask using two separate metrics. First, define the square of the maximum column norm of the mask matrix:

$$\|\boldsymbol{M}\|_{1\to 2}^2 := \max\nolimits_j \left(\sum\nolimits_i m_{ij}^2\right).$$

Roughly, the parenthesis reflects the number of interactions we want to estimate that involve the variable j, and the maximum computes a bound over all p variables. The second metric is the spectral norm $\|\boldsymbol{M}\|$ of the mask matrix, which provides a more global view of the complexity of the interactions that we estimate.

Some examples may illuminate how these metrics reflect the properties of the mask.

First, suppose that we estimate the entire covariance matrix, so the mask is the matrix of ones:

$$\boldsymbol{M} = \text{matrix of ones} \quad \Longrightarrow \quad \|\boldsymbol{M}\|_{1\to 2}^2 = p \quad \text{and} \quad \|\boldsymbol{M}\| = p.$$

We will see that the value p here corresponds with the factor p in the sample complexity bound (3.1.4). Next, consider the mask that arises in banded covariance estimation:

$$\boldsymbol{M} = \text{0–1 matrix, bandwidth } B \quad \Longrightarrow \quad \|\boldsymbol{M}\|_{1\to 2}^2 \le B \quad \text{and} \quad \|\boldsymbol{M}\| \le B$$

because there are at most B ones in each row and column. When $B \ll p$, the banded mask is much less complex than the matrix of ones, and estimation is commensurately easier. Third, assuming the mask is a Kac matrix, we have

$$\boldsymbol{M} = \text{Kac matrix, parameter } \varphi \quad \Longrightarrow \quad \|\boldsymbol{M}\|_{1\to 2}^2 \le \frac{1}{1-\varphi^2} \quad \text{and} \quad \|\boldsymbol{M}\| \le \frac{1}{1-\varphi}.$$

For a fixed value of φ, neither quantity depends on the total number of variables, so covariance estimation with this mask should require very few samples.

Remark 3.1.3. In each example above, the two metrics take very similar values, but this coincidence does not always occur. Although the spectral norm dominates the maximum column norm, the *square* of the maximum column norm can be substantially larger or substantially smaller than the spectral norm. We have omitted examples to support this point because they do not seem to arise naturally in the setting of masked covariance estimation.

3.1.6 Masked Covariance Estimation for Gaussian Distributions

This paper develops a bound for the estimation error (3.1.7) when the random vector $\boldsymbol{x}$ follows a subgaussian distribution with zero mean. For illustrative purposes, this section focuses on the simpler case where the random vector has a normal distribution. The general results appear in Section 3.3.

Theorem 3.1.4 (Masked Covariance Estimation for Gaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a Gaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in* (3.1.1) *and* (3.1.2)*. Then the expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\le 8\left[\left(\frac{\|\boldsymbol{M}\|_{1\to 2}^2\log(6p)}{n}\right)^{1/2}+\frac{\|\boldsymbol{M}\|\log^2(6np)}{n}\right]\|\boldsymbol{\Sigma}\|. \qquad (3.1.9)$$

Theorem 3.1.4 is a simplified version of Corollary 3.3.3. The reader is encouraged to examine the full result, which includes several substantial refinements.

Remark 3.1.5. In the actual practice of covariance estimation, we center each sample empirically by subtracting the sample mean $\bar{\boldsymbol{x}}=n^{-1}\sum_{i=1}^n\boldsymbol{x}_i$. The sample covariance (3.1.2) is computed using the centered samples $\widetilde{\boldsymbol{x}}_i=\boldsymbol{x}_i-\bar{\boldsymbol{x}}$ instead of the original samples $\boldsymbol{x}_i$. The theory in this paper can be extended to cover the masked covariance estimator formed with centered samples; see [128, Rem. 4] for the details of the argument.

3.1.6.1 Sample Complexity Bound

Theorem 3.1.4 allows us to develop conditions on the number n of samples that we need to control the estimation error with high probability. Markov's inequality can be used to convert (3.1.9) into an error bound that holds in probability. For example, with probability at least 99%,

$$\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\le \mathrm{C}\left[\left(\frac{\|\boldsymbol{M}\|_{1\to 2}^2\log p}{n}\right)^{1/2}+\frac{\|\boldsymbol{M}\|\log^2(np)}{n}\right]\|\boldsymbol{\Sigma}\|. \qquad (3.1.10)$$

For stronger exponential error bounds, we refer to Corollary 3.3.3. To obtain the sample complexity, assume that $n\le p$, and let $\varepsilon\in(0,1)$. Then (3.1.10) yields the statement

$$n\ge \mathrm{C}\left[\varepsilon^{-2}\|\boldsymbol{M}\|_{1\to 2}^2\log p+\varepsilon^{-1}\|\boldsymbol{M}\|\log^2 p\right]\quad\Longrightarrow\quad\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\le\varepsilon\|\boldsymbol{\Sigma}\| \qquad (3.1.11)$$

with probability at least 99%.

3.1.6.2 Is This Sample Complexity Bound Optimal?

Levina and Vershynin show that the sample complexity of masked covariance estimation must exhibit a logarithmic dependence on the number p of variables [128, Rem. 3].

They also argue that there should be a linear dependence on the maximum number of interactions that involve a single variable [128, Eqn. (1.4) et seq.]; this term appears in (3.1.11) in the guise of $\|\boldsymbol{M}\|_{1\to 2}^2$. As a consequence of these observations, it seems plausible that the first summand in the sample bound (3.1.11) has the optimal form. On the other hand, we believe that the factor $\log^2 p$ in the second summand could probably be reduced to $\log p$.

The discussion in Example 3.1.1 suggests that it may be possible to improve the dependence of the sample complexity bound (3.1.11) on the spectral norm $\|\boldsymbol{\Sigma}\|$ of the covariance. Indeed, we have obtained a refinement of this type. See Corollary 3.3.3 for details.

3.1.6.3 Application Example

Consider the banded covariance estimation problem, with the mask

$$\boldsymbol{M} = \text{0–1 matrix with bandwidth } B.$$

See the matrix $\boldsymbol{M}_{\text{band}}$ displayed on page 82 for an instance with $B = 3$ and $p = 5$. The sample complexity bound (3.1.11) and the norm calculations from Section 3.1.5 demonstrate that

$$n \geq \mathrm{C}\left[\varepsilon^{-2} B \log p + \varepsilon^{-1} B \log^2 p\right] \tag{3.1.12}$$

is sufficient to provide a relative estimation error ε in spectral norm with 99% probability. For comparison, recall the sufficient condition (3.1.4) that the sample complexity for estimating the entire covariance with relative error ε satisfies

$$n \geq \mathrm{C}\,\varepsilon^{-2} p.$$

When the bandwidth is much smaller than the number of variables ($B \ll p$), the masked covariance estimator outperforms the classical covariance estimator. On the other hand, when the bandwidth is comparable with the number of variables, the analysis of the masked covariance estimator gives a sample complexity bound (3.1.12) that is worse by a polylogarithmic factor.

We remark that, when ε is constant, the second summand in (3.1.12) always dominates the first as $p \to \infty$. On the other hand, the first summand is larger when $\varepsilon \le \log^{-1} p$. In other words, the excess logarithm in the second term of (3.1.12) does not have an impact on the sample complexity when we are seeking highly accurate covariance estimates.

3.1.6.4 Comparison with Bounds of Levina and Vershynin

Theorem 3.1.4 should be compared with the main result of Levina and Vershynin [128, Thm. 2.1], which states that

$$\mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \le \mathrm{C}\left[\frac{\|\boldsymbol{M}\|_{1\to 2}\log^{5/2} p}{\sqrt{n}} + \frac{\|\boldsymbol{M}\|\log^3 p}{n}\right]\|\boldsymbol{\Sigma}\|\,.$$

The associated sample complexity bound is

$$n \ge \mathrm{C}\left[\varepsilon^{-2}\|\boldsymbol{M}\|_{1\to 2}^2 \log^5 p + \varepsilon^{-1}\|\boldsymbol{M}\|\log^3 p\right]. \tag{3.1.13}$$

Our sample complexity bound (3.1.11) has exactly the same structure as (3.1.13), but we have managed to remove a moderate number of logarithms.

We do not feel that chopping down logs is an interesting pursuit *per se*. Instead, the value of this work stems from the fact that we have applied an argument that is completely different from previous work on masked covariance estimation. Our approach provides some qualitative refinements over Levina and Vershynin's bound in the Gaussian setting (Corollary 3.3.3), and it also extends to the general subgaussian distributions (Theorem 3.3.2).

3.1.6.5 Proof Techniques

The argument in this paper is based on a recent set of ideas, collectively known as the *matrix Laplace transform method*. This approach can be regarded as a generalization of the classical technique, attributed to Bernstein, that develops probability inequalities for a random variable in terms of bounds for its cumulant generating function. Tropp [233], building on work of Ahlswede and Winter [1], demonstrates that the scalar approach admits a tight analogy in the matrix setting. See Section 3.2.4 for an overview of this technique.

The matrix Laplace transform method is particularly well suited for studying sums of independent random matrices. To apply these techniques, we express the error as a sum of i.i.d. random matrices, each with zero mean:

$$\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{M} \odot (\boldsymbol{x}_i \boldsymbol{x}_i^* - \mathbb{E}\, \boldsymbol{x}\boldsymbol{x}^*).$$

The main challenge is to study the matrix cumulant generating function of each summand:

$$\log \mathbb{E} \exp\left(\theta \boldsymbol{M} \odot (\boldsymbol{x}_i \boldsymbol{x}_i^* - \mathbb{E}\, \boldsymbol{x}\boldsymbol{x}^*)\right) \quad \text{for } \theta > 0. \tag{3.1.14}$$

The key technical result of this paper is a semidefinite upper bound for the matrix cgf (3.1.14). This estimate requires a number of substantial new ideas, including a symmetrization argument, a careful analysis of the variance of the random matrix in the exponent of (3.1.14), and a delicate truncation bound.

3.1.7 Organization of the paper

The rest of the paper is organized as follows. Section 3.2 introduces our notation and some preliminaries. Section 3.3 presents the main result for zero-mean subgaussian distributions, together with its proof and the proof of Theorem 3.1.4. In Section 3.4, we deal with the technical challenge of estimating the matrix cumulant generating function (3.1.14).

3.2 Preliminaries

This section sets out the background material we require for the proof. The argument depends on a very recent set of ideas, collectively known as the *matrix Laplace transform method*. We introduce the main results from this theory in Section 3.2.4, and we provide references to the primary sources. The rest of the material here is more or less standard. Section 3.2.1 states our notational conventions, Section 3.2.2 describes some basic properties of the Schur product, and Section 3.2.3 includes key facts about subgaussian random variables.

3.2.1 Notation and Conventions

In this paper, we work exclusively with real numbers. Plain italic letters always refer to scalars. Bold italic lowercase letters, such as $\boldsymbol{a}$, refer to column vectors. Bold italic uppercase letters, such as $\boldsymbol{A}$, denote matrices. All matrices in this work are square; the dimensions are determined by context. We write $\mathbf{0}$ for the zero matrix and $\mathbf{I}$ for the identity matrix. The matrix unit $\mathbf{E}_{ij}$ has a unit entry in the (i, j) position and zeros elsewhere.

The symbol * denotes the transpose operation on vectors and matrices. We use the term *self-adjoint* to refer to a matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$ to avoid confusion between symmetric matrices and symmetric random variables. Curly inequalities refer to the positive-semidefinite partial ordering on self-adjoint matrices: $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ if and only if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

The function $\operatorname{diag}(\cdot)$ maps a vector $\boldsymbol{a}$ to a matrix whose diagonal entries correspond with the entries of $\boldsymbol{a}$. We write $\operatorname{tr}(\cdot)$ for the trace of a matrix. The symbol $\odot$ denotes the componentwise (i.e., Schur or Hadamard) product of two matrices.

We write $\|\cdot\|$ for both the ℓ_2 vector norm and the associated operator norm, which is usually called the *spectral norm*. The norm $\|\cdot\|_\infty$ returns the absolute maximum entry of a vector. For clarity, we use a separate notation $\|\cdot\|_{\max}$ for the absolute maximum entry of a matrix. The maximum column norm $\|\cdot\|_{1\to 2}$ is defined as

$$\|\boldsymbol{A}\|_{1\to 2} := \max_j \left(\sum_i |a_{ij}|^2\right)^{1/2}.$$

The notation reflects the fact that this is the natural norm for linear maps from ℓ_1 into ℓ_2.

We reserve the symbol ε for a *Rademacher random variable*, which takes the two values ±1 with equal probability. We also assume that all random variables are sufficiently regular that we are justified in computing expectations, interchanging limits, and so forth.

3.2.2 Facts about the Schur Product

The proof depends on some basic properties of Schur products. The first result is a simple but useful algebraic identity. For each square matrix $\boldsymbol{A}$ and each conforming vector $\boldsymbol{x}$,

$$\boldsymbol{A} \odot \boldsymbol{x}\boldsymbol{x}^* = \operatorname{diag}(\boldsymbol{x})\boldsymbol{A}\operatorname{diag}(\boldsymbol{x}). \tag{3.2.1}$$

The second result states that the Schur product with a positive-semidefinite matrix is order preserving. That is, for a fixed positive-semidefinite matrix $\boldsymbol{A}$,

$$\boldsymbol{B}_1 \preccurlyeq \boldsymbol{B}_2 \quad \text{implies} \quad \boldsymbol{A} \odot \boldsymbol{B}_1 \preccurlyeq \boldsymbol{A} \odot \boldsymbol{B}_2. \tag{3.2.2}$$

This property follows from Schur's theorem [98, Thm. 7.5.3], which demonstrates that the Schur product of two positive-semidefinite matrices remains positive semidefinite.

3.2.3 Subgaussian Random Variables

There are several different ways to formalize the concept of a random variable that decays faster than a Gaussian random variable [239]. For the purposes of this paper, the following definition is most convenient.

Definition 3.2.1 (Subgaussian random variable)**.** *A random variable X is* subgaussian *if there exists a positive constant K such that*

$$\mathbb{P}\{|X| > t\} \le 2\,\mathrm{e}^{-t^2/K^2} \quad \textit{for all } t \ge 0.$$

The subgaussian coefficient $\kappa(X)$ *is defined to be the infimal K for which this inequality holds.*

We can bound all the moments of a subgaussian random variable X in terms of its subgaussian coefficient:

$$\mathbb{E}\,|X|^q = \int_0^\infty qt^{q-1}\,\mathbb{P}\{|X| > t\}\,\mathrm{d}t \le \int_0^\infty qt^{q-1} \cdot 2\,\mathrm{e}^{-t^2/\kappa(X)^2}\,\mathrm{d}t = 2\kappa(X)^q\,\Gamma(q/2+1).$$

In particular, the raw fourth moment of X satisfies

$$\mathbb{E}\,|X|^4 \le 4\kappa(X)^4. \tag{3.2.3}$$

3.2.4 The Matrix Laplace Transform Method

In classical probability, the Laplace transform method is a powerful tool for obtaining tail bounds for a sum of independent random variables. In their influential paper [1], Ahlswede and Winter describe a generalization of the Laplace transform method that applies to a sum of independent random matrices. Subsequent papers by Oliveira [163, 164], by Tropp [233, 231], and by Hsu et al. [103] all contain substantial refinements and extensions of the original idea. Altogether, these tools are easy to use, remarkably effective, and widely applicable.

In analogy with the scalar case, we study large deviations using a matrix version of the moment generating function (mgf) and the cumulant generating function (cgf). Let $\boldsymbol{Z}$ be a self-adjoint random matrix. Using the matrix exponential, we define the matrix mgf and matrix cgf, respectively, to be

$$\boldsymbol{M}_{\boldsymbol{Z}}(\theta) := \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{Z}} \quad \text{and} \quad \boldsymbol{\Xi}_{\boldsymbol{Z}}(\theta) := \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{Z}} \quad \text{for } \theta \in \mathbb{R}.$$

Note that these expectations may not exist for all values of θ. The matrix cgf can be interpreted as an *exponential mean*, an average that emphasizes large deviations of the spectrum with the same sign as the parameter θ.

The matrix mgf contains valuable information about the behavior of the maximum eigenvalue of a symmetric random matrix. The following result is a matrix analog of the classical approach to large deviations, which is attributed to Bernstein.

Proposition 3.2.2 (Matrix Laplace transform bound). *Let $\boldsymbol{Z}$ be a random, self-adjoint matrix. For each $t \in \mathbb{R}$,*

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Z}) \geq t\right\} \leq \inf_{\theta > 0} \left\{ \mathrm{e}^{-\theta t} \cdot \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Z}} \right\}. \tag{3.2.4}$$

In this form, Proposition 4.7.4 is due to Oliveira [164, Sec. 3], but the main idea goes back to the paper [1] of Ahlswede and Winter. See [233, Prop. 3.1] for a succinct proof.

In our application, the random matrix $\boldsymbol{Z}$ can be expressed as a sum of i.i.d. zero-mean random, self-adjoint matrices. The argument relies on a symmetrization procedure, which introduces additional randomness into the series.

Proposition 3.2.3 (Symmetrization bound)**.** *Consider a sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$ of independent, random, self-adjoint matrices. For each $\theta \in \mathbb{R}$,*

$$\mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{i=1}^n \theta(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\right) \leq \mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{i=1}^n 2\theta\varepsilon_i\boldsymbol{Y}_i\right),$$

where $\{\varepsilon_i\}$ are independent Rademacher random variables that are also independent from $\{\boldsymbol{Y}_i\}$.

The proof of Proposition 3.2.3 is essentially identical with the proof of Lemma 7.6 in [233], so we omit the argument.

The matrix Laplace transform method derives its power from a deep technical result that allows us to bound the mgf of a sum of independent random matrices in terms of the cgfs of the summands. We state a simplified version of this fact that suits our needs.

Proposition 3.2.4 (Subadditivity of cgfs)**.** *Let $\boldsymbol{Y}$ be a random, self-adjoint matrix. Consider a finite sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$ of independent copies of $\boldsymbol{Y}$. For each $\theta \in \mathbb{R}$,*

$$\mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{i=1}^n \theta\boldsymbol{Y}_i\right) \leq \operatorname{tr}\exp\left(n\log\mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{Y}}\right).$$

Proposition 3.2.4 is due to Tropp [233, Lem. 3.4]. The main ingredient in the proof is a celebrated concavity theorem established by Lieb [131, Thm. 6].

We use these techniques to develop a matrix Bernstein inequality that is adapted for partial covariance estimation. The final ingredient in our argument is a matrix mgf bound that parallels the classical mgf bound underlying Bernstein's inequality.

Proposition 3.2.5 (Bernstein matrix mgf bound)**.** *Let $\boldsymbol{Y}$ be a random, self-adjoint matrix that satisfies*

$$\mathbb{E}\,\boldsymbol{Y} = \mathbf{0} \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{Y}) \leq R \quad \textit{almost surely.}$$

When $\theta \in (0, R^{-1})$,

$$\mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{Y}} \preccurlyeq \mathbf{I} + \frac{\theta^2}{2(1-\theta R)} \cdot \mathbb{E}(\boldsymbol{Y}^2).$$

Proposition 3.2.5 follows immediately from [233, Lem. 6.7] and the classical inequality

$$\frac{\mathrm{e}^{\theta R} - \theta R - 1}{R^2} \leq \frac{\theta^2}{2(1-\theta R)} \quad \text{valid for } \theta \in (0, R^{-1}).$$

We can verify this bound by comparing derivatives. The constants in this inequality can be improved, but we have chosen the version here to streamline other aspects of the argument.

3.3 Masked Covariance Estimation for a Subgaussian Distribution

In this section, we state and prove our main error estimates for masked covariance estimation. Section 3.3.1 defines two concentration parameters that measure the spread of the distribution. We present the main theorem for subgaussian distributions in Section 3.3.2, and we specialize to Gaussian distributions in Section 3.3.3. Section 3.3.4 shows how to derive the result for Gaussian matrices from the main theorem. Finally, we establish the main result in Section 3.3.5.

3.3.1 Concentration Parameters

The effectiveness of the masked sample covariance estimator depends on the concentration properties of the distribution of $\boldsymbol{x}$. Let us introduce two quantities that measure different facets of the variation of the random vector.

The *subgaussian coefficient* $\kappa(\boldsymbol{x})$ of the distribution is defined to be the maximum subgaussian coefficient of a single component of the vector:

$$\kappa(\boldsymbol{x}) := \max_i \kappa(X_i). \tag{3.3.1}$$

In other words, we assume that each component of the distribution exhibit subgaussian decay with variance controlled by $\kappa(\boldsymbol{x})^2$.

We do not need every marginal of the distribution to be subgaussian with controlled variance, but we do require some information on the spread of the distribution in other directions. Define the *uniform fourth moment* $\nu(\boldsymbol{x})$ by the formula

$$\nu(\boldsymbol{x}) := \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/4}. \tag{3.3.2}$$

The uniform fourth moment measures how much the worst marginal varies.

Note that both $\kappa(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ have the same homogeneity as the random vector $\boldsymbol{x}$. (This property is sometimes expressed by saying that the quantities have the same dimension, the same units, or the same scaling.) As a consequence, $\kappa^2(\boldsymbol{x})$ and $\nu^2(\boldsymbol{x})$ have the same homogeneity as the covariance matrix $\boldsymbol{\Sigma}$.

In the sequel, we abbreviate $\kappa := \kappa(\boldsymbol{x})$ and $\nu := \nu(\boldsymbol{x})$ whenever the distribution of the random vector $\boldsymbol{x}$ is clear.

Remark 3.3.1. For Gaussian distributions, the uniform fourth moment ν always dominates the subgaussian coefficient κ. In the worst case, ν can be much larger than κ. Indeed, suppose that X is a standard normal random variable, and consider the random vector $\boldsymbol{x} = (X, X, \dots, X)^* \in \mathbb{R}^p$. Although the subgaussian coefficient $\kappa(\boldsymbol{x}) = \sqrt{2}$, the directional fourth moment $\nu(\boldsymbol{x}) = 12^{1/4}\sqrt{p}$.

For other kinds of distributions, the subgaussian coefficient κ may be substantially larger than the uniform fourth moment ν. Examples of this phenomenon already emerge in the univariate case.

3.3.2 Main Result for Masked Covariance Estimation

The following theorem provides detailed information about the expectation and tail behavior of the error in the masked sample covariance estimator for a zero-mean subgaussian distribution.

Theorem 3.3.2 (Masked Covariance Estimation for Subgaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a subgaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in* (3.1.1) *and* (3.1.2)*. Then the expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M}\odot\boldsymbol{\Sigma}\right\| \le \left[\frac{16\kappa^2\nu^2\left\|\boldsymbol{M}\right\|_{1\to 2}^2 \log(2\mathrm{e}p)}{n}\right]^{1/2} + \frac{4\kappa^2\left\|\boldsymbol{M}\right\|\log^2(2\mathrm{e}np)}{n}. \tag{3.3.3}$$

Furthermore, for each $t > 0$, the estimation error satisfies the tail bound

$$\mathbb{P}\left\{\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M}\odot\boldsymbol{\Sigma}\right\| \ge t\right\} \le 2\mathrm{e}p\cdot\exp\left(\frac{-nt^2/2}{8\kappa^2\nu^2\left\|\boldsymbol{M}\right\|_{1\to 2}^2 + 4\kappa^2\left\|\boldsymbol{M}\right\|\log(4np)\cdot t}\right). \tag{3.3.4}$$

The subgaussian coefficient κ and the uniform fourth moment ν are defined in (3.3.1) *and*

(3.3.2).

The proof of Theorem 3.3.2 appears in Section 3.3.5. We can extend this result to the case where we center the observations using the sample mean before computing the sample covariance; the argument is identical with the one described by Levina and Vershynin [128, Rem. 4] for the Gaussian case.

3.3.2.1 Interpretation and Consequences

Let us take a moment to discuss Theorem 3.3.2. First, we note that the error in the masked sample covariance estimator can be expressed as

$$\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{M} \odot (\boldsymbol{x}_i \boldsymbol{x}_i^* - \mathbb{E}\, \boldsymbol{x}\boldsymbol{x}^*), \tag{3.3.5}$$

using the definitions (3.1.1) and (3.1.2) of the covariance and sample covariance. For each i, the parenthesis in (3.3.5) has subexponential tails because the random vector $\boldsymbol{x}_i$ is subgaussian. Therefore, the formula (3.3.5) expresses the error as an average of subexponential random variables.

Consequently, we expect the estimation error to obey a probability inequality just like (3.3.4). For moderate values of t, the error (3.3.4) exhibits subgaussian decay, an intimation of the normal profile that emerges when the number of samples tends to infinity. For large values of t, the error has subexponential decay, owing to the heavier tails of the summands in (3.3.5). Likewise, the two terms in the expected error bound (3.3.3) correspond with the two regimes in the tail bound. The first term reflects the subgaussian decay, while the second term comes from the subexponential decay.

The scale for subgaussian decay is controlled by a measure of the variance σ^2 of each summand:

$$\sigma^2 = 8\kappa^2\nu^2 \left\| \boldsymbol{M} \right\|_{1\to 2}^2 .$$

We see that moderate deviations depend on the local properties of the mask, as encapsulated in $\|\boldsymbol{M}\|_{1\to 2}^2$. The appearance of the subgaussian coefficient κ in σ^2 reflects the variance of each component of the random vector. The presence of the uniform fourth moment ν shows that there is also a role for the spread of the random vector in every direction.

The scale for subexponential decay is controlled by a second quantity,

$$R = 4\kappa^2 \|\boldsymbol{M}\| \log(4np).$$

Large deviations reflect more global properties of the mask owing to the presence of $\|\boldsymbol{M}\|$. The subgaussian coefficient κ arises here because the tails of the distribution drive the tails of the error. Note that the large-deviation behavior only depends on the individual components of the random vector being subgaussian; we attribute this fact to the basis-dependent nature of the Schur product. The logarithmic factor in R emerges from a truncation argument, and we believe it is parasitic.

We can obtain a sample complexity bound directly from the probability inequality (3.3.4) in Theorem 3.3.2. Assume that $n \leq p$ and that $\varepsilon \in (0,1)$. Then

$$n \geq \mathrm{C} \cdot \frac{\kappa^2}{\nu^2} \left[\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log p}{\varepsilon^2} + \frac{\|\boldsymbol{M}\| \log^2 p}{\varepsilon} \right] \quad \Longrightarrow \quad \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\| \leq \varepsilon \nu^2 \qquad (3.3.6)$$

with high probability. The square ν^2 of the uniform fourth moment has the same homogeneity as the covariance matrix, so (3.3.6) is a type of relative error bound. As before, the first summand reflects the subgaussian part of the tail, while the second summand comes from the subexponential part. A novel feature of the sample bound (3.3.6) is the presence of the ratio κ^2/ν^2, which is a dimensionless measure of the shape of the distribution. This ratio can be very large or very small, so it should be assessed within the scope of a particular application.

3.3.3 Specialization to Gaussian Distributions

It is natural to apply Theorem 3.3.2 to study the performance of masked covariance estimation for a zero-mean Gaussian random vector. In this case, the covariance matrix determines the distribution completely, so we can obtain a more transparent statement that does not involve the concentration parameters κ and ν.

Corollary 3.3.3 (Masked Covariance Estimation for Gaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a Gaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in* (3.1.1) *and* (3.1.2). *Then the expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\leq\sqrt{\frac{56\left\|\boldsymbol{\Sigma}\right\|_{\max}\left\|\boldsymbol{\Sigma}\right\|\left\|\boldsymbol{M}\right\|_{1\to2}^2\log(6p)}{n}}+\frac{8\left\|\boldsymbol{\Sigma}\right\|_{\max}\left\|\boldsymbol{M}\right\|\log^2(6np)}{n}.$$

Furthermore, for each $t>0$*, the estimation error satisfies the tail bound*

$$\mathbb{P}\left\{\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\geq t\right\}$$
$$\leq 6p\cdot\exp\left(\frac{-nt^2}{56\left\|\boldsymbol{\Sigma}\right\|_{\max}\left\|\boldsymbol{\Sigma}\right\|\left\|\boldsymbol{M}\right\|_{1\to2}^2+16\left\|\boldsymbol{\Sigma}\right\|_{\max}\left\|\boldsymbol{M}\right\|\log(4np)\cdot t}\right).$$

The proof of Corollary 3.3.3 appears below in Section 3.3.4. Theorem 3.1.4 of the Introduction follows quickly from this result when we apply the inequality $\left\|\boldsymbol{\Sigma}\right\|_{\max}\leq\left\|\boldsymbol{\Sigma}\right\|$ and complete some numerical estimates.

It is fruitful to compare Corollary 3.3.3 directly with earlier work on masked covariance estimation for a Gaussian distribution. Assume that $n\leq p$ and $\varepsilon\in(0,1)$. Then Corollary 3.3.3 delivers a sample complexity bound of the form

$$n\geq \mathrm{C}\cdot\frac{\left\|\boldsymbol{\Sigma}\right\|_{\max}}{\left\|\boldsymbol{\Sigma}\right\|}\left[\frac{\left\|\boldsymbol{M}\right\|_{1\to2}^2\log p}{\varepsilon^2}+\frac{\left\|\boldsymbol{M}\right\|\log^2 p}{\varepsilon}\right]\quad\Longrightarrow\quad\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\leq\varepsilon\left\|\boldsymbol{\Sigma}\right\| \tag{3.3.7}$$

with high probability. The bound (3.3.7) is similar with the results of Levina and Vershynin [128], stated in (3.1.13), but two improvements are worth mentioning.

First, recall that the sample complexity bound (3.1.6) we present in Example 3.1.1 depends on the absolute maximum entry of the covariance matrix, rather than its spectral norm. A similar refinement appears in the bound (3.3.7) on account of the ratio of the two norms. This ratio never exceeds one, and it can be as small as p^{-1} for particular choices of the covariance matrix. We interpret this term as saying that covariance estimation is easier when the variables are highly correlated with each other. This represents a new phenomenon that previous authors have not identified.

The second improvement over (3.1.13), which has less conceptual significance, is the reduction of the number of logarithmic factors.

3.3.4 Proof of Corollary 3.3.3 from Theorem 3.3.2

The result for Gaussian distributions is a direct consequence of the main theorem because the covariance matrix $\mathbf{\Sigma}$ of a zero-mean Gaussian vector $\boldsymbol{x}$ characterizes the distribution completely. As a consequence, we just need to estimate the concentration parameters $\kappa(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ in terms of $\mathbf{\Sigma}$.

First, we compute the subgaussian coefficient $\kappa(\boldsymbol{x})$. Observe that the ith component X_i of the vector $\boldsymbol{x}$ is a Gaussian random variable with variance σ_{ii}, where σ_{ii} denotes the ith diagonal entry of $\mathbf{\Sigma}$. The usual Gaussian tail bound demonstrates that

$$\mathbb{P}\left\{|X_i| > t\right\} \leq 2\,\mathrm{e}^{-t^2/2\sigma_{ii}}.$$

According to Definition 3.2.1, the subgaussian coefficient $\kappa(X_i)^2 \leq 2\sigma_{ii}$, and so the subgaussian coefficient of the vector satisfies

$$\kappa(\boldsymbol{x})^2 \leq \max_i 2\sigma_{ii} = 2\left\|\mathbf{\Sigma}\right\|_{\max}.$$

The latter equality holds because the absolute maximum entry of a positive-definite matrix occurs on its diagonal.

Next, we compute the uniform fourth moment $\nu(\boldsymbol{x})$. Fix a unit vector $\boldsymbol{u}$. The distribution of the marginal $\boldsymbol{u}^*\boldsymbol{x}$ is Gaussian with mean zero. To compute the variance $\sigma_{\boldsymbol{u}}^2$ of the marginal, we write $\boldsymbol{x} = \mathbf{\Sigma}^{1/2}\boldsymbol{g}$, where $\boldsymbol{g}$ is a standard Gaussian vector. Then

$$\sigma_{\boldsymbol{u}}^2 = \mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^2 = \mathbb{E}\,|\boldsymbol{u}^*(\mathbf{\Sigma}^{1/2}\boldsymbol{g})|^2 = \boldsymbol{u}^*\mathbf{\Sigma}^{1/2}(\mathbb{E}\,\boldsymbol{g}\boldsymbol{g}^*)\mathbf{\Sigma}^{1/2}\boldsymbol{u} = \boldsymbol{u}^*\mathbf{\Sigma}\boldsymbol{u} \leq \left\|\mathbf{\Sigma}\right\|.$$

The fourth moment of a Gaussian variable equals three times its squared variance, so

$$\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4 = 3\sigma_{\boldsymbol{u}}^4 \leq 3\left\|\mathbf{\Sigma}\right\|^2.$$

We conclude that the uniform fourth moment satisfies

$$\nu(\boldsymbol{x}) = \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/4} \leq 3^{1/4}\left\|\mathbf{\Sigma}\right\|^{1/2}.$$

To complete the argument, substitute the estimates for $\kappa(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ into Theorem 3.3.2 and make some numerical estimates.

3.3.5 Proof of Theorem 3.3.2

The argument follows the same lines as the classical Laplace transform technique. For clarity, we break the presentation into discrete steps.

3.3.5.1 The Matrix Laplace Transform Method

We begin with the proof of the probability inequality (3.3.4). First, split the tail bound for the spectral norm into two pieces:

$$\begin{aligned}\mathbb{P}\left\{\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\geq t\right\}\\ \leq \mathbb{P}\left\{\lambda_{\max}(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\geq t\right\}+\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{M}\odot\boldsymbol{\Sigma}-\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n)\geq t\right\}.\end{aligned}\tag{3.3.8}$$

This inequality depends on the fact $\|\boldsymbol{A}\|=\max\{\lambda_{\max}(\boldsymbol{A}),\lambda_{\max}(-\boldsymbol{A})\}$, valid for each self-adjoint matrix $\boldsymbol{A}$, and an invocation of the union bound. We develop an estimate for the first term on the right-hand side of (3.3.8); an essentially identical argument applies to the second term.

The matrix Laplace transform bound, Proposition 4.7.4, allows us to control the first term on the right-hand side of (3.3.8) in terms of a matrix mgf.

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\geq t\right\}&=\mathbb{P}\left\{\lambda_{\max}\left(n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\right)\geq nt\right\}\\ &\leq\inf_{\theta>0}\left\{\mathrm{e}^{-\theta nt}\cdot\mathbb{E}\operatorname{tr}\exp\left(\theta n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\right)\right\}.\end{aligned}\tag{3.3.9}$$

In the first line of (3.3.9), we have rescaled both sides of the event and applied the positive homogeneity of the maximum eigenvalue. Let us introduce notation for the trace of the matrix mgf:

$$E(\theta):=\mathbb{E}\operatorname{tr}\exp\left(\theta n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\right).\tag{3.3.10}$$

Our main task is to obtain a suitable bound for $E(\theta)$.

3.3.5.2 Symmetrizing the Random Sum

The random matrix appearing in (3.3.10) admits a natural expression as a sum of centered, independent random matrices. To see why, substitute the definitions (3.1.1) and (3.1.2) of the population covariance matrix $\mathbf{\Sigma}$ and the sample covariance matrix $\widehat{\mathbf{\Sigma}}_n$ to obtain

$$E(\theta) = \mathbb{E} \operatorname{tr} \exp \left(\sum\nolimits_{i=1}^{n} \theta \big(\boldsymbol{M} \odot \boldsymbol{x}_i \boldsymbol{x}_i^* - \mathbb{E}\, \boldsymbol{M} \odot \boldsymbol{x}_i \boldsymbol{x}_i^* \big) \right).$$

The samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are statistically independent, so the summands are independent, centered random matrices. Therefore, we may apply the symmetrization lemma, Proposition 3.2.3, to reach

$$E(\theta) \le \mathbb{E} \operatorname{tr} \exp \left(\sum\nolimits_{i=1}^{n} 2\theta \varepsilon_i (\boldsymbol{M} \odot \boldsymbol{x}_i \boldsymbol{x}_i^*) \right), \tag{3.3.11}$$

where $\{\varepsilon_i\}$ is a sequence of independent Rademacher random variables that is also independent from the sequence $\{\boldsymbol{x}_i\}$ of samples. The benefit of the estimate (3.3.11) is that each Schur product involves a rank-one matrix, which greatly simplifies our computations.

3.3.5.3 Matrix cgf Bound for the Matrix mgf

The summands on the right-hand side of (3.3.11) are i.i.d., so we can apply Proposition 3.2.4 on the subadditivity of matrix cgfs to see that

$$E(\theta) \le \operatorname{tr} \exp(n \cdot \log \mathbb{E} \exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)). \tag{3.3.12}$$

The chief technical contribution of this paper consists in the following matrix cgf bound:

$$\log \mathbb{E} \exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I} \quad \text{when } \theta \in (0, R^{-1}), \tag{3.3.13}$$

where

$$\sigma^2 := 8\kappa^2\nu^2 \left\| \boldsymbol{M} \right\|_{1\to 2}^2 \quad \text{and} \quad R := 4\kappa^2 \left\| \boldsymbol{M} \right\| \log(4np). \tag{3.3.14}$$

The concentration parameters κ and ν that characterize $\boldsymbol{x}$ are defined as in (3.3.1) and (3.3.2). The calculation underlying (3.3.13) requires several pages and some substantial new ideas. We encapsulate the details in Lemma 3.4.1, which is the subject of Section 3.4.

The trace exponential is monotone with respect to the semidefinite order [174, Prop. 1], so we can substitute the cgf bound (3.3.13) into our estimate (3.3.12) for $E(\theta)$. Thus,

$$E(\theta) \leq \operatorname{tr} \exp\left(\frac{\theta^2 \sigma^2 n}{2(1-\theta R)} \cdot \mathbf{I} + \mathbf{I}\right) = \mathrm{e} p \cdot \exp\left(\frac{\theta^2 \sigma^2 n}{2(1-\theta R)}\right). \tag{3.3.15}$$

The second relation depends on the fact that the identity matrix has dimension p. The inequality (3.3.15) is just what we need to establish the probability inequality and the expectation bound that constitute the conclusions of Theorem 3.3.2.

3.3.5.4 Probability Bound for the Estimation Error

We are now prepared to complete our bound for the tail probability, initiated in (3.3.8). Substitute the estimate (3.3.15) for the matrix mgf into the Laplace transform bound (3.3.9) to discover that

$$\mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right) \geq t\right\} \leq \mathrm{e} p \cdot \inf_{\theta > 0} \exp\left(-\theta n t + \frac{\theta^2 \sigma^2 n}{2(1-\theta R)}\right).$$

Select the classical value for the parameter: $\theta = t/(\sigma^2 + Rt)$. This choice yields an upper bound for the first term on the right-hand side of (3.3.8):

$$\mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right) \geq t\right\} \leq \mathrm{e} p \cdot \exp\left(\frac{-n t^2}{2(\sigma^2 + R t)}\right). \tag{3.3.16}$$

The second term on the right-hand side of (3.3.8) admits the same upper bound:

$$\mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n\right) \geq t\right\} \leq \mathrm{e} p \cdot \exp\left(\frac{-n t^2}{2(\sigma^2 + R t)}\right). \tag{3.3.17}$$

The proof of (3.3.17) is essentially identical with the proof of (3.3.16), so we omit the details.

Finally, recall the definition (3.3.14) for the quantities σ^2 and R. Then introduce the relations (3.3.16) and (3.3.17) into the probability inequality (3.3.8) to establish the tail bound (3.3.4) stated in Theorem 3.3.2.

3.3.5.5 Bound for the Expected Estimation Error

Although it is possible to control the expected error by integrating the tail bound (3.3.4), we obtain somewhat better results through a direct application of the estimate (3.3.15) for the matrix mgf $E(\theta)$.

The argument is based on the following inequality, of independent interest, which provides a way to bound the expected spectral norm of a matrix in terms of its mgf. Let $\boldsymbol{Z}$ be a random, self-adjoint matrix, and fix a positive number θ. We have the following chain of relations:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\| &\le \theta^{-1}\log\mathbb{E}\,\mathrm{e}^{\theta\|\boldsymbol{Z}\|}\\ &= \theta^{-1}\log\mathbb{E}\,\mathrm{e}^{\max\{\lambda_{\max}(\theta\boldsymbol{Z}),\ \lambda_{\max}(-\theta\boldsymbol{Z})\}}\\ &= \theta^{-1}\log\mathbb{E}\max\left\{\lambda_{\max}(\mathrm{e}^{\theta\boldsymbol{Z}}),\lambda_{\max}(\mathrm{e}^{-\theta\boldsymbol{Z}})\right\}\\ &\le \theta^{-1}\log\left(\mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta\boldsymbol{Z}}+\mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{-\theta\boldsymbol{Z}}\right). \qquad (3.3.18)\end{aligned}$$

For the first inequality, multiply and divide by θ; then invoke Jensen's inequality to bound the expectation by an exponential mean. The second relation expresses the spectral norm of a symmetric matrix in terms of eigenvalues. In the third line, we pull the maximum through the exponential and then apply the spectral mapping theorem to draw out the eigenvalue maps. Finally, replace the maximum by a sum, and bound the maximum eigenvalue of the matrix exponential, which is positive definite, by the trace.

We intend to apply (3.3.18) to the random matrix

$$\boldsymbol{Z} = n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M}\odot\boldsymbol{\Sigma}).$$

According to the definition (3.3.10) of the function $E(\theta)$, the trace of the mgf of the matrix $\boldsymbol{Z}$ coincides with $E(\theta)$. Therefore, when the parameter $\theta \in (0, R^{-1})$, our upper bound (3.3.15) for $E(\theta)$ demonstrates that

$$\mathbb{E}\operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Z}} = E(\theta) \leq \mathrm{e}p \cdot \exp\left(\frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right). \tag{3.3.19}$$

The argument underlying the bound (3.3.15) for the trace mgf of $\boldsymbol{Z}$ also applies to $-\boldsymbol{Z}$, whereby

$$\mathbb{E}\operatorname{tr} \mathrm{e}^{-\theta \boldsymbol{Z}} \leq \mathrm{e}p \cdot \exp\left(\frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right). \tag{3.3.20}$$

Introduce (3.3.19) and (3.3.20) into the norm bound (3.3.18) to reach

$$n \cdot \mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq \theta^{-1}\left(\log(2\mathrm{e}p) + \frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right).$$

Minimize the right-hand side over admissible values of θ, ideally with a computer algebra system. This computation yields

$$n \cdot \mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq \sqrt{2\sigma^2 n \log(2\mathrm{e}p)} + R\log(2\mathrm{e}p).$$

Divide through by n and recall the definition (3.3.14) of the quantities σ^2 and R. Combine the two logarithms in the second term to complete the proof of the expected error bound (3.3.3) from Theorem 3.3.2.

3.4 The Matrix cgf of a Schur Product

In this section, we work out the details of the matrix cgf bound (3.3.13) that stands at the center of Theorem 3.3.2. The following lemma contains a complete statement of the result.

Lemma 3.4.1 (Matrix cgf Bound for a Schur Product)**.** *Fix a self-adjoint matrix $\boldsymbol{M}$. Let $\boldsymbol{x} = (X_1, \dots, X_p)^*$ be a random vector, and let ε be a Rademacher variable, independent from $\boldsymbol{x}$. For each positive integer n,*

$$\log \mathbb{E}\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I} \quad \textit{when } \theta \in (0, R^{-1}),$$

where

$$\sigma^2 := 8\kappa^2\nu^2 \left\|\boldsymbol{M}\right\|_{1\to 2}^2 \quad \textit{and} \quad R := 4\kappa^2 \left\|\boldsymbol{M}\right\| \log(4np).$$

The concentration parameters κ and ν associated with $\boldsymbol{x}$ are defined as in (3.3.1) *and* (3.3.2).

To prove Lemma 3.4.1, we would like to invoke the Bernstein mgf bound, Proposition 3.2.5, but several obstacles stand in the way. First, estimating the variance of the random matrix $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$ involves a surprisingly delicate calculation. Second, this random matrix is typically unbounded, which requires us to develop a new type of truncation argument. We address ourselves to these tasks in the next two subsections.

3.4.1 Computing the Variance

The Bernstein mgf bound demands that we compute the variance of the random matrix $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$. The following lemma contains this estimate. Our key insight is that the monotonicity (3.2.2) of the Schur product allows us to replace one factor in the product by a scalar matrix. This act of diagonalization simplifies the estimate tremendously because we erase the off-diagonal entries when we take the Schur product with an identity matrix.

Lemma 3.4.2 (Semidefinite variance bound)**.** *Under the assumptions of Lemma 3.4.1, it holds that*

$$\mathbb{E}(2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq 8\kappa^2\nu^2 \left\|\boldsymbol{M}\right\|_{1\to 2}^2 \cdot \mathbf{I}.$$

Proof. First, we treat the leading constant and the Rademacher random variable.

$$\mathbb{E}(2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = 4\,\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2. \tag{3.4.1}$$

The expectation with respect to $\boldsymbol{x}$ is not so easy to handle. To begin, we perform some algebraic manipulations to consolidate the remaining randomness. The Schur product identity (3.2.1) implies that

$$\begin{aligned}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 &= (\operatorname{diag}(\boldsymbol{x})\boldsymbol{M}\operatorname{diag}(\boldsymbol{x}))^2 \\ &= \operatorname{diag}(\boldsymbol{x})(\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})^2\boldsymbol{M})\operatorname{diag}(\boldsymbol{x}) = (\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})^2\boldsymbol{M}) \odot \boldsymbol{x}\boldsymbol{x}^*.\end{aligned}$$

Rewrite the diagonal matrix as a linear combination of matrix units: $\operatorname{diag}(\boldsymbol{x})^2 = \sum_i X_i^2\,\mathbf{E}_{ii}$.

The bilinearity of the Schur product now yields

$$(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = \left[\boldsymbol{M}\left(\sum\nolimits_i X_i^2\,\mathbf{E}_{ii}\right)\boldsymbol{M}\right] \odot \boldsymbol{x}\boldsymbol{x}^* = \sum\nolimits_i (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot (X_i^2\,\boldsymbol{x}\boldsymbol{x}^*).$$

Take the expectation of this expression to reach

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = \sum\nolimits_i (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\,\boldsymbol{x}\boldsymbol{x}^*)]. \tag{3.4.2}$$

Next, we invoke the monotonicity (3.2.2) of the Schur product to make a diagonal estimate for each summand in (3.4.2):

$$(\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)] \preccurlyeq \lambda_{\max}(\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)) \cdot (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot \mathbf{I}.$$

The Rayleigh–Ritz variational formula [17, Cor. III.1.2] allows us to write the maximum eigenvalue as a supremum. Thus,

$$\begin{aligned}\lambda_{\max}(\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)) &= \sup_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^*\,[\mathbb{E}(X_i^2\,\boldsymbol{x}\boldsymbol{x}^*)]\,\boldsymbol{u} = \sup_{\|\boldsymbol{u}\|=1} \mathbb{E}\left[X_i^2\,|\boldsymbol{u}^*\boldsymbol{x}|^2\right] \\ &\le \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,X_i^4)^{1/2}\,(\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/2} \le 2\kappa(X_i)^2 \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/2} \le 2\kappa^2\nu^2.\end{aligned}$$

The first inequality is Cauchy–Schwarz. For the second inequality, we apply (3.2.3) to bound the fourth moment of X_i in terms of the subgaussian coefficient. The final inequality follows from the definitions (3.3.1) and (3.3.2) of the concentration parameters. Combine the last two displays to obtain

$$(\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)] \preccurlyeq 2\kappa^2\nu^2 \cdot (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot \mathbf{I}. \tag{3.4.3}$$

To complete our bound for the variance, we introduce (3.4.3) into (3.4.2), which delivers

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq 2\kappa^2\nu^2 \cdot \boldsymbol{M}^2 \odot \mathbf{I}.$$

The remaining matrix is diagonal, so we can control it using only its maximum entry:

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq 2\kappa^2\nu^2 \max\nolimits_i(\boldsymbol{M}^2)_{ii} \cdot \mathbf{I} = 2\kappa^2\nu^2\,\|\boldsymbol{M}\|_{1\to 2}^2 \cdot \mathbf{I}.$$

The second relation follows from the fact that the diagonal entries of $\boldsymbol{M}^2$ list the squared norms of the columns of $\boldsymbol{M}$, and $\|\boldsymbol{M}\|_{1\to 2}$ computes the maximum column norm of $\boldsymbol{M}$. Substitute the latter expression into (3.4.1) to conclude. □

3.4.2 Proof of Lemma 3.4.1

This subsection contains the main steps in the proof of Lemma 3.4.1. We begin by explaining the motivation behind our approach.

We would like to invoke the Bernstein matrix mgf inequality, Proposition 3.2.5, to control the mgf of $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$. This proposition requires the maximum eigenvalue of the random matrix to satisfy an almost sure bound. Using the Schur product identity (3.2.1), we can develop a simple estimate for the maximum eigenvalue:

$$\lambda_{\max}(2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \leq 2\left\|\operatorname{diag}(\boldsymbol{x})\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})\right\| \leq 2\left\|\boldsymbol{M}\right\|\left\|\operatorname{diag}(\boldsymbol{x})\right\|^2 = 2\left\|\boldsymbol{M}\right\|\left\|\boldsymbol{x}\right\|_\infty^2. \quad (3.4.4)$$

Unfortunately, the random variable $\|\boldsymbol{x}\|_\infty$ is typically unbounded, which suggests that we cannot apply the Bernstein approach directly.

To tackle this problem, we develop a truncation argument in Section 3.4.2.1, which splits the distribution of the random matrix $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$ into two pieces, depending on the size of $\|\boldsymbol{x}\|_\infty$. This technique allows us to apply the Bernstein estimate to the bounded part of the random matrix (Section 3.4.2.2). To handle the unbounded part, we use the inequality (3.4.4) to develop a coarse tail estimate that we can integrate directly (Section 3.4.2.3). Section 3.4.2.4 combines these results to complete the argument.

3.4.2.1 The Truncation Argument

As we have explained, we intend to decompose the random matrix $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$ based on the magnitude of the random variable $\|\boldsymbol{x}\|_\infty$. To that end, define the event

$$\mathcal{A} := \{\|\boldsymbol{x}\|_\infty^2 \leq B\}, \quad (3.4.5)$$

where we determine a suitable truncation level B later.

Now, let us split the matrix mgf into expectations over $\mathcal{A}$ and $\mathcal{A}^c$:

$$\begin{aligned}\mathbb{E}\exp(2\theta\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*) &= \mathbb{E}\left[\exp(2\theta\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}\right] + \mathbb{E}\left[\exp(2\theta\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}^c}\right] \\ &\preccurlyeq \mathbb{E}\exp((2\theta\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}) + \mathbb{E}\left[\exp(2\theta\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}^c}\right]. \qquad (3.4.6)\end{aligned}$$

The first identity follows because the two indicators form a partition of unity. In the second line, notice that the first term can only increase in the semidefinite order when we draw the indicator $\mathbb{1}_{\mathcal{A}}$ into the exponential.

3.4.2.2 Bernstein Estimate for the Bounded Part of the Random Matrix

We can interpret the first term on the right-hand side of (3.4.6) as the mgf of a random matrix whose maximum eigenvalue is bounded; this matrix mgf admits a Bernstein-type estimate.

We must verify that the truncated matrix $(2\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}$ satisfies the hypotheses of Proposition 3.2.5. First, note that

$$\mathbb{E}[(2\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}] = \mathbf{0}$$

because the Rademacher variable ε is independent from $\boldsymbol{x}$ and hence from $\mathcal{A}$. Second, continuing the calculation (3.4.4), we determine that the maximum eigenvalue is bounded.

$$\lambda_{\max}((2\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}) \le 2\left\|\boldsymbol{M}\right\|\left\|\boldsymbol{x}\right\|_\infty^2 \cdot \mathbb{1}_{\mathcal{A}} \le 2B\left\|\boldsymbol{M}\right\|. \qquad (3.4.7)$$

The second inequality in (3.4.7) relies on the definition (3.4.5) of the truncation event. Third, we apply Lemma 3.4.2 to obtain a semidefinite bound for the variance.

$$\mathbb{E}[(2\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}]^2 \preccurlyeq \mathbb{E}[(2\varepsilon \boldsymbol{M}\odot \boldsymbol{x}\boldsymbol{x}^*)^2] \preccurlyeq 8\nu^2\kappa^2\left\|\boldsymbol{M}\right\|_{1\to 2}\cdot \mathbf{I}. \qquad (3.4.8)$$

Of course, discarding the indicator in (3.4.8) only increases the semidefinite order.

In view of (3.4.7) and (3.4.8), we define a variance parameter and a uniform bound parameter

$$\sigma^2 := 8\kappa^2\nu^2\left\|\boldsymbol{M}\right\|_{1\to 2}^2 \quad \text{and} \quad R := 2B\left\|\boldsymbol{M}\right\|. \qquad (3.4.9)$$

Finally, we apply Proposition 3.2.5 and the variance estimate (3.4.8) to achieve

$$\mathbb{E}\exp((2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}}) \preccurlyeq \mathbf{I} + \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I}. \tag{3.4.10}$$

The relation (3.4.10) is valid for all $\theta \in (0, R^{-1})$.

3.4.2.3 Controlling the Unbounded Part of the Random Matrix

We treat the second term on the right-hand side of (3.4.6) by making a rough bound that we can integrate directly. First, observe that

$$\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \exp(2\theta \cdot \lambda_{\max}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)) \cdot \mathbf{I} \preccurlyeq \exp(2\theta \left\| \boldsymbol{M} \right\| \left\| \boldsymbol{x} \right\|_\infty^2) \cdot \mathbf{I}.$$

We have applied the semidefinite relation $\mathrm{e}^{\boldsymbol{A}} \preccurlyeq \mathrm{e}^{\lambda_{\max}(\boldsymbol{A})} \cdot \mathbf{I}$, valid for each self-adjoint matrix $\boldsymbol{A}$, followed by the eigenvalue bound (3.4.4). Multiply both sides by the indicator $\mathbb{1}_{\mathcal{A}^c}$, and take the expectation to reach

$$\begin{aligned}\mathbb{E}[\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}^c}] &\preccurlyeq \mathbb{E}[\exp(2\theta \left\| \boldsymbol{M} \right\| \left\| \boldsymbol{x} \right\|_\infty^2)\mathbb{1}_{\mathcal{A}^c}] \cdot \mathbf{I} \\ &=: \mathbb{E}[\mathrm{e}^{\alpha W}\mathbb{1}_{\mathcal{A}^c}] \cdot \mathbf{I}. \end{aligned} \tag{3.4.11}$$

In the expression (3.4.11), we have abbreviated

$$\alpha := 2\theta \left\| \boldsymbol{M} \right\| \quad \text{and} \quad W := \left\| \boldsymbol{x} \right\|_\infty^2. \tag{3.4.12}$$

We apply classical techniques to bound the remaining expectation.

Observe that we can control the tail probability of W using the subgaussian coefficient κ. Indeed,

$$\begin{aligned}\mathbb{P}\left\{W > w\right\} &= \mathbb{P}\left\{\max_i \left|X_i\right|^2 > w\right\} \\ &\leq \sum\nolimits_{i=1}^p \mathbb{P}\left\{\left|X_i\right|^2 > w\right\} \leq \sum\nolimits_{i=1}^p 2\,\mathrm{e}^{-w/\kappa(X_i)^2} \leq 2p\,\mathrm{e}^{-w/\kappa^2}. \end{aligned} \tag{3.4.13}$$

The second relation is the union bound. The third follows from Definition 3.2.1 of the subgaussian coefficient $\kappa(X)$ of a random variable X, while the last depends on the definition (3.3.1) of the subgaussian coefficient κ of the random vector $\boldsymbol{x}$.

Next, we invoke a standard integration-by-parts argument [22, Eqn. (21.10)] to study the expectation in (3.4.11). Since $\mathcal{A}^c = \{W > B\}$,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\alpha W}\mathbb{1}_{\mathcal{A}^c}] &= \mathrm{e}^{\alpha B}\cdot\mathbb{P}\{W > B\} + \alpha\int_B^\infty \mathrm{e}^{\alpha w}\cdot\mathbb{P}\{W > w\}\,\mathrm{d}w \\ &\leq \mathrm{e}^{\alpha B}\cdot 2p\,\mathrm{e}^{-B/\kappa^2} + \alpha\int_B^\infty \mathrm{e}^{\alpha w}\cdot 2p\,\mathrm{e}^{-w/\kappa^2}\,\mathrm{d}w \\ &= 2p\left[1 + \frac{\alpha}{1/\kappa^2 - \alpha}\right]\mathrm{e}^{-(1/\kappa^2-\alpha)B}. \end{aligned} \tag{3.4.14}$$

We have used the tail bound (3.4.13) twice to obtain the inequality in the second line. The third line follows when we evaluate the definite integral under the assumption that $\alpha < 1/\kappa^2$.

To continue the bound on the right-hand side of (3.4.14), we need to make a careful estimate. Owing to definition (3.4.12) of α, the condition

$$\theta \leq \frac{1}{4\kappa^2\,\|\boldsymbol{M}\|} \quad\Longrightarrow\quad \alpha \leq \frac{1}{2\kappa^2}. \tag{3.4.15}$$

Assume that θ satisfies the hypothesis of (3.4.15). Now, observe that the right-hand side of the inequality (3.4.14) is an increasing function of α. Therefore, we may increase α to $1/2\kappa^2$ on the right-hand side of (3.4.14) and then set the truncation level

$$B = 2\kappa^2\log(4np) \tag{3.4.16}$$

to obtain the bound

$$\mathbb{E}[\mathrm{e}^{\alpha W}\mathbb{1}_{\mathcal{A}^c}] \leq 4p\,\mathrm{e}^{-B/2\kappa^2} = \frac{1}{n}.$$

Introduce this expression into (3.4.11) to conclude that

$$\mathbb{E}[\exp(2\theta\varepsilon\boldsymbol{M}\odot\boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathcal{A}^c}] \preccurlyeq \frac{1}{n}\cdot\mathbf{I}. \tag{3.4.17}$$

Finally, we verify that the truncation level B forces the parameter θ to satisfy the hypothesis of (3.4.15). Recall the definition (3.4.9) of the bound parameter and the definition (3.4.16) of the truncation level to see that

$$R = 2B \left\| \boldsymbol{M} \right\| = 4\kappa^2 \left\| \boldsymbol{M} \right\| \log(4np).$$

We have already assumed that $\theta < R^{-1}$. It follows that

$$\theta < \frac{1}{R} = \frac{1}{\log(4np)} \cdot \frac{1}{4\kappa^2 \left\| \boldsymbol{M} \right\|} \leq \frac{1}{4\kappa^2 \left\| \boldsymbol{M} \right\|}.$$

This observation completes the tail estimate.

3.4.2.4 Combining the Results

We have obtained estimates for the two terms in our truncation bound (3.4.6). Introduce (3.4.10) and (3.4.17) into (3.4.6) to reach

$$\mathbb{E} \exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \mathbf{I} + \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I},$$

where σ^2 and R are defined in (3.4.9). We have also assumed that $\theta \in (0, R^{-1})$. The logarithm is operator monotone [18, Exer. 4.2.5], so

$$\log \mathbb{E} \exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \log \left[\mathbf{I} + \frac{\theta^2\sigma^2}{2(1-R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I} \right].$$

To complete the proof of Lemma 3.4.1, we invoke the semidefinite relation $\log(\mathbf{I}+\boldsymbol{A}) \preccurlyeq \boldsymbol{A}$, which holds for each positive semidefinite matrix $\boldsymbol{A}$.

Chapter 4

Subadditivity of Matrix φ-Entropy and Concentration of Random Matrices

Preface

This chapter is adapted from the work [54] which appears in the Electronic Journal of Probability, coauthored by the candidate and the candidate's advisor, Joel A. Tropp.

4.1 Introduction and Related Work

Entropy and related functions quantify the uncertainty inherent in a probability distribution. Measures of entropy have the property that the total entropy of a "product" is bounded by the sum of the entropies of the "factors." This fundamental fact is called *subadditivity of entropy*, or sometimes *tensorization*, and it drives many applications of entropy. The survey [132] contains a discussion of subadditivity in statistical mechanics, and the monograph [182] describes examples in information theory. In this work, we focus on the role of subadditivity of entropy in probability.

4.1.1 Subadditivity and Concentration

A *concentration inequality* states that a random variable is unlikely to exhibit a significant deviation from its mean value. The current intuition holds that a random variable concentrates whenever it depends smoothly on many independent random variables [215]. Ledoux [123, 125] and Bobkov & Ledoux [24] initiated a line of research that uses methods based on entropy to derive concentration inequalities. A few of the many authors who have contributed include Massart [152, 153], Rio [189], Bousquet [31], and Boucheron et al. [28, 29]. See the book [30] for a comprehensive treatment of this theory and its bibliography.

Let us summarize the ideas that lead from entropy to concentration. In this setting, we define the *entropy functional* for each nonnegative, real random variable Z:

$$H(Z) := \mathbb{E}(Z \log Z) - (\mathbb{E}\, Z) \log(\mathbb{E}\, Z). \tag{4.1.1}$$

Heuristically, $H(Z)$ quantifies our uncertainty about the precise value of Z. We typically consider the situation where $Z = \mathrm{e}^{\theta Y}$ for a zero-mean random variable Y. In this case, we have the identity

$$\log \mathbb{E}\, \mathrm{e}^{\theta Y} = \theta \int_0^\theta \frac{H(\mathrm{e}^{\beta Y})}{\mathbb{E}\, \mathrm{e}^{\beta Y}} \cdot \frac{\mathrm{d}\beta}{\beta^2}. \tag{4.1.2}$$

Through Markov's inequality, bounds on the left-hand side imply that Y takes a large value with exponentially small probability. Therefore, we might hope to invoke inequalities for the entropy functional H to analyze the fluctuations of Y.

Indeed, the entropy functional exhibits a subadditivity property that allows us to implement this program. Suppose that Z is a function of mutually independent random variables $X_1, \dots, X_n$. We can define conditional entropy functionals

$$H_i(Z) := \mathbb{E}_i(Z \log Z) - (\mathbb{E}_i\, Z) \log(\mathbb{E}_i\, Z),$$

where $\mathbb{E}_i$ denotes the expectation with respect to X_i, holding X_j fixed for $j \neq i$. The conditional entropy H_i reflects the uncertainty about Z that is attributable to our lack of knowledge about X_i. Subadditivity is the nontrivial result that

$$H(Z) \leq \sum\nolimits_{i=1}^n \mathbb{E}[H_i(Z)]. \tag{4.1.3}$$

In other words, our uncertainty about Z does not exceed the total (average) uncertainty due to each X_i individually. Combining the identity (4.1.2) and the subadditivity property (4.1.3) with bounds for the conditional entropy, we can establish exponential concentration inequalities for functions of independent random variables.

The idea of considering alternative forms of entropy can be traced at least as far as the work of Rényi [188], Bregman [33], and Csiszár [58]. In the early 2000s, researchers [122, 47, 26, 48] recognized that generalized entropy functionals can exhibit subadditivity properties similar with those of the entropy functional (4.1.1). Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function. The *φ-entropy functional* is defined for each nonnegative random variable Z by the formula

$$H_\varphi(Z) := \mathbb{E}\,\varphi(Z) - \varphi(\mathbb{E}\,Z).$$

The functional (4.1.1) derives from the choice $\varphi : t \mapsto t \log t$. Under stringent conditions on φ, it can be shown that the φ-entropy functional satisfies a subadditivity property analogous with (4.1.3). In particular, the function $\varphi : t \mapsto t^p$ yields a subadditive φ-entropy when $1 \le p \le 2$, a fact that leads to polynomial concentration inequalities [26].

4.1.2 Subadditivity of Matrix Entropies

The purpose of this paper is to explore the subadditivity properties of entropy functionals defined on matrix-valued random variables. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function. For a positive-semidefinite (psd) random matrix $\boldsymbol{Z}$, we can consider the *matrix φ-entropy functional*

$$H_\varphi(\boldsymbol{Z}) := \mathbb{E}\,\bar{\operatorname{tr}}\,\varphi(\boldsymbol{Z}) - \bar{\operatorname{tr}}\,\varphi(\mathbb{E}\,\boldsymbol{Z}),$$

where φ refers to a standard matrix function and $\bar{\operatorname{tr}}$ denotes the normalized trace. See Section 4.2.1 for definitions. It may be helpful to note some alternative presentations of the matrix φ-entropy. First, the expression has the same structure as the scalar entropy (4.1.1) because

$$H_\varphi(\boldsymbol{Z}) = \mathbb{E}\,\Phi(\boldsymbol{Z}) - \Phi(\mathbb{E}\,\boldsymbol{Z}) \quad \text{where } \Phi := \bar{\operatorname{tr}}\,\varphi \text{ is convex.}$$

Second, we can decompose the matrix entropy as

$$H_\varphi(\boldsymbol{Z}) = \big[\mathbb{E}\,\bar{\operatorname{tr}}\,\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{Z})\big] + \big[\varphi(\bar{\operatorname{tr}}\,\mathbb{E}\,\boldsymbol{Z}) - \bar{\operatorname{tr}}\,\varphi(\mathbb{E}\,\boldsymbol{Z})\big].$$

In other words, the matrix entropy quantifies the total loss in two different averaging operations on the matrix.

This work contains two main contributions:

1. We develop conditions on φ which ensure that the matrix φ-entropy is subadditive.
2. We verify these conditions for the functions $\varphi : t \mapsto t \log t$ and $\varphi : t \mapsto t^p$ where $p \in [1, 2]$.

The arguments parallel the analysis of scalar φ-entropies in Boucheron et al. [26], but the technical difficulties are more formidable in the matrix setting.

There are several areas that may benefit from this investigation.

Random matrix theory In the scalar setting, subadditivity of φ-entropy leads to powerful concentration inequalities. The subadditivity of matrix φ-entropy allows us to adapt these arguments to obtain some concentration inequalities for random matrices.

Convex analysis We derive subadditivity of the matrix φ-entropy functional H_φ from its convexity properties; see Lemma 4.4.1 et seq. These results may be useful in other contexts. For example, the convexity of scalar φ-entropy plays a role in machine learning [187, Sec. 2.5 et seq.].

Operator theory To prove that specific examples of matrix φ-entropy are subadditive, we rely on sophisticated methods from operator theory. In return, the results here may be relevant for problems in operator theory.

Quantum theory In quantum statistical mechanics and quantum information theory, entropies are defined for positive-definite matrices. Subadditivity of the quantum relative entropy function plays an important role in these fields, and this same result is closely connected with subadditivity of the matrix entropy H_φ where $\varphi : t \mapsto t \log t$. As such, subadditivity of other matrix φ-entropies may be relevant for quantum theory.

4.2 Main Results

In this section, we lay out detailed definitions and statements of our main results on subadditivity of matrix φ-entropy and its application to prove concentration inequalities for random matrices.

4.2.1 Notation and Background

Let us instate some notation. The set $\mathbb{R}_+$ contains the nonnegative real numbers, and $\mathbb{R}_{++}$ consists of all positive real numbers. We write $\mathbb{M}^d$ for the complex Banach space of $d \times d$ complex matrices, equipped with the usual ℓ_2 operator norm $\|\cdot\|$. The *normalized trace* is the function

$$\bar{\operatorname{tr}}\, \boldsymbol{B} := \frac{1}{d} \sum\nolimits_{j=1}^{d} b_{jj} \quad \text{for } \boldsymbol{B} \in \mathbb{M}^d.$$

The theory can be developed using the standard trace, but additional complications arise.

The set $\mathbb{H}^d$ refers to the real-linear subspace of $d \times d$ Hermitian matrices in $\mathbb{M}^d$. For a matrix $\boldsymbol{A} \in \mathbb{H}^d$, we write $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ for the algebraic minimum and maximum eigenvalues. For each interval $I \subset \mathbb{R}$, we define the set of Hermitian matrices whose eigenvalues fall in that interval:

$$\mathbb{H}^d(I) := \{\boldsymbol{A} \in \mathbb{H}^d : [\lambda_{\min}(\boldsymbol{A}), \lambda_{\max}(\boldsymbol{A})] \subset I\}.$$

We also introduce the set $\mathbb{H}^d_+$ of $d \times d$ positive-semidefinite matrices and the set $\mathbb{H}^d_{++}$ of $d \times d$ positive-definite matrices. Curly inequalities refer to the positive-semidefinite order. For example, $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

Next, let us explain how to extend scalar functions to matrices. Recall that each Hermitian matrix $\boldsymbol{A} \in \mathbb{H}^d$ has a *spectral resolution*

$$\boldsymbol{A} = \sum\nolimits_{i=1}^{d} \lambda_i \boldsymbol{P}_i, \tag{4.2.1}$$

where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of $\boldsymbol{A}$. The matrices $\boldsymbol{P}_1, \dots, \boldsymbol{P}_d$ are orthogonal projectors that satisfy the orthogonality relations

$$\boldsymbol{P}_i \boldsymbol{P}_j = \delta_{ij} \boldsymbol{P}_j \quad \text{and} \quad \sum\nolimits_{i=1}^{d} \boldsymbol{P}_i = \mathbf{I},$$

where δ_{ij} is the Kronecker delta and $\mathbf{I}$ is the identity matrix. One obtains a standard matrix function by applying a scalar function to the spectrum of a Hermitian matrix.

Definition 4.2.1 (Standard Matrix Function)**.** *Let $f : I \mapsto \mathbb{R}$ be a function on an interval I of the real line. Suppose that $\boldsymbol{A} \in \mathbb{H}^d(I)$ has the spectral decomposition* (4.2.1)*. Then*

$$f(\boldsymbol{A}) := \sum\nolimits_{i=1}^{d} f(\lambda_i)\boldsymbol{P}_i.$$

We use lowercase Roman and Greek letters to refer to standard matrix functions. When we apply a familiar real-valued function to an Hermitian matrix, we are referring to the associated standard matrix function. Bold capital letters such as $\boldsymbol{Y}, \boldsymbol{Z}$ denote general matrix functions that are not necessarily standard.

4.2.2 Subadditivity of Matrix Entropies

In this section, we provide an overview of the theory of matrix φ-entropies. At a high level, our approach has a strong parallel with the work of Boucheron et al. [26]. Nevertheless, there are interesting differences between the scalar and the matrix setting.

4.2.2.1 The Class of Matrix Entropies

First, we carve out a class of standard matrix functions that we can use to construct matrix entropies with the same subadditivity properties as their scalar counterparts.

Definition 4.2.2 (Φ_d Function Class)**.** *Let d be a natural number. The class Φ_d contains each function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ that is either affine or else satisfies the following three conditions.*

1. *φ is convex and continuous at zero.*

2. *φ has two continuous derivatives on $\mathbb{R}_{++}$.*

3. *Define $\psi(t) = \varphi'(t)$ for $t \in \mathbb{R}_{++}$. The derivative $\mathsf{D}\psi$ of the standard matrix function $\psi : \mathbb{H}^d_{++} \to \mathbb{H}^d$ is an invertible linear operator on $\mathbb{H}^d_{++}$, and the map $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is concave with respect to the semidefinite order on operators.*

The technical definitions that support requirement (3) appear in Section 4.3. For now, we just remark that the scalar equivalent of (3) is the statement that $t \mapsto [\varphi''(t)]^{-1}$ is concave on $\mathbb{R}_{++}$.

The class Φ_1 coincides with the Φ function class considered in [26]. It can be shown that $\Phi_{d+1} \subseteq \Phi_d$ for each natural number d, so it is appropriate to introduce the *class of matrix entropies*:

$$\Phi_\infty := \bigcap\nolimits_{d=1}^{\infty} \Phi_d.$$

This class consists of scalar functions that satisfy the conditions of Definition 4.2.2 for an arbitrary choice of dimension d. Note that Φ_∞ is a convex cone: it contains all positive multiples and all finite sums of its elements.

In contrast to the scalar setting, it is quite hard to determine what functions are contained in Φ_∞. The main technical achievement of this paper is to demonstrate that the standard entropy and certain power functions belong to the matrix entropy class.

Theorem 4.2.3 (Elements of the Matrix Entropy Class)**.** *The following functions are members of the Φ_∞ class.*

1. *The standard entropy $t \mapsto t \log t$.*

2. *The power function $t \mapsto t^p$ for each $p \in [1, 2]$.*

The proof of Theorem 4.2.3 appears in Section 4.6. The statement about classical entropy can be obtained from standard results in matrix theory after some argument, but the result for power functions demands new effort. In fact, the claim about the classical entropy follows from the result for power functions because of the representation $t \log t = \lim_{p \downarrow 1} (t^p - t)/(p - 1)$.

See the independent work [92, Sec. 4] for closely related material. Very recently, Hansen and Zhang have developed an elegant characterization of the matrix entropy class [93].

4.2.2.2 Matrix φ-Entropy

For each function in the matrix entropy class, we can introduce a generalized entropy functional that measures the amount of fluctuation in a random matrix.

Definition 4.2.4 (Matrix φ-Entropy)**.** *Let $\varphi \in \Phi_\infty$. Consider a random matrix $\boldsymbol{Z}$ taking values in $\mathbb{H}_+^d$, and assume that $\mathbb{E}\,\|\boldsymbol{Z}\| < \infty$ and $\mathbb{E}\,\|\varphi(\boldsymbol{Z})\| < \infty$. The matrix φ-entropy* *functional H_φ is*

$$H_\varphi(\boldsymbol{Z}) := \mathbb{E}\,\bar{\operatorname{tr}}\,\varphi(\boldsymbol{Z}) - \bar{\operatorname{tr}}\,\varphi(\mathbb{E}\,\boldsymbol{Z}). \tag{4.2.2}$$

Similarly, the conditional matrix φ-entropy functional is

$$H_\varphi(\boldsymbol{Z}\,|\,\mathcal{F}) := \mathbb{E}\left[\,\bar{\operatorname{tr}}\,\varphi(\boldsymbol{Z})\,|\,\mathcal{F}\right] - \bar{\operatorname{tr}}\,\varphi\big(\mathbb{E}[\boldsymbol{Z}\,|\,\mathcal{F}]\big),$$

where $\mathcal{F}$ is a subalgebra of the master sigma algebra.

For each convex function φ, the trace function $\bar{\operatorname{tr}}\,\varphi : \mathbb{H}_+^d \to \mathbb{R}$ is also convex [46, Sec. 2.2]. Therefore, Jensen's inequality implies that the matrix φ-entropy is nonnegative:

$$H_\varphi(\boldsymbol{Z}) \geq 0.$$

For concreteness, here are some basic examples of matrix φ-entropy functionals.

$$\begin{aligned}
&H_\varphi(\boldsymbol{Z}) = \bar{\operatorname{tr}}\left[\mathbb{E}(\boldsymbol{Z}\log\boldsymbol{Z}) - (\mathbb{E}\,\boldsymbol{Z})\log(\mathbb{E}\,\boldsymbol{Z})\right] &&\text{when } \varphi(t) = t\log t.\\
&H_\varphi(\boldsymbol{Z}) = \bar{\operatorname{tr}}\left[\mathbb{E}(\boldsymbol{Z}^p) - (\mathbb{E}\,\boldsymbol{Z})^p\right] &&\text{when } \varphi(t) = t^p \text{ for } p \in [1,2].\\
&H_\varphi(\boldsymbol{Z}) = 0 &&\text{when } \varphi \text{ is affine.}
\end{aligned}$$

4.2.2.3 Subadditivity of Matrix φ-Entropy

The key fact about matrix φ-entropies is that they satisfy a subadditivity property. Let $\boldsymbol{x} := (X_1, \dots, X_n)$ denote a vector of independent random variables taking values in a Polish space, and write $\boldsymbol{x}_{-i}$ for the random vector obtained by deleting the ith entry of $\boldsymbol{x}$.

$$\boldsymbol{x}_{-i} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

Consider a positive-semidefinite random matrix $\boldsymbol{Z}$ that can be expressed as a measurable function of the random vector $\boldsymbol{x}$.

$$\boldsymbol{Z} := \boldsymbol{Z}(X_1, \dots, X_n) \in \mathbb{H}_+^d.$$

We instate the integrability conditions $\mathbb{E}\,\|\boldsymbol{Z}\| < \infty$ and $\mathbb{E}\,\|\varphi(\boldsymbol{Z})\| < \infty$.

Theorem 4.2.5 (Subadditivity of Matrix φ-Entropy)**.** *Fix a function $\varphi \in \Phi_\infty$. Under the* *prevailing assumptions,*

$$H_\varphi(\boldsymbol{Z}) \leq \sum\nolimits_{i=1}^n \mathbb{E}\left[H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i})\right]. \tag{4.2.3}$$

Typically, we apply Theorem 4.2.5 by way of a corollary. Let $X_1', \dots, X_n'$ denote independent copies of $X_1, \dots, X_n$, and form the random matrix

$$\boldsymbol{Z}_i' := \boldsymbol{Z}(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n) \in \mathbb{H}_+^d.$$

Then $\boldsymbol{Z}_i'$ and $\boldsymbol{Z}$ are independent and identically distributed, conditional on the sigma algebra generated by $\boldsymbol{x}_{-i}$. In particular, these two random matrices are exchangeable counterparts.

Corollary 4.2.6 (Entropy Bounds via Exchangeability)**.** *Fix a function $\varphi \in \Phi_\infty$, and write $\psi = \varphi'$. With the prevailing notation,*

$$H_\varphi(\boldsymbol{Z}) \leq \frac{1}{2} \sum\nolimits_{i=1}^n \mathbb{E}\, \bar{\operatorname{tr}} \left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i'))\right].$$

Theorem 4.2.5 and Corollary 4.2.6 are matrix counterparts of the foundational results from Boucheron et al. [26, Sec. 3], which establish that scalar φ-entropies satisfy a similar subadditivity property. We devote Section 4.4 to the proof of these results.

4.2.3 Some Matrix Concentration Inequalities

Using Corollary 4.2.6, we can derive concentration inequalities for random matrices. In contrast to some previous approaches to matrix concentration, we need to place some significant restrictions on the type of random matrices we consider.

Definition 4.2.7 (Invariance under Signed Permutation)**.** *A random matrix $\boldsymbol{Y} \in \mathbb{H}^d$ is invariant under signed permutation if we have the equality of distribution*

$$\boldsymbol{Y} \sim \boldsymbol{\Pi}^* \boldsymbol{Y} \boldsymbol{\Pi} \quad \textit{for each signed permutation } \boldsymbol{\Pi}.$$

A signed permutation $\boldsymbol{\Pi} \in \mathbb{M}^d$ *is a matrix with the properties that (i) each row and each column contains exactly one nonzero entry and (ii) the nonzero entries only take values $+1$ and -1.*

In particular, consider a random matrix that is invariant under orthogonal conjugation:

$$\boldsymbol{Y} \sim \boldsymbol{U}^* \boldsymbol{Y} \boldsymbol{U} \quad \text{for each orthogonal matrix } \boldsymbol{U}.$$

A matrix that satisfies this condition always verifies the requirement of Definition 4.2.7. Many classical ensembles, such as the GOE, satisfy this orthogonal invariance condition. Similar remarks apply to random matrices that are invariant under unitary conjugation.

4.2.3.1 A Bounded Difference Inequality

Let us present an exponential tail bound for a random matrix whose distribution is invariant under signed permutation.

Theorem 4.2.8 (Bounded Differences)**.** *Let $\boldsymbol{x} := (X_1, \dots, X_n)$ be a vector of independent random variables, and let $\boldsymbol{x}' := (X_1', \dots, X_n')$ be an independent copy of $\boldsymbol{x}$. Consider random matrices*

$$\begin{aligned} \boldsymbol{Y} &:= \boldsymbol{Y}(X_1, \dots, X_i, \dots, X_n) \in \mathbb{H}^d \quad \textit{and} \\ \boldsymbol{Y}_i' &:= \boldsymbol{Y}(X_1, \dots, X_i', \dots, X_n) \in \mathbb{H}^d \quad \textit{for } i = 1, \dots, n. \end{aligned}$$

Assume that $\boldsymbol{Y}$ is invariant under signed permutation and that $\|\boldsymbol{Y}\|$ is bounded almost surely. Introduce the variance measure

$$V_{\boldsymbol{Y}} := \sup \left\| \mathbb{E} \left[\sum\nolimits_{i=1}^n (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,\middle|\, \boldsymbol{x} \right] \right\|, \tag{4.2.4}$$

where the supremum occurs over all possible values of $\boldsymbol{x}$. For each $t \geq 0$,

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \geq t\right\} &\leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}, \quad \textit{and} \\ \mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \leq -t\right\} &\leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}. \end{aligned}$$

Theorem 4.2.8 follows from Corollary 4.2.6 with the choice $\varphi(t) = t \log t$. See Section 4.7 for the proof. This result can be viewed as a type of matrix bounded difference inequality. Closely related inequalities already appear in the literature; see [233, Cor. 7.5], [138, Cor. 11.1], and [172, Cor. 4.1]. In fact, Theorem 4.2.8 is dominated by [172, Cor. 4.1], which is not restricted to random matrices that are invariant under signed permutation.

4.2.3.2 Example: Sample covariance matrices

It may be helpful to sketch a short example that indicates the scope of Theorem 4.2.8. Consider a random vector of the form

$$\boldsymbol{w} := (\varepsilon_1 W_1, \varepsilon_2 W_2, \dots, \varepsilon_p W_p)^*,$$

where (W_k) is an exchangeable family of random variables and (ε_k) is a sequence of independent Rademacher random variables. We also require that the random vector is bounded: $\|\boldsymbol{w}\|^2 \leq B$.

Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$ be iid copies of $\boldsymbol{w}$, and consider the sample covariance matrix

$$\boldsymbol{Y} := \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i^*.$$

Our assumptions on $\boldsymbol{w}$ ensure that $\boldsymbol{Y}$ is invariant under signed permutation and that $\|\boldsymbol{Y}\|$ is bounded. Note that $\mathbb{E}\,\boldsymbol{Y} = c\mathbf{I}$ for a positive number c. It is also easy to check that the variance measure (4.2.4) satisfies $V_{\boldsymbol{Y}} \leq 2B^2/n$. An application of Theorem 4.2.8 delivers

$$\mathbb{P}\left\{\|\boldsymbol{Y} - c\mathbf{I}\| \geq t\right\} \leq 2d \cdot \mathrm{e}^{-nt^2/(4B^2)}.$$

The bound is informative when $c^2 > t^2 > 4B^2 \log(2d)/n$. In other words, the number n of samples should satisfy $n > 4B^2 \log(2d)/c^2$. Modulo constants, this estimate cannot be improved when $\boldsymbol{w}$ has the uniform distribution on $\{\pm\mathbf{e}_1, \dots, \pm\mathbf{e}_p\}$, the set of signed standard basis vectors.

The main result of Rudelson's paper [193] is a concentration bound for sample covariance matrices based on the noncommutative Khintchine inequality [136]. Rudelson allows any bounded random vector $\boldsymbol{w}$ with a scalar covariance matrix, and he achieves the same result derived here.

4.2.3.3 Matrix Moment Bounds

We can also establish moment inequalities for a random matrix whose distribution is invariant under signed permutation.

Theorem 4.2.9 (Matrix Moment Bound)**.** *Fix a number $q \in \{2, 3, 4, \dots\}$. Let $\boldsymbol{x} := (X_1, \dots, X_n)$ be a vector of independent random variables, and let $\boldsymbol{x}' := (X_1', \dots, X_n')$ be an independent copy of $\boldsymbol{x}$. Consider positive-semidefinite random matrices*

$$\begin{aligned} \boldsymbol{Y} &:= \boldsymbol{Y}(X_1, \dots, X_i, \dots, X_n) \in \mathbb{H}_+^d \quad \textit{and} \\ \boldsymbol{Y}_i' &:= \boldsymbol{Y}(X_1, \dots, X_i', \dots, X_n) \in \mathbb{H}_+^d \quad \textit{for } i = 1, \dots, n. \end{aligned}$$

Assume that $\boldsymbol{Y}$ is invariant under signed permutation and that $\mathbb{E}(\|\boldsymbol{Y}\|^q) < \infty$. Suppose that there is a constant $c \geq 0$ with the property

$$\boldsymbol{V_Y} := \mathbb{E}\left[\sum\nolimits_{i=1}^n (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,\middle|\, \boldsymbol{x}\right] \preccurlyeq c\,\boldsymbol{Y}. \tag{4.2.5}$$

Then the random matrix $\boldsymbol{Y}$ satisfies the moment inequality

$$[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^q)]^{1/q} \leq \mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{Y} + \frac{q-1}{2} \cdot c.$$

Theorem 4.2.9 follows from Corollary 4.2.6 with the choice $\varphi(t) = t^{q/(q-1)}$. See Section 4.8 for the proof. This result can be regarded as a matrix extension of a moment inequality for real random variables [26, Cor. 1]. The paper [172] contains similar moment inequalities for random matrices that need not satisfy the condition of signed permutation invariance. See also [110, 112, 113].

4.2.4 Generalized Subadditivity of Matrix φ-Entropy

Theorem 4.2.5 is the shadow of a more sophisticated subadditivity property. We outline the simplest form of this more general result. See the lecture notes of Carlen [46] for more background on the topics in this section.

We work in the $*$-algebra $\mathbb{M}^d$ of $d \times d$ complex matrices, equipped with the conjugate transpose operation $*$ and the normalized trace inner product $\langle \boldsymbol{A},\, \boldsymbol{B} \rangle := \bar{\operatorname{tr}}(\boldsymbol{A}^*\boldsymbol{B})$. We say that a subspace $\mathfrak{A} \subset \mathbb{M}^d$ is a $*$-*subalgebra* when $\mathfrak{A}$ contains the identity matrix, $\mathfrak{A}$ is closed under matrix multiplication, and $\mathfrak{A}$ is closed under conjugate transposition. In other terms, $\mathbf{I} \in \mathfrak{A}$ and $\boldsymbol{AB} \in \mathfrak{A}$ and $\boldsymbol{A}^* \in \mathfrak{A}$ whenever $\boldsymbol{A}, \boldsymbol{B} \in \mathfrak{A}$.

In this setting, there is an elegant notion of conditional expectation. The orthogonal projector $\mathbb{E}_{\mathfrak{A}} : \mathbb{M}^d \to \mathfrak{A}$ onto the $*$-subalgebra $\mathfrak{A}$ is called the *conditional expectation* with respect to the $*$-subalgebra. For $*$-subalgebras $\mathfrak{A}$ and $\mathfrak{B}$, we say that the conditional expectations $\mathbb{E}_{\mathfrak{A}}$ and $\mathbb{E}_{\mathfrak{B}}$ *commute* when

$$(\mathbb{E}_{\mathfrak{A}}\, \mathbb{E}_{\mathfrak{B}})(\boldsymbol{M}) = (\mathbb{E}_{\mathfrak{B}}\, \mathbb{E}_{\mathfrak{A}})(\boldsymbol{M}) \quad \text{for every } \boldsymbol{M} \in \mathbb{M}^d.$$

This construction generalizes the concept of independence in a probability space.

We can define the matrix φ-entropy conditional on a $*$-subalgebra $\mathfrak{A}$:

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}) := \bar{\operatorname{tr}}[\varphi(\boldsymbol{A}) - \varphi(\mathbb{E}_{\mathfrak{A}}\, \boldsymbol{A})] \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d.$$

Note that $\bar{\operatorname{tr}}(\mathbb{E}_{\mathfrak{A}}\, \boldsymbol{A}) = \bar{\operatorname{tr}}\, \boldsymbol{A}$ for each matrix $\boldsymbol{A}$ in $\mathbb{H}_+^d$, so we do not need to include a conditional expectation in the leading term. Let $\mathfrak{A}_1, \dots, \mathfrak{A}_n$ be $*$-subalgebras whose conditional expectations commute. Then we can extend the definition of the matrix φ-entropy to read

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_1, \dots, \mathfrak{A}_n) := \bar{\operatorname{tr}}[\varphi(\boldsymbol{A}) - \varphi(\mathbb{E}_{\mathfrak{A}_1} \cdots \mathbb{E}_{\mathfrak{A}_n}\, \boldsymbol{A})] \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d.$$

Because of commutativity, the order of the conditional expectations has no effect on the calculation. It turns out that matrix φ-entropy admits the following subadditivity property.

Theorem 4.2.10 (Subaddivity of Matrix φ-Entropy II)**.** *Fix a function $\varphi \in \Phi_\infty$. Let $\mathfrak{A}_1, \dots, \mathfrak{A}_n$ be $*$-subalgebras of $\mathbb{M}^d$ whose conditional expectations commute. Then*

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_1, \dots, \mathfrak{A}_n) \le \sum\nolimits_{i=1}^n H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_i) \quad \textit{for } \boldsymbol{A} \in \mathbb{H}_+^d. \tag{4.2.6}$$

We omit the proof of this result. The argument involves considerations similar with Theorem 4.2.5, but it requires an extra dose of operator theory. The work in this paper already addresses the more challenging aspects of the proof. Note that the case

$\varphi : t \mapsto t \log t$ is essentially a consequence of the classical results in [133].

Theorem 4.2.10 can be seen as a formal extension of the subadditivity of matrix φ-entropy expressed in Theorem 4.2.5. To see why, let $\Omega := \Omega_1 \times \cdots \times \Omega_n$ be a product probability space. The space $L_2(\Omega; \mathbb{M}^d)$ of random matrices is a $*$-algebra with the normalized trace functional $\mathbb{E}\,\bar{\mathrm{tr}}$. For each $i = 1, \dots, n$, we can form a $*$-subalgebra $\mathfrak{A}_i$ consisting of the random matrices that do not depend on the ith factor Ω_i of the product. The conditional expectation $\mathbb{E}_{\mathfrak{A}_i}$ simply integrates out the ith random variable. By independence, the family of conditional expectations $\mathbb{E}_{\mathfrak{A}_1}, \dots, \mathbb{E}_{\mathfrak{A}_n}$ commutes. Using this dictionary, compare the statement of Theorem 4.2.10 with Theorem 4.2.5.

4.3 Operators and Functions acting on Matrices

This work involves a substantial amount of operator theory. This section contains a short treatment of the basic facts. See [17, 18] for a more complete introduction.

4.3.1 Linear Operators on Matrices

Let $\mathbb{C}^d$ be the complex Hilbert space of dimension d, equipped with the standard inner product $\langle \boldsymbol{a},\ \boldsymbol{b} \rangle := \boldsymbol{a}^* \boldsymbol{b}$. We usually identify $\mathbb{M}^d$ with $\mathbb{B}(\mathbb{C}^d)$, the complex Banach space of linear operators acting on $\mathbb{C}^d$, equipped with the ℓ_2 operator norm $\|\cdot\|$.

We can also endow $\mathbb{M}^d$ with the normalized trace inner product $\langle \boldsymbol{A},\ \boldsymbol{B} \rangle := \bar{\mathrm{tr}}(\boldsymbol{A}^* \boldsymbol{B})$ to form a Hilbert space. As a Hilbert space, $\mathbb{M}^d$ is isometrically isomorphic with $\mathbb{C}^{d^2}$. Let $\mathbb{B}(\mathbb{M}^d)$ denote the complex Banach space of linear operators that map the Hilbert space $\mathbb{M}^d$ into itself, equipped with the induced operator norm. The Banach space $\mathbb{B}(\mathbb{M}^d)$ is isometrically isomorphic with the Banach space $\mathbb{M}^{d^2}$.

As a consequence of this construction, every concept from matrix analysis has an immediate analog for linear operators on matrices. An operator $\mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$ is *self-adjoint* when

$$\langle \boldsymbol{A},\ \mathsf{T}(\boldsymbol{B}) \rangle = \langle \mathsf{T}(\boldsymbol{A}),\ \boldsymbol{B} \rangle \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{B}(\mathbb{M}^d).$$

A self-adjoint operator $\mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$ is *positive semidefinite* when

$$\langle \boldsymbol{A},\ \mathsf{T}(\boldsymbol{A}) \rangle \geq 0 \quad \text{for all } \boldsymbol{A} \in \mathbb{M}^d.$$

For self-adjoint operators $\mathsf{S}, \mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$, the notation $\mathsf{S} \preccurlyeq \mathsf{T}$ means that $\mathsf{T} - \mathsf{S}$ is positive semidefinite.

Each self-adjoint matrix operator $\mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$ has a spectral resolution of the form

$$\mathsf{T} = \sum\nolimits_{i=1}^{d^2} \lambda_i \mathsf{P}_i, \tag{4.3.1}$$

where $\lambda_1, \dots, \lambda_{d^2}$ are the eigenvalues of T and the spectral projectors $\mathsf{P}_1, \dots, \mathsf{P}_{d^2}$ are positive-semidefinite operators that satisfy

$$\mathsf{P}_i \mathsf{P}_j = \delta_{ij} \mathsf{P}_j \quad \text{and} \quad \sum\nolimits_{i=1}^{d^2} \mathsf{P}_i = \mathsf{I},$$

where δ_{ij} is the Kronecker delta and I is the identity operator. As in the matrix case, a self-adjoint operator with nonnegative eigenvalues is the same thing as a positive-semidefinite operator.

We can extend a scalar function $f : I \to \mathbb{R}$ on an interval I of the real line to obtain a standard operator function. Indeed, if T has the spectral resolution (4.3.1) and the eigenvalues of T fall in the interval I, we define

$$f(\mathsf{T}) := \sum\nolimits_{i=1}^{d^2} f(\lambda_i) \mathsf{P}_i.$$

This definition, of course, parallels the definition for matrices.

4.3.2 Monotonicity and Convexity

Let X and Y be sets of self-adjoint operators, such as $\mathbb{H}^d(I)$ or the set of self-adjoint operators in $\mathbb{B}(\mathbb{M}^d)$. We can introduce notions of monotonicity and convexity for a general function $\Psi : X \to Y$ using the semidefinite order on the spaces of operators.

Definition 4.3.1 (Monotone Operator-Valued Function)**.** *The function* $\Psi : X \to Y$ *is* monotone *when*

$$\mathsf{S} \preccurlyeq \mathsf{T} \quad \Longrightarrow \quad \Psi(\mathsf{S}) \preccurlyeq \Psi(\mathsf{T}) \quad \textit{for all } \mathsf{S}, \mathsf{T} \in X.$$

Definition 4.3.2 (Convex Operator-Valued Function)**.** *The function* $\Psi : X \to Y$ *is* convex *when X is a convex set and*

$$\mathsf{\Psi}(\alpha\mathsf{S} + \bar{\alpha}\mathsf{T}) \preccurlyeq \alpha \cdot \mathsf{\Psi}(\mathsf{S}) + \bar{\alpha} \cdot \mathsf{\Psi}(\mathsf{T}) \quad \textit{for all } \alpha \in [0,1] \textit{ and all } \mathsf{S}, \mathsf{T} \in X.$$

We have written $\bar{\alpha} := 1 - \alpha$. The function $\mathsf{\Psi}$ is concave *when $-\mathsf{\Psi}$ is convex.*

The convexity of an operator-valued function $\mathsf{\Psi}$ is equivalent with a Jensen-type relation:

$$\mathsf{\Psi}(\mathbb{E}\,\mathsf{X}) \preccurlyeq \mathbb{E}\,\mathsf{\Psi}(\mathsf{X}) \tag{4.3.2}$$

whenever X is an integrable random operator taking values in X.

In particular, we can apply these definitions to standard matrix and operator functions. Let I be an interval of the real line. We say that the function $f : I \to \mathbb{R}$ is *operator monotone* when the lifted map $f : \mathbb{H}^d(I) \to \mathbb{H}^d$ is monotone for each natural number d. Likewise, the function $f : I \to \mathbb{R}$ is *operator convex* when the lifted map $f : \mathbb{H}^d(I) \to \mathbb{H}^d$ is convex for each natural number d.

Although scalar monotonicity and convexity are quite common, they are much rarer in the matrix setting [17, Chap. 4]. For present purposes, we note that the power functions $t \mapsto t^p$ with $p \in [0,1]$ are operator monotone and operator concave. The power functions $t \mapsto t^p$ with $p \in [1,2]$ and the standard entropy $t \mapsto t \log t$ are all operator convex.

4.3.3 The Derivative of a Vector-Valued Function

The definition of the Φ_∞ function class involves a requirement that a certain standard matrix function is differentiable. For completeness, we include the background needed to interpret this condition.

Definition 4.3.3 (Derivative of a Vector-Valued Function)**.** *Let X and Y be Banach spaces, and let U be an open subset of X. A function $\boldsymbol{F} : U \to Y$ is* differentiable *at a point $\boldsymbol{A} \in U$ if there exists a bounded linear operator $\mathsf{T} : X \to Y$ for which*

$$\lim_{\boldsymbol{B} \to \boldsymbol{0}} \frac{\|\boldsymbol{F}(\boldsymbol{A} + \boldsymbol{B}) - \boldsymbol{F}(\boldsymbol{A}) - \mathsf{T}(\boldsymbol{B})\|_Y}{\|\boldsymbol{B}\|_X} = 0.$$

When $\boldsymbol{F}$ is differentiable at $\boldsymbol{A}$, the operator T is called the derivative *of $\boldsymbol{F}$ at $\boldsymbol{A}$, and we* *define* $\mathsf{D}\boldsymbol{F}(\boldsymbol{A}) := \mathsf{T}$.

The derivative and the directional derivative have the following relationship:

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{F}(\boldsymbol{A} + s\boldsymbol{B})\Big|_{s=0} = \mathsf{D}\boldsymbol{F}(\boldsymbol{A})(\boldsymbol{B}). \tag{4.3.3}$$

In Section 4.6.2, we present an explicit formula for the derivative of a standard matrix function.

4.4 Subadditivity of Matrix φ-Entropy

In this section, we establish Theorem 4.2.5, which states that the matrix φ-entropy is subadditive for every function in the Φ_∞ class. This result depends on a variational representation for the matrix φ-entropy that appears in Section 4.4.1. We use the variational formula to derive a Jensen-type inequality in Section 4.4.2. The proof of Theorem 4.2.5 appears in Section 4.4.3.

4.4.1 Representation of Matrix φ-Entropy as a Supremum

The fundamental fact behind the subadditivity theorem is a representation of the matrix φ-entropy as a supremum of affine functions.

Lemma 4.4.1 (Supremum Representation for Entropy)**.** *Fix a function $\varphi \in \Phi_\infty$, and introduce the scalar derivative $\psi = \varphi'$. Suppose that $\boldsymbol{Z}$ is a random positive-semidefinite matrix for which $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Then*

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}}\ \mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}) - \psi(\mathbb{E}\,\boldsymbol{T}))(\boldsymbol{Z} - \boldsymbol{T}) + \varphi(\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T})\right]. \tag{4.4.1}$$

The range of the supremum contains each random positive-definite matrix $\boldsymbol{T}$ for which $\|\boldsymbol{T}\|$ and $\|\varphi(\boldsymbol{T})\|$ are integrable. In particular, the matrix φ-entropy H_φ can be written in the dual form

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}}\ \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{\Upsilon}_1(\boldsymbol{T}) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\right], \tag{4.4.2}$$

where $\boldsymbol{\Upsilon}_i : \mathbb{H}^d_+ \to \mathbb{H}^d$ for $i = 1, 2$.

This result implies that H_φ is a convex function on the space of random positive-semidefinite matrices. The dual representation of H_φ is well suited for establishing a form of Jensen's inequality, Lemma 4.4.3, which is the main ingredient in the proof of the subadditivity property, Theorem 4.2.5.

It may be valuable to see some particular instances of the dual representation of the matrix φ-entropy:

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}} \ \mathbb{E}\,\bar{\operatorname{tr}}\left[(\log \boldsymbol{T} - \log(\mathbb{E}\,\boldsymbol{T}))\cdot \boldsymbol{Z}\right] \qquad \text{when } \varphi(t) = t\log t.$$

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}} \ \mathbb{E}\,\bar{\operatorname{tr}}\left[p(\boldsymbol{T}^{p-1} - (\mathbb{E}\,\boldsymbol{T})^{p-1})\cdot \boldsymbol{Z} - (p-1)(\boldsymbol{T}^p - (\mathbb{E}\,\boldsymbol{T})^p)\right] \quad \text{when } \varphi(t) = t^p \text{ for } p \in [1,2].$$

The first formula is the matrix version of a well-known variational principle for the classical entropy, cf. [26, p. 525]. In the matrix setting, this result can be derived from the joint convexity of quantum relative entropy [133].

4.4.1.1 The Convexity Lemma

To establish the variational formula, we require a convexity result for a quadratic form connected with the function φ.

Lemma 4.4.2. *Fix a function $\varphi \in \Phi_\infty$, and let $\psi = \varphi'$. Suppose that $\boldsymbol{Y}$ is a random matrix taking values in $\mathbb{H}^d_+$, and let $\boldsymbol{K}$ be a random matrix taking values in $\mathbb{M}^d$. Assume that $\|\boldsymbol{Y}\|$ and $\|\boldsymbol{K}\|$ are integrable. Then*

$$\mathbb{E}\left\langle \boldsymbol{K},\ \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\right\rangle \geq \left\langle (\mathbb{E}\,\boldsymbol{K}),\ \mathsf{D}\psi(\mathbb{E}\,\boldsymbol{Y})(\mathbb{E}\,\boldsymbol{K})\right\rangle.$$

Proof. The proof hinges on a basic convexity property of quadratic forms. Define a map that takes a matrix $\boldsymbol{A}$ in $\mathbb{H}^d$ and a positive-definite operator T on $\mathbb{M}^d$ to a nonnegative number:

$$\mathcal{Q} : (\boldsymbol{A}, \mathsf{T}) \mapsto \left\langle \boldsymbol{A},\ \mathsf{T}^{-1}(\boldsymbol{A})\right\rangle.$$

We assert that the function $\mathcal{Q}$ is convex. Indeed, the same result is well known when $\boldsymbol{A}$ and T are replaced by a vector and a positive-definite matrix [18, Exer. 1.5.1], and the extension is immediate from the isometric isomorphism between operators and matrices.

Recall that the Φ_∞ class requires $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ to be a concave map on $\mathbb{H}^d_{++}$. With these observations at hand, we can make the following calculation:

$$\begin{aligned}\mathbb{E}\left\langle \boldsymbol{K},\ \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\right\rangle &= \mathbb{E}\left\langle \boldsymbol{K},\ ([\mathsf{D}\psi(\boldsymbol{Y})]^{-1})^{-1}(\boldsymbol{K})\right\rangle \\ &\geq \left\langle (\mathbb{E}\,\boldsymbol{K}),\ (\mathbb{E}[\mathsf{D}\psi(\boldsymbol{Y})]^{-1})^{-1}(\mathbb{E}\,\boldsymbol{K})\right\rangle \\ &\geq \left\langle (\mathbb{E}\,\boldsymbol{K}),\ ([\mathsf{D}\psi(\mathbb{E}\,\boldsymbol{Y})]^{-1})^{-1}(\mathbb{E}\,\boldsymbol{K})\right\rangle \\ &= \left\langle (\mathbb{E}\,\boldsymbol{K}),\ \mathsf{D}\psi(\mathbb{E}\,\boldsymbol{Y})(\mathbb{E}\,\boldsymbol{K})\right\rangle .\end{aligned}$$

We obtain the second relation when we apply Jensen's inequality to the convex function $\mathcal{Q}$. The third relation depends on the semidefinite Jensen inequality (4.3.2) for the concave function $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$, coupled with the fact [17, Prop. V.1.6] that the operator inverse reverses the semidefinite order. □

4.4.1.2 Proof of Lemma 4.4.1

The argument parallels the proof of [26, Lem. 1]. We begin with some reductions. The case where φ is an affine function is immediate, so we may require the derivative $\psi = \varphi'$ to be non-constant. By approximation, we may also assume that the random matrix $\boldsymbol{Z}$ is strictly positive definite.

[Indeed, since φ is continuous on $\mathbb{R}_+$, the Dominated Convergence Theorem implies that the matrix φ-entropy H_φ is continuous on the set containing each positive-semidefinite random matrix $\boldsymbol{Y}$ where $\|\boldsymbol{Y}\|$ and $\|\varphi(\boldsymbol{Y})\|$ are integrable. Therefore, we can approximate a positive-semidefinite random matrix $\boldsymbol{Z}$ by a sequence $\{\boldsymbol{Y}_n\}$ of positive-definite random matrices where $\boldsymbol{Y}_n \to \boldsymbol{Z}$ and be confident that $H_\varphi(\boldsymbol{Y}_n) \to H_\varphi(\boldsymbol{Z})$.]

When $\boldsymbol{T} = \boldsymbol{Z}$, the argument of the supremum in (4.4.1) equals $H_\varphi(\boldsymbol{Z})$. Therefore, our burden is to verify the inequality

$$H_\varphi(\boldsymbol{Z}) \geq \mathbb{E}\,\bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}) - \psi(\mathbb{E}\,\boldsymbol{T}))(\boldsymbol{Z} - \boldsymbol{T}) + \mathbb{E}\,\varphi(\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T})\right] \tag{4.4.3}$$

for each random positive-definite matrix $\boldsymbol{T}$ that satisfies the same integrability requirements as $\boldsymbol{Z}$. For simplicity, we assume that the eigenvalues of both $\boldsymbol{Z}$ and $\boldsymbol{T}$ are bounded and bounded away from zero. See Appendix 4.9 for the extension to the general case.

We use an interpolation argument to establish (4.4.3). Define the family of random matrices

$$\boldsymbol{T}_s := (1-s)\cdot\boldsymbol{Z} + s\cdot\boldsymbol{T} \quad \text{for } s\in[0,1].$$

Introduce the real-valued function

$$F(s) := \mathbb{E}\,\bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}_s)-\psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{Z}-\boldsymbol{T}_s)\right] + H_\varphi(\boldsymbol{T}_s).$$

Observe that $F(0) = H_\varphi(\boldsymbol{Z})$, while $F(1)$ coincides with the right-hand side of (4.4.3). Therefore, to establish (4.4.3), it suffices to show that the function $F(s)$ is weakly decreasing on the interval $[0,1]$.

We intend to prove that $F'(s)\le 0$ for $s\in[0,1]$. Since $\boldsymbol{Z}-\boldsymbol{T}_s = -s\cdot(\boldsymbol{T}-\boldsymbol{Z})$, we can rewrite the function F in the form

$$F(s) = -s\cdot\mathbb{E}\,\bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}_s)-\psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + \mathbb{E}\,\bar{\operatorname{tr}}\left[\varphi(\boldsymbol{T}_s)-\varphi(\mathbb{E}\,\boldsymbol{T}_s))\right]. \tag{4.4.4}$$

We differentiate the function F to obtain

$$\begin{aligned} F'(s) = -s\cdot\mathbb{E}\,\bar{\operatorname{tr}}\left[\mathsf{D}\psi(\boldsymbol{T}_s)(\boldsymbol{T}-\boldsymbol{Z})\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + s\cdot\bar{\operatorname{tr}}\left[\mathsf{D}\psi(\mathbb{E}\,\boldsymbol{T}_s)(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z}))\cdot(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z}))\right] \\ - \mathbb{E}\,\bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}_s)-\psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + \mathbb{E}\,\bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}_s)-\psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right]. \end{aligned} \tag{4.4.5}$$

To handle the first term in (4.4.4), we applied the product rule, the rule (4.3.3) for directional derivatives, and the expression $\mathrm{d}\boldsymbol{T}_s/\mathrm{d}s = \boldsymbol{T}-\boldsymbol{Z}$. We used the identity $\mathsf{D}\operatorname{tr}\varphi(\boldsymbol{A}) = \psi(\boldsymbol{A})$ to differentiate the second term. We also relied on the Dominated Convergence Theorem to pass derivatives through expectations, which is justified because φ and ψ are continuously differentiable on $\mathbb{H}^d_{++}$ and the eigenvalues of the random matrices are bounded and bounded away from zero. Now, the last two terms in (4.4.5) cancel, and we can rewrite the first two terms using the trace inner product:

$$F'(s) = s\cdot\left[\,\langle(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z})),\ \mathsf{D}\psi(\mathbb{E}\,\boldsymbol{T}_s)(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z}))\rangle - \mathbb{E}\,\langle(\boldsymbol{T}-\boldsymbol{Z}),\ \mathsf{D}\psi(\boldsymbol{T}_s)(\boldsymbol{T}-\boldsymbol{Z})\rangle\,\right].$$

Invoke Lemma 4.4.2 to conclude that $F'(s)\le 0$ for $s\in[0,1]$.

4.4.2 A Conditional Jensen Inequality

The variational inequality in Lemma 4.4.1 leads directly to a Jensen inequality for the matrix φ-entropy.

Lemma 4.4.3 (Conditional Jensen Inequality)**.** *Fix a function $\varphi \in \Phi_\infty$. Suppose that (X_1, X_2) is a pair of independent random variables taking values in a Polish space, and let $\boldsymbol{Z} = \boldsymbol{Z}(X_1, X_2)$ be a random positive-semidefinite matrix for which $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Then*

$$H_\varphi\left(\mathbb{E}_1 \boldsymbol{Z}\right) \le \mathbb{E}_1 H_\varphi\left(\boldsymbol{Z} \,|\, X_1\right),$$

where $\mathbb{E}_1$ is the expectation with respect to the first variable X_1.

Proof. Let $\mathbb{E}_2$ denote the expectation with respect to the second variable X_2. The result is a simple consequence of the dual representation (4.10.2) of the matrix φ-entropy:

$$H_\varphi\left(\mathbb{E}_1 \boldsymbol{Z}\right) = \sup_{\boldsymbol{T}} \; \mathbb{E}_2 \, \bar{\mathrm{tr}}\left[\boldsymbol{\Upsilon}_1\big(\boldsymbol{T}(X_2)\big) \cdot \left(\mathbb{E}_1 \boldsymbol{Z}\right) + \boldsymbol{\Upsilon}_2\big(\boldsymbol{T}(X_2)\big)\right]. \tag{4.4.6}$$

We have written $\boldsymbol{T}(X_2)$ to emphasize that this matrix depends only on the randomness in X_2. To control (4.4.6), we apply Fubini's theorem to interchange the order of $\mathbb{E}_1$ and $\mathbb{E}_2$, and then we exploit the convexity of the supremum to draw out the expectation $\mathbb{E}_1$.

$$\begin{aligned}
H_\varphi\left(\mathbb{E}_1 \boldsymbol{Z}\right) &= \sup_{\boldsymbol{T}} \; \mathbb{E}_1 \, \mathbb{E}_2 \, \bar{\mathrm{tr}}\left[\boldsymbol{\Upsilon}_1(\boldsymbol{T}(X_2)) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T}(X_2))\right] \\
&\le \mathbb{E}_1 \sup_{\boldsymbol{T}} \; \mathbb{E}_2 \, \bar{\mathrm{tr}}\left[\boldsymbol{\Upsilon}_1(\boldsymbol{T}(X_2)) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T}(X_2)\right] \\
&= \mathbb{E}_1 \sup_{\boldsymbol{T}} \; \mathbb{E}\left[\, \bar{\mathrm{tr}}[\boldsymbol{\Upsilon}_1(\boldsymbol{T}(X_2)) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T}(X_2)] \,|\, X_1\right] \\
&= \mathbb{E}_1 \, H_\varphi(\boldsymbol{Z} \,|\, X_1).
\end{aligned}$$

The last relation is the duality formula (4.10.2), applied conditionally. □

4.4.3 Proof of Theorem 4.2.5

We are now prepared to establish the main result on subadditivity of matrix φ-entropy. This theorem is a direct consequence of the conditional Jensen inequality, Lemma 4.4.3. In this argument, we write $\mathbb{E}_i$ for the expectation with respect to the variable X_i. Using the notation from Section 4.2.2.3, we see that $\mathbb{E}_i = \mathbb{E}[\,\cdot\,|\,\boldsymbol{x}_{-i}]$.

First, separate the matrix φ-entropy into two parts by adding and subtracting terms:

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\, \bar{\operatorname{tr}}\, [\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}_1 \boldsymbol{Z}) + \varphi(\mathbb{E}_1 \boldsymbol{Z}) - \varphi(\mathbb{E}\, \boldsymbol{Z})]. \\ &= \mathbb{E}\left[\mathbb{E}_1\, \bar{\operatorname{tr}}\, [\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}_1 \boldsymbol{Z})]\right] + \mathbb{E}\, \bar{\operatorname{tr}}\, [\varphi(\mathbb{E}_1 \boldsymbol{Z}) - \varphi(\mathbb{E}\, \mathbb{E}_1 \boldsymbol{Z})]. \end{aligned} \tag{4.4.7}$$

We can rewrite this expression as

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-1}) + H_\varphi(\mathbb{E}_1 \boldsymbol{Z}) \\ &\le \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-1}) + \mathbb{E}_1\, H_\varphi(\boldsymbol{Z} \,|\, X_1). \end{aligned} \tag{4.4.8}$$

The inequality follows from Lemma 4.4.3 because $\boldsymbol{Z} = \boldsymbol{Z}(X_1, \boldsymbol{x}_{-1})$ where X_1 and $\boldsymbol{x}_{-1}$ are independent random variables.

The first term on the right-hand side of (4.4.8) coincides with the first summand on the right-hand side of the subadditivity inequality (4.2.3). We must argue that the remaining summands are contained in the second term on the right-hand side of (4.4.8). Repeating the argument in the previous paragraph, conditioning on X_1, we obtain

$$H_\varphi(\boldsymbol{Z} \,|\, X_1) \le \mathbb{E}\left[H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-2}) \,|\, X_1 \right] + \mathbb{E}_2\, H_\varphi(\boldsymbol{Z} \,|\, X_1, X_2).$$

Substituting this expression into (4.4.8), we obtain

$$H_\varphi(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^{2} \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}) + \mathbb{E}_1\, \mathbb{E}_2\, H_\varphi(\boldsymbol{Z} \,|\, X_1, X_2).$$

Continuing in this fashion, we arrive at the subadditivity inequality (4.2.3):

$$H_\varphi(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^{n} \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}).$$

This completes the proof of Theorem 4.2.5.

4.5 Entropy Bounds via Exchangeability

In this section, we derive Corollary 4.2.6, which uses exchangeable pairs to bound the conditional entropies that appear in Theorem 4.2.5. This result follows from another variational representation of the matrix φ-entropy.

4.5.1 Representation of the Matrix φ-Entropy as an Infimum

In this section, we present another formula for the matrix φ-entropy.

Lemma 4.5.1 (Infimum Representation for Entropy)**.** *Fix a function $\varphi \in \Phi_\infty$, and let $\psi = \varphi'$. Assume that $\boldsymbol{Z}$ is a random positive-semidefinite matrix where $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Then*

$$H_\varphi(\boldsymbol{Z}) = \inf_{\boldsymbol{A} \in \mathbb{H}_+^d} \mathbb{E}\,\bar{\operatorname{tr}}\,[\varphi(\boldsymbol{Z}) - \varphi(\boldsymbol{A}) - (\boldsymbol{Z} - \boldsymbol{A}) \cdot \psi(\boldsymbol{A})]\,. \tag{4.5.1}$$

Let $\boldsymbol{Z}'$ be an independent copy of $\boldsymbol{Z}$. Then

$$H_\varphi(\boldsymbol{Z}) \le \frac{1}{2} \cdot \mathbb{E}\,\bar{\operatorname{tr}}\left[(\boldsymbol{Z} - \boldsymbol{Z}')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}'))\right]. \tag{4.5.2}$$

We require a familiar trace inequality [46, Thm. 2.11]. This bound simply restates the fact that a convex function lies above its tangents.

Proposition 4.5.2 (Klein's Inequality)**.** *Let $f : I \to \mathbb{R}$ be a differentiable convex function on an interval I of the real line. Then*

$$\bar{\operatorname{tr}}\left[f(\boldsymbol{B}) - f(\boldsymbol{A}) - (\boldsymbol{B} - \boldsymbol{A}) \cdot f'(\boldsymbol{A})\right] \ge 0 \quad \textit{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d(I).$$

With Klein's inequality at hand, the variational inequality follows quickly.

Proof of Lemma 4.5.1. Every function $\varphi \in \Phi_\infty$ is convex and differentiable, so Proposition 4.5.2 with $\boldsymbol{B} = \mathbb{E}\,\boldsymbol{Z}$ implies that

$$\bar{\operatorname{tr}}\,[-\varphi(\mathbb{E}\,\boldsymbol{Z})] \le \bar{\operatorname{tr}}\,[-\varphi(\boldsymbol{A}) - (\mathbb{E}\,\boldsymbol{Z} - \boldsymbol{A}) \cdot \psi(\boldsymbol{A})]$$

for each fixed matrix $\boldsymbol{A} \in \mathbb{H}_+^d$. Substitute this bound into the definition (4.2.2) of the matrix

φ-entropy, and draw the expectation out of the trace to reach

$$H_{\varphi}(\boldsymbol{Z}) \le \mathbb{E}\,\bar{\operatorname{tr}}\,[\varphi(\boldsymbol{Z}) - \varphi(\boldsymbol{A}) - (\boldsymbol{Z} - \boldsymbol{A}) \cdot \psi(\boldsymbol{A})]. \tag{4.5.3}$$

The inequality (4.5.3) becomes an equality when $\boldsymbol{A} = \mathbb{E}\,\boldsymbol{Z}$, which establishes the variational representation (4.5.1).

The symmetrized bound (4.5.2) follows from an exchangeability argument. Select $\boldsymbol{A} = \boldsymbol{Z}'$ in the expression (4.5.3), and apply the fact that $\mathbb{E}\,\varphi(\boldsymbol{Z}) = \mathbb{E}\,\varphi(\boldsymbol{Z}')$ to obtain

$$H_{\varphi}(\boldsymbol{Z}) \le -\,\mathbb{E}\,\bar{\operatorname{tr}}\,[(\boldsymbol{Z} - \boldsymbol{Z}') \cdot \psi(\boldsymbol{Z}')]. \tag{4.5.4}$$

Since $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ are exchangeable, we can also bound the matrix φ-entropy as

$$H_{\varphi}(\boldsymbol{Z}) \le -\,\mathbb{E}\,\bar{\operatorname{tr}}\,[(\boldsymbol{Z}' - \boldsymbol{Z}) \cdot \psi(\boldsymbol{Z})]. \tag{4.5.5}$$

Take the average of the two bounds (4.5.4) and (4.5.5) to reach the desired inequality (4.5.2). □

In the scalar case, stronger bounds are available. For a function $\varphi \in \Phi_1$,

$$\varphi(b) - \varphi(a) - (b-a)\varphi'(a) \le (b-a)(\varphi'(b) - \varphi'(a)) \quad \text{for all } a, b \ge 0.$$

See [48, Lem. 4.2] for details.

4.5.2 Proof of Corollary 4.2.6

Lemma 4.5.1 leads to a succinct proof of Corollary 4.2.6. We continue to use the notation from Section 4.2.2.3. Apply the inequality (4.5.2) conditionally to control the conditional matrix φ-entropy:

$$H_{\varphi}(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}) \le \frac{1}{2} \cdot \mathbb{E}\,\bar{\operatorname{tr}}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i')) \,|\, \boldsymbol{x}_{-i}\right] \tag{4.5.6}$$

because $\boldsymbol{Z}_i'$ and $\boldsymbol{Z}$ are conditionally iid, given $\boldsymbol{x}_{-i}$. Take the expectation on both sides of (4.5.6), and invoke the tower property of conditional expectation:

$$\mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}) \le \frac{1}{2} \cdot \mathbb{E}\,\bar{\operatorname{tr}} \left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i')) \right]. \tag{4.5.7}$$

To complete the proof, substitute (4.5.7) into the right-hand side of the bound (4.2.3) from the subadditivity result, Theorem 4.2.5.

4.6 Members of the Φ_∞ function class

In this section, we demonstrate that the classical entropy and certain power functions belong to the Φ_∞ function class. The main challenge is to verify that $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is a concave operator-valued map. We establish this result for the classical entropy in Section 4.6.4 and for the power function in Section 4.6.5. See the independent work [92, Sec. 4] for closely related results.

4.6.1 Tensor Product Operators

First, we explain the tensor product construction of an operator. The tensor product will allow us to represent the derivative of a standard matrix function compactly.

Definition 4.6.1 (Tensor Product)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d$. The operator $\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{B}(\mathbb{M}^d)$ is defined by the relation*

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{M}) = \boldsymbol{AMB} \quad \textit{for each } \boldsymbol{M} \in \mathbb{M}^d. \tag{4.6.1}$$

The operator $\boldsymbol{A} \otimes \boldsymbol{B}$ is self-adjoint because we assume the factors are Hermitian matrices.

Suppose that $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d$ are Hermitian matrices with spectral resolutions

$$\boldsymbol{A} = \sum\nolimits_{i=1}^d \lambda_i \boldsymbol{P}_i \quad \text{and} \quad \boldsymbol{B} = \sum\nolimits_{j=1}^d \mu_j \boldsymbol{Q}_j. \tag{4.6.2}$$

Then the tensor product $\boldsymbol{A} \otimes \boldsymbol{B}$ has the spectral resolution

$$\boldsymbol{A} \otimes \boldsymbol{B} = \sum\nolimits_{i,j=1}^d \lambda_i \mu_j \boldsymbol{P}_i \otimes \boldsymbol{Q}_j.$$

In particular, the tensor product of two positive-definite matrices is a positive-definite operator.

4.6.2 The Derivative of a Standard Matrix Function

Next, we present some classical results on the derivative of a standard matrix function. See [17, Sec. V.3] for further details.

Definition 4.6.2 (Divided Difference)**.** *Let $f : I \to \mathbb{R}$ be a continuously differentiable function on an interval I of the real line. The* first divided difference *is the map $f^{[1]} : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$f^{[1]}(\lambda, \mu) := \begin{cases} f'(\lambda), & \lambda = \mu, \\ \frac{f(\lambda)-f(\mu)}{\lambda-\mu}, & \lambda \neq \mu. \end{cases}$$

We also require the Hermite representation of the divided difference:

$$f^{[1]}(\lambda, \mu) = \int_0^1 f'(\tau\lambda + \bar{\tau}\mu)\,\mathrm{d}\tau, \tag{4.6.3}$$

where we have written $\bar{\tau} := 1 - \tau$.

The following result gives an explicit expression for the derivative of a standard matrix function in terms of a divided difference.

Proposition 4.6.3 (Daleckiĭ–Kreĭn Formula)**.** *Let $f : I \to \mathbb{R}$ be a continuously differentiable function of an interval I of the real line. Suppose that $\boldsymbol{A} \in \mathbb{H}^d(I)$ is a diagonal matrix with $\boldsymbol{A} = \operatorname{diag}(a_1, \dots, a_d)$. The derivative $\mathsf{D}f(\boldsymbol{A}) \in \mathbb{B}(\mathbb{M}^d)$, and*

$$\mathsf{D}f(\boldsymbol{A})(\boldsymbol{H}) = f^{[1]}(\boldsymbol{A}) \odot \boldsymbol{H} \quad \text{for } \boldsymbol{H} \in \mathbb{M}^d,$$

where $\odot$ denotes the Schur (i.e., componentwise) product and $f^{[1]}(\boldsymbol{A})$ refers to the matrix of divided differences:

$$\left[f^{[1]}(\boldsymbol{A})\right]_{ij} = f^{[1]}(a_i, a_j) \quad \text{for } i, j = 1, \dots, d.$$

4.6.3 Operator Means

Our approach also relies on the concept of an operator mean. The following definition is due to Kubo & Ando [120].

Definition 4.6.4 (Operator Mean)**.** *Let $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ be an operator concave function that satisfies $f(1) = 1$. Fix a natural number d. Let S and T be positive-definite operators in $\mathbb{B}(\mathbb{M}^d)$. We define the mean of the operators:*

$$\mathsf{M}_f(\mathsf{S}, \mathsf{T}) := \mathsf{T}^{1/2} \cdot f(\mathsf{T}^{-1/2}\mathsf{S}\mathsf{T}^{-1/2}) \cdot \mathsf{T}^{1/2} \in \mathbb{B}(\mathbb{M}^d).$$

When S and T commute, the formula simplifies to

$$\mathsf{M}_f(\mathsf{S}, \mathsf{T}) = \mathsf{T} \cdot f(\mathsf{S}\mathsf{T}^{-1}).$$

A few examples may be helpful. The function $f(s) = (1 + s)/2$ represents the usual arithmetic mean, the function $f(s) = s^{1/2}$ gives the geometric mean, and the function $f(s) = 2s/(1 + s)$ yields the harmonic mean. Operator means have a concavity property, which was established in the paper [120].

Proposition 4.6.5 (Operator Means are Concave)**.** *Let $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ be an operator monotone function with $f(1) = 1$. Fix a natural number d. Suppose that $\mathsf{S}_1, \mathsf{S}_2, \mathsf{T}_1, \mathsf{T}_2$ are positive-definite operators in $\mathbb{B}(\mathbb{M}^d)$. Then*

$$\alpha \cdot \mathsf{M}_f(\mathsf{S}_1, \mathsf{T}_1) + \bar{\alpha} \cdot \mathsf{M}_f(\mathsf{S}_2, \mathsf{T}_2) \preccurlyeq \mathsf{M}_f(\alpha\mathsf{S}_1 + \bar{\alpha}\mathsf{S}_2, \alpha\mathsf{T}_1 + \bar{\alpha}\mathsf{T}_2)$$

for $\alpha \in [0, 1]$ and $\bar{\alpha} = 1 - \alpha$.

4.6.4 Entropy

In this section, we demonstrate that the standard entropy function is a member of the Φ_∞ function class.

Theorem 4.6.6. *The function $\varphi : t \mapsto t \log t - t$ is a member of the Φ_∞ class.*

This result immediately implies Theorem 4.2.3(1), which states that $t \mapsto t \log t$ belongs to Φ_∞. Indeed, the matrix entropy class contains all affine functions and all finite sums of its elements.

Theorem 4.6.6 follows easily from (deep) classical results because the variational representation of the standard entropy from Lemma 4.4.1 is equivalent with the joint convexity of quantum relative entropy [133]. Instead of pursuing this idea, we present an argument that parallels the approach we use to study the power function. Some of the calculations below also appear in [131, Proof of Cor. 2.1], albeit in compressed form.

Proof. Fix a positive integer d. We plan to show that the function $\varphi : t \mapsto t \log t - t$ is a member of the class Φ_d. Evidently, φ is continuous and convex on $\mathbb{R}_+$, and it has two continuous derivatives on $\mathbb{R}_{++}$. It remains to verify the concavity condition for the second derivative.

Write $\psi(t) = \varphi'(t) = \log t$, and let $\boldsymbol{A} \in \mathbb{H}^d_{++}$. Without loss of generality, we may choose a basis where $\boldsymbol{A} = \operatorname{diag}(a_1, \dots, a_d)$. The Daleckiĭ–Kreĭn formula, Proposition 4.6.3, tells us

$$\mathsf{D}\psi(\boldsymbol{A})(\boldsymbol{H}) = \psi^{[1]}(\boldsymbol{A}) \odot \boldsymbol{H} = \left[\psi^{[1]}(a_i, a_j) \cdot h_{ij}\right]_{ij}.$$

As an operator, the derivative acts by Schur multiplication. This formula also makes it clear that the inverse of this operator acts by Schur multiplication:

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \left[\frac{1}{\psi^{[1]}(a_i, a_j)} \cdot h_{ij}\right]_{ij}.$$

Using the Hermite representation (4.6.3) of the first divided difference of $t \mapsto \mathrm{e}^t$, we find

$$\frac{1}{\psi^{[1]}(\mu, \lambda)} = \frac{\lambda - \mu}{\log \lambda - \log \mu} = \int_0^1 \mathrm{e}^{\tau \log \lambda + \bar{\tau} \log \mu} \, \mathrm{d}\tau = \int_0^1 \lambda^\tau \mu^{\bar{\tau}} \, \mathrm{d}\tau.$$

The latter calculation assumes that $\mu \neq \lambda$; it extends to the case $\mu = \lambda$ because both sides of the identity are continuous. As a consequence,

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \int_0^1 \left[a_i^\tau h_{ij} a_j^{\bar{\tau}}\right]_{ij} \mathrm{d}\tau = \int_0^1 \boldsymbol{A}^\tau \boldsymbol{H} \boldsymbol{A}^{\bar{\tau}} \, \mathrm{d}\tau = \int_0^1 (\boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}})(\boldsymbol{H}) \, \mathrm{d}\tau.$$

We discover the expression

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1} = \int_0^1 \boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}} \, \mathrm{d}\tau. \tag{4.6.4}$$

This formula is correct for every positive-definite matrix.

For each $\tau \in [0,1]$, consider the operator monotone function $f : t \mapsto t^\tau$ defined on $\mathbb{R}_+$. Since $f(1) = 1$, we can construct the operator mean M_f associated with the function f. Note that $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{A}$ are commuting positive operators. Thus,

$$\mathsf{M}_f(\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A}) = (\mathbf{I} \otimes \boldsymbol{A}) \cdot f((\boldsymbol{A} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{A})^{-1}) = \boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}}.$$

The map $\boldsymbol{A} \mapsto (\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A})$ is linear, so Proposition 4.6.5 guarantees that $\boldsymbol{A} \mapsto \boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}}$ is concave for each $\tau \in [0,1]$. This result is usually called the Lieb Concavity Theorem [17, Thm. IX.6.1]. Combine this fact with the integral representation (4.6.4) to reach the conclusion that $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is a concave map on the cone $\mathbb{H}^d_{++}$ of positive-definite matrices. □

4.6.5 Power Functions

In this section, we prove that certain power functions belong to the Φ_∞ function class.

Theorem 4.6.7. *For each $p \in [0,1]$, the function $\varphi : t \mapsto t^{p+1}/(p+1)$ is a member of the Φ_∞ class.*

This result immediately implies Theorem 4.2.3(2), which states that $t \mapsto t^{p+1}$ belongs to the class Φ_∞. Indeed, the matrix entropy class contains all positive multiples of its elements.

The proof of Theorem 4.6.7 follows the same path as Theorem 4.6.6, but it is somewhat more involved. First, we derive an expression for the function $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ where $\psi = \varphi'$.

Lemma 4.6.8. *Fix $p \in (0,1]$, and let $\psi(t) = t^p$ for $t \geq 0$. For each matrix $\boldsymbol{A} \in \mathbb{H}^d_+$,*

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1} = \frac{1}{p} \int_0^1 (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{(1-p)/p} \, \mathrm{d}\tau, \tag{4.6.5}$$

where $\bar{\tau} := 1 - \tau$.

Proof. As before, we may assume without loss of generality that the matrix $\boldsymbol{A} = \operatorname{diag}(a_1, \dots, a_d)$.

Using the Daleckiĭ–Kreĭn formula, Proposition 4.6.3, we see that

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \left[\frac{1}{\psi^{[1]}(a_i, a_j)} \cdot h_{ij}\right].$$

The Hermite representation (4.6.3) of the first divided difference of $t \mapsto t^{1/p}$ gives

$$\frac{1}{\psi^{[1]}(\mu, \lambda)} = \frac{\mu - \lambda}{\mu^p - \lambda^p} = \frac{1}{p}\int_0^1 (\tau \cdot \lambda^p + \bar{\tau} \cdot \mu^p)^{(1-p)/p}\, \mathrm{d}\tau =: g(\lambda, \mu).$$

We use continuity to verify that the latter calculation remains valid when $\mu = \lambda$. Using this function g, we can identify a compact representation of the operator:

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \sum\nolimits_{ij} g(a_i, a_j) h_{ij} \mathbf{E}_{ij} = \left[\sum\nolimits_{ij} g(a_i, a_j)(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj})\right](\boldsymbol{H}),$$

where we write $\mathbf{E}_{ij}$ for the matrix with a one in the (i, j) position and zeros elsewhere. It remains to verify that the bracket coincides with the expression (4.6.5). Indeed,

$$\begin{aligned}\sum\nolimits_{ij} g(a_i, a_j)(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj}) &= \frac{1}{p}\int_0^1 \sum\nolimits_{ij} (\tau \cdot a_i^p + \bar{\tau} \cdot a_j^p)^{(1-p)/p} \,(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj})\, \mathrm{d}\tau \\ &= \frac{1}{p}\int_0^1 \left[\sum\nolimits_{ij} (\tau \cdot a_i^p + \bar{\tau} \cdot a_j^p)(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj})\right]^{(1-p)/p} \mathrm{d}\tau \\ &= \frac{1}{p}\int_0^1 (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{(1-p)/p}\, \mathrm{d}\tau.\end{aligned}$$

The second relation follows from the definition of the standard operator function associated with $t \mapsto t^{(1-p)/p}$. To confirm that the third line equals the second, expand the matrices $\boldsymbol{A} = \sum_i a_i \mathbf{E}_{ii}$ and $\mathbf{I} = \sum_j \mathbf{E}_{jj}$ and invoke the bilinearity of the tensor product. □

Proof of Theorem 4.6.7. We are now prepared to prove that certain power functions belong to the Φ_∞ function class. Fix an exponent $p \in [0, 1]$, and let d be a fixed positive integer. We intend to show that the function $\varphi(t) = t^{p+1}/(p + 1)$ belongs to the Φ_d class. When $p = 0$, the function φ is affine, so we may assume that $p > 0$. It is clear that φ is continuous and convex on $\mathbb{R}_+$, and φ has two continuous derivatives on $\mathbb{R}_{++}$. It remains to verify that the second derivative has the required concavity property.

Let $\psi(t) = \varphi'(t) = t^p$ for $t \geq 0$, and consider a matrix $\boldsymbol{A} \in \mathbb{H}_{++}^d$. Lemma 4.6.8

demonstrates that

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1} = \frac{1}{p}\int_0^1 (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{(1/p)(1-p)} \, \mathrm{d}\tau, \tag{4.6.6}$$

where we maintain the usage $\bar{\tau} := 1 - \tau$. For each $\tau \in [0,1]$, the scalar function $a \mapsto \tau a + \bar{\tau}$ is operator monotone because it is affine and increasing. On account of the result [3, Cor. 4.3], the function

$$f : a \mapsto (\tau \cdot a^p + \bar{\tau})^{1/p}$$

is also operator monotone. A short calculation shows that $f(1) = 1$. Therefore, we can use f to construct an operator mean M_f. Since $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{A}$ are commuting positive operators, we have

$$\mathsf{M}_f(\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A}) = (\mathbf{I} \otimes \boldsymbol{A}) \cdot f((\boldsymbol{A} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{A})^{-1}) = \left(\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p\right)^{1/p}.$$

The map $\boldsymbol{A} \mapsto (\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A})$ is linear, so Proposition 4.6.5 ensures that

$$\boldsymbol{A} \mapsto (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{1/p} \tag{4.6.7}$$

is a concave map.

We are now prepared to check that (4.6.6) defines a concave operator. Let $\boldsymbol{S}, \boldsymbol{T}$ be arbitrary positive-definite matrices, and choose $\alpha \in [0,1]$. Suppose that $\boldsymbol{Z}$ is the random matrix that takes value $\boldsymbol{S}$ with probability α and value $\boldsymbol{T}$ with probability $1 - \alpha$. For each $\tau \in [0,1]$, we compute

$$\begin{aligned}\mathbb{E}\left[(\tau \cdot \boldsymbol{Z}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{Z}^p)^{1/p}\right]^{1-p} &\preccurlyeq \left[\mathbb{E}\,(\tau \cdot \boldsymbol{Z}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{Z}^p)^{1/p}\right]^{1-p} \\ &\preccurlyeq \left[\left(\tau \cdot (\mathbb{E}\,\boldsymbol{Z})^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes (\mathbb{E}\,\boldsymbol{Z})^p\right)^{1/p}\right]^{1-p}.\end{aligned}$$

The first relation holds because $t \mapsto t^{1-p}$ is operator concave [17, Thm. V.1.9 and Thm. V.2.5]. To obtain the second relation, we apply the concavity property of the map (4.6.7), followed by the fact that $t \mapsto t^{1-p}$ is operator monotone [17, Thm. V.1.9]. This calculation establishes the claim that

$$\boldsymbol{A} \mapsto \left(\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p\right)^{(1-p)/p}$$

is concave on $\mathbb{H}^d_{++}$ for each $\tau \in [0,1]$. In view of the integral representation (4.6.6), we may conclude that $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is concave on the cone $\mathbb{H}^d_{++}$ of positive-definite matrices. □

4.7 A Bounded Difference Inequality for Random Matrices

In this section, we prove Theorem 4.2.8, a bounded difference inequality for a random matrix whose distribution is invariant under signed permutation. We begin with some preliminaries that support the proof, and we establish the main result in Section 4.7.2.

4.7.1 Preliminaries

First, we describe how to compute the expectation of a function of a random matrix whose distribution is invariant under signed permutation. See Definition 4.2.7 for a reminder of what this requirement means.

Lemma 4.7.1. *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line. Assume that $\boldsymbol{X} \in \mathbb{H}^d(I)$ is a random matrix whose distribution is invariant under signed permutation. Then*

$$\mathbb{E}\, f(\boldsymbol{X}) = \bar{\operatorname{tr}}[\mathbb{E}\, f(\boldsymbol{X})] \cdot \mathbf{I}.$$

Proof. Let $\boldsymbol{\Pi} \in \mathbb{H}^d$ be an arbitrary signed permutation matrix. Observe that

$$\mathbb{E}\, f(\boldsymbol{X}) = \mathbb{E}\, f(\boldsymbol{\Pi}^* \boldsymbol{X} \boldsymbol{\Pi}) = \boldsymbol{\Pi}^*[\mathbb{E}\, f(\boldsymbol{X})]\boldsymbol{\Pi}. \tag{4.7.1}$$

The first relation holds because the distribution of $\boldsymbol{X}$ is invariant under conjugation by $\boldsymbol{\Pi}$. The second relation follows from the definition of a standard matrix function and the fact that $\boldsymbol{\Pi}$ is unitary. We may average (4.7.1) over $\boldsymbol{\Pi}$ drawn from the uniform distribution on the set of signed permutation matrices. A direct calculation shows that the resulting matrix is diagonal, and its diagonal entries are identically equal to $\bar{\operatorname{tr}}[\mathbb{E}\, f(\boldsymbol{X})]$. □

We also require a trace inequality that is related to the mean value theorem. This result specializes [138, Lem. 3.4].

Proposition 4.7.2 (Mean Value Trace Inequality)**.** *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line whose derivative f' is convex. For all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d(I)$,*

$$\bar{\operatorname{tr}}[(\boldsymbol{A}-\boldsymbol{B})(f(\boldsymbol{A})-f(\boldsymbol{B}))] \le \frac{1}{2}\,\bar{\operatorname{tr}}[(\boldsymbol{A}-\boldsymbol{B})^2 \cdot (f'(\boldsymbol{A})+f'(\boldsymbol{B}))].$$

4.7.2 Proof of Theorem 4.2.8

The argument proceeds in three steps. First, we present some elements of the matrix Laplace transform method. Second, we use the subaddivity of matrix φ-entropy to deduce a differential inequality for the trace moment generating function of the random matrix. Finally, we explain how to integrate the differential inequality to obtain the concentration result.

4.7.2.1 The Matrix Laplace Transform Method

We begin with a matrix extension of the moment generating function (mgf), which has played a major role in recent work on matrix concentration.

Definition 4.7.3 (Trace Mgf)**.** *Let $\boldsymbol{Y}$ be a random Hermitian matrix. The* normalized trace moment generating function *of $\boldsymbol{Y}$ is defined as*

$$m(\theta) := m_{\boldsymbol{Y}}(\theta) := \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta \boldsymbol{Y}} \quad \text{for } \theta \in \mathbb{R}.$$

The expectation need not exist for all values of θ.

The following proposition explains how the trace mgf can be used to study the maximum eigenvalue of a random Hermitian matrix [231, Prop. 3.1].

Proposition 4.7.4 (Matrix Laplace Transform Method)**.** *Let $\boldsymbol{Y} \in \mathbb{H}^d$ be a random matrix with normalized trace mgf $m(\theta) := \bar{\operatorname{tr}}\,\mathrm{e}^{\theta \boldsymbol{Y}}$. For each $t \in \mathbb{R}$,*

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge t\right\} \le d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + \log m(\theta)}.$$

4.7.2.2 A Differential Inequality for the Trace Mgf

Suppose that $\boldsymbol{Y} \in \mathbb{H}^d$ is a random Hermitian matrix that depends on a random vector $\boldsymbol{x} := (X_1, \dots, X_n)$. We require the distribution of $\boldsymbol{Y}$ to be invariant under signed permutations, and we insist that $\|\boldsymbol{Y}\|$ is bounded. Without loss of generality, assume that $\boldsymbol{Y}$ has zero mean. Throughout the argument, we let the notation of Section 4.2.2.3 and Theorem 4.2.8 prevail.

Let us explain how to use the subadditivity of matrix φ-entropy to derive a differential inequality for the trace mgf. Consider the function $\varphi(t) = t \log t$, which belongs to the Φ_∞ class because of Theorem 4.2.3(1). Introduce the random positive-definite matrix $\boldsymbol{Z} := \mathrm{e}^{\theta \boldsymbol{Y}}$, where $\theta > 0$. We write out an expression for the matrix φ-entropy of $\boldsymbol{Z}$:

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E} \,\bar{\operatorname{tr}}[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}\, \boldsymbol{Z})] \\ &= \mathbb{E} \,\bar{\operatorname{tr}} \left[(\theta \boldsymbol{Y}) \mathrm{e}^{\theta \boldsymbol{Y}} - \mathrm{e}^{\theta \boldsymbol{Y}} \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{Y}} \right] \\ &= \theta \cdot \mathbb{E} \,\bar{\operatorname{tr}} \left[\boldsymbol{Y} \mathrm{e}^{\theta \boldsymbol{Y}} \right] - (\mathbb{E} \,\bar{\operatorname{tr}}\, \mathrm{e}^{\theta \boldsymbol{Y}}) \log (\mathbb{E} \,\bar{\operatorname{tr}}\, \mathrm{e}^{\theta \boldsymbol{Y}}) \\ &= \theta m'(\theta) - m(\theta) \log m(\theta). \end{aligned} \tag{4.7.2}$$

In the third line, we have applied Lemma 4.7.1 to the logarithm in the second term, relying on the fact that $\boldsymbol{Y}$ is invariant under signed permutations. To reach the last line, we recognize that $m'(\theta) = \mathbb{E} \,\bar{\operatorname{tr}}(\boldsymbol{Y} \mathrm{e}^{\theta \boldsymbol{Y}})$. We have used the boundedness of $\|\boldsymbol{Y}\|$ to justify this derivative calculation.

Corollary 4.2.6 provides an upper bound for the matrix φ-entropy. Define the derivative $\psi(t) = \varphi'(t) = 1 + \log t$. Then

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{1}{2} \sum\nolimits_{i=1}^n \mathbb{E} \,\bar{\operatorname{tr}} \left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i') \right] \\ &= \frac{\theta}{2} \sum\nolimits_{i=1}^n \mathbb{E} \,\bar{\operatorname{tr}} \left[(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathrm{e}^{\theta \boldsymbol{Y}_i'})(\boldsymbol{Y} - \boldsymbol{Y}_i') \right]. \end{aligned}$$

Consider the function $f : t \mapsto \mathrm{e}^{\theta t}$. Its derivative $f' : t \mapsto \theta \mathrm{e}^{\theta t}$ is convex because $\theta > 0$, so Proposition 4.7.2 delivers the bound

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{\theta^2}{4} \sum\nolimits_{i=1}^n \mathbb{E} \,\bar{\operatorname{tr}} \left[(\mathrm{e}^{\theta \boldsymbol{Y}} + \mathrm{e}^{\theta \boldsymbol{Y}_i'})(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \right] \\ &= \frac{\theta^2}{2} \sum\nolimits_{i=1}^n \mathbb{E} \,\bar{\operatorname{tr}} \left[\mathrm{e}^{\theta \boldsymbol{Y}} (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \right] \\ &= \frac{\theta^2}{2} \sum\nolimits_{i=1}^n \mathbb{E} \,\bar{\operatorname{tr}} \left[\mathrm{e}^{\theta \boldsymbol{Y}} \cdot \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}] \right]. \end{aligned}$$

The second relation follows from the fact that $\boldsymbol{Y}$ and $\boldsymbol{Y}_i'$ are exchangeable, conditional on $\boldsymbol{x}_{-i}$. The last line is just the tower property of conditional expectation, combined with the observation that $\boldsymbol{Y}$ is a function of $\boldsymbol{x}$. To continue, we simplify the expression and make some additional bounds.

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{\theta^2}{2}\,\mathbb{E}\,\bar{\operatorname{tr}}\left[\mathrm{e}^{\theta\boldsymbol{Y}}\cdot\sum\nolimits_{i=1}^n \mathbb{E}[(\boldsymbol{Y}-\boldsymbol{Y}_i')^2\,|\,\boldsymbol{x}]\right] \\ &\le \frac{\theta^2}{2}(\mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\boldsymbol{Y}})\left\|\sum\nolimits_{i=1}^n \mathbb{E}[(\boldsymbol{Y}-\boldsymbol{Y}_i')^2\,|\,\boldsymbol{x}]\right\| \\ &\le \frac{\theta^2 V_{\boldsymbol{Y}}}{2}\cdot m(\theta). \end{aligned} \tag{4.7.3}$$

The second relation follows from a standard trace inequality and the observation that $\mathrm{e}^{\theta\boldsymbol{Y}}$ is positive definite. Last, we identify the variance measure $V_{\boldsymbol{Y}}$ defined in (4.2.4) and the trace mgf $m(\theta)$.

Combine the expression (4.7.2) with the inequality (4.7.3) to arrive at the estimate

$$\theta m'(\theta) - m(\theta)\log m(\theta) \le \frac{\theta^2 V_{\boldsymbol{Y}}}{2}\cdot m(\theta) \quad \text{for } \theta > 0. \tag{4.7.4}$$

We can use this differential inequality to obtain bounds on the trace mgf $m(\theta)$.

4.7.2.3 Solving the Differential Inequality

Rearrange the differential inequality (4.7.4) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{\log m(\theta)}{\theta}\right] = \frac{m'(\theta)}{\theta m(\theta)} - \frac{\log m(\theta)}{\theta^2} \le \frac{V_{\boldsymbol{Y}}}{2}. \tag{4.7.5}$$

The l'Hôpital rule allows us to calculate the value of $\theta^{-1}\log m(\theta)$ at zero. Since $m(0) = 1$,

$$\lim_{\theta\to 0}\frac{\log m(\theta)}{\theta} = \lim_{\theta\to 0}\frac{m'(\theta)}{m(\theta)} = \lim_{\theta\to 0}\frac{\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}\mathrm{e}^{\theta\boldsymbol{Y}})}{\mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\boldsymbol{Y}}} = \mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{Y} = 0.$$

This is where we use the hypothesis that $\boldsymbol{Y}$ has mean zero. Now, we integrate (4.7.5) from zero to some positive value θ to find that the trace mgf satisfies

$$\frac{\log m(\theta)}{\theta} \le \frac{\theta V_{\boldsymbol{Y}}}{2} \quad \text{when } \theta > 0. \tag{4.7.6}$$

The approach in this section is usually referred to as the Herbst argument [124].

4.7.2.4 The Laplace Transform Argument

We are now prepared to finish the argument. Combine the matrix Laplace transform method, Proposition 4.7.4, with the trace mgf bound (4.7.6) to reach

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \geq t\right\} \leq d \cdot \inf_{\theta>0} \mathrm{e}^{-\theta t+\log m(\theta)} \leq d \cdot \inf_{\theta>0} \mathrm{e}^{-\theta t+\theta^2 V_{\boldsymbol{Y}}/2} = d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}. \tag{4.7.7}$$

To obtain the result for the minimum eigenvalue, we note that

$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y}) \leq -t\right\} = \mathbb{P}\left\{\lambda_{\max}(-\boldsymbol{Y}) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}.$$

The inequality follows when we apply (4.7.7) to the random matrix $-\boldsymbol{Y}$. This completes the proof of Theorem 4.2.8.

4.8 Moment Inequalities for Random Matrices with Bounded Differences

In this section, we prove Theorem 4.2.9, which gives information about the moments of a random matrix that satisfies a kind of self-bounding property.

Proof of Theorem 4.2.9. Fix a number $q \in \{2, 3, 4, \dots\}$. Suppose that $\boldsymbol{Y} \in \mathbb{H}_+^d$ is a random positive-semidefinite matrix that depends on a random vector $\boldsymbol{x} := (X_1, \dots, X_n)$. We require the distribution of $\boldsymbol{Y}$ to be invariant under signed permutations, and we assume that $\mathbb{E}(\|\boldsymbol{Y}\|^q) < \infty$. The notation of Section 4.2.2.3 and Theorem 4.2.9 remains in force.

Let us explain how the subadditivity of matrix φ-entropy leads to a bound on the qth trace moment of $\boldsymbol{Y}$. Consider the power function $\varphi(t) = t^{q/(q-1)}$. Theorem 4.6.7 ensures that $\varphi \in \Phi_\infty$ because $q/(q-1) \in (1, 2]$. Introduce the random positive-semidefinite matrix $\boldsymbol{Z} := \boldsymbol{Y}^{q-1}$. Then

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\,\bar{\operatorname{tr}}\left[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}\,\boldsymbol{Z})\right] \\ &= \mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^q) - \bar{\operatorname{tr}}\left[(\mathbb{E}(\boldsymbol{Y}^{q-1}))^{q/(q-1)}\right] \\ &= \mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^q) - \left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)}. \end{aligned} \tag{4.8.1}$$

The transition to the last line requires Lemma 4.7.1.

Corollary 4.2.6 provides an upper bound for the matrix φ-entropy. Define the derivative $\psi(t) = \varphi'(t) = (q/(q-1)) \cdot t^{1/(q-1)}$. We have

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\leq \frac{1}{2} \sum\nolimits_{i=1}^n \mathbb{E}\, \bar{\operatorname{tr}} \left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i')) \right] \\ &= \frac{q}{2(q-1)} \sum\nolimits_{i=1}^n \mathbb{E}\, \bar{\operatorname{tr}} \left[\big(\boldsymbol{Y}^{q-1} - (\boldsymbol{Y}_i')^{q-1}\big)(\boldsymbol{Y} - \boldsymbol{Y}_i') \right]. \end{aligned}$$

The function $f : t \mapsto t^{q-1}$ has the derivative $f' : t \mapsto (q-1)t^{q-2}$, which is convex because $q \in \{2, 3, 4, \dots\}$. Therefore, the mean value trace inequality, Proposition 4.7.2, delivers the bound

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\leq \frac{q}{4} \sum\nolimits_{i=1}^n \mathbb{E}\, \bar{\operatorname{tr}} \left[\big(\boldsymbol{Y}^{q-2} + (\boldsymbol{Y}_i')^{q-2}\big)(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \right] \\ &= \frac{q}{2} \sum\nolimits_{i=1}^n \mathbb{E}\, \bar{\operatorname{tr}} \left[\boldsymbol{Y}^{q-2} (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \right] \\ &= \frac{q}{2} \sum\nolimits_{i=1}^n \mathbb{E}\, \bar{\operatorname{tr}} \left[\boldsymbol{Y}^{q-2}\, \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}] \right]. \end{aligned}$$

The second identity holds because $\boldsymbol{Y}$ and $\boldsymbol{Y}_i'$ are exchangeable, conditional on $\boldsymbol{x}_{-i}$. The last line follows from the tower property of conditional expectation. We simplify this expression as follows.

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\leq \frac{q}{2}\, \mathbb{E}\, \bar{\operatorname{tr}} \left[\boldsymbol{Y}^{q-2} \cdot \sum\nolimits_{i=1}^n \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}] \right] \\ &\leq \frac{q}{2}\, \mathbb{E}\, \bar{\operatorname{tr}} \left[\boldsymbol{Y}^{q-2} \cdot c\boldsymbol{Y} \right] \\ &= \frac{cq}{2}\, \mathbb{E}\, \bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1}). \end{aligned} \tag{4.8.2}$$

The second inequality derives from the hypothesis (4.2.5) that $\boldsymbol{V}_{\boldsymbol{Y}} \preccurlyeq c\boldsymbol{Y}$. Note that this bound requires the fact that $\boldsymbol{Y}^{q-2}$ is positive semidefinite.

Combine the expression (4.8.1) for the matrix φ-entropy with the upper bound (4.8.2) to achieve the estimate

$$\mathbb{E}\, \bar{\operatorname{tr}}(\boldsymbol{Y}^q) - \left[\mathbb{E}\, \bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1}) \right]^{q/(q-1)} \leq \frac{cq}{2}\, \mathbb{E}\, \bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1}).$$

Rewrite this bound, and invoke the numerical fact $1 + aq \leq (1+a)^q$ to obtain

$$\begin{aligned}\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^q) &\leq \left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)} \left(1 + \frac{cq/2}{\left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1})\right]^{1/q-1}}\right) \\ &\leq \left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)} \left(1 + \frac{c/2}{\left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1})\right]^{1/q-1}}\right)^q.\end{aligned}$$

Extract the qth root from both sides to reach

$$\left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^q)\right]^{1/q} \leq \left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^{q-1})\right]^{1/(q-1)} + \frac{c}{2}.$$

We have compared the qth trace moment of $\boldsymbol{Y}$ with the $(q-1)$th trace moment. Proceeding by iteration, we arrive at

$$\left[\mathbb{E}\,\bar{\operatorname{tr}}(\boldsymbol{Y}^q)\right]^{1/q} \leq \mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{Y} + \frac{q-1}{2}\cdot c.$$

This observation completes the proof of Theorem 4.2.9. □

4.9 Lemma 4.4.1, The General Case

In this appendix, we explain how to prove Lemma 4.4.1 in full generality. The argument calls for a simple but powerful result, known as the generalized Klein inequality [174, Prop. 3], which allows us to lift a large class of scalar inequalities to matrices.

Proposition 4.9.1 (Generalized Klein Inequality)**.** *For each $k = 1, \dots, n$, suppose that $f_k : I_1 \to \mathbb{R}$ and $g_k : I_2 \to \mathbb{R}$ are functions on intervals I_1 and I_2 of the real line. Suppose that*

$$\sum\nolimits_{k=1}^{n} f_k(a)\, g_k(b) \geq 0 \quad \textit{for all } a \in I_1 \textit{ and } b \in I_2.$$

Then, for each natural number d,

$$\sum\nolimits_{k=1}^{n} \bar{\operatorname{tr}}[f_k(\boldsymbol{A})\, g_k(\boldsymbol{B})] \geq 0 \quad \textit{for all } \boldsymbol{A} \in \mathbb{H}^d(I_1) \textit{ and } \boldsymbol{B} \in \mathbb{H}^d(I_2).$$

Proof of Lemma 4.4.1, General Case. We retain the notation from Lemma 4.4.1. In particular, we assume that $\boldsymbol{Z}$ is a random positive-definite matrix for which $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are both integrable. We also assume that $\boldsymbol{T}$ is a random positive-definite matrix with $\|\boldsymbol{T}\|$ and

$\|\varphi(\boldsymbol{T})\|$ integrable.

For $n \in \mathbb{N}$, define the function $l_n(a) := (a \vee 1/n) \wedge n$, where $\vee$ denotes the maximum operator and $\wedge$ denotes the minimum operator. Consider the random matrices $\boldsymbol{Z}_n := l_n(\boldsymbol{T})$ and $\boldsymbol{T}_k := l_k(\boldsymbol{T})$ for each $k, n \in \mathbb{N}$. These matrices have eigenvalues that are bounded and bounded away from zero, so these entities satisfy the inequality (4.4.3) we have already established.

$$H_\varphi(\boldsymbol{Z}_n) \geq \mathbb{E}\,\bar{\mathrm{tr}}\,\big[(\psi(\boldsymbol{T}_k) - \psi(\mathbb{E}\,\boldsymbol{T}_k))(\boldsymbol{Z}_n - \boldsymbol{T}_k) + \mathbb{E}\,\varphi(\boldsymbol{T}_k - \varphi(\mathbb{E}\,\boldsymbol{T}_k)\big].$$

Rearrange the terms in this inequality to obtain

$$\mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) \geq \bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T}_k)(\mathbb{E}\,\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{T}_k) - \varphi(\mathbb{E}\,\boldsymbol{T}_k) + \varphi(\mathbb{E}\,\boldsymbol{Z}_n)\big], \tag{4.9.1}$$

where we have introduced the function

$$\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{B}) := \varphi(\boldsymbol{A}) - \varphi(\boldsymbol{B}) - (\boldsymbol{A} - \boldsymbol{B})\psi(\boldsymbol{B}) \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d_{++}.$$

To complete the proof of Lemma 4.4.1, we must develop the bound

$$\mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}) \geq \bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T})(\mathbb{E}\,\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T}) + \varphi(\mathbb{E}\,\boldsymbol{Z})\big] \tag{4.9.2}$$

by driving $k, n \to \infty$ in (4.9.1).

Let us begin with the right-hand side of (4.9.1). We have the sure limit $\boldsymbol{Z}_n \to \boldsymbol{Z}$. Therefore, the Dominated Convergence Theorem guarantees that $\mathbb{E}\,\boldsymbol{Z}_n \to \mathbb{E}\,\boldsymbol{Z}$ because $\|\boldsymbol{Z}\|$ is integrable and $\|\boldsymbol{Z}_n\| \leq \|\boldsymbol{Z}\|$. Likewise, $\mathbb{E}\,\boldsymbol{T}_k \to \mathbb{E}\,\boldsymbol{T}$. The functions φ and ψ are continuous, so the limit of the right-hand side of (4.9.1) satisfies

$$\begin{aligned}&\bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T}_k)(\mathbb{E}\,\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{T}_k) - \varphi(\mathbb{E}\,\boldsymbol{T}_k) + \varphi(\mathbb{E}\,\boldsymbol{Z}_n)\big] \\ &\qquad\qquad \to \bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T})(\mathbb{E}\,\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T}) + \varphi(\mathbb{E}\,\boldsymbol{Z})\big].\end{aligned} \tag{4.9.3}$$

This expression coincides with the right-hand side of (4.9.2).

Taking the limit of the left-hand side of (4.9.1) is more involved because the function ψ may grow quickly at zero and infinity. We accomplish our goal in two steps. First, we take the limit as $n \to \infty$. Afterward, we take the limit as $k \to \infty$.

Introduce the nonnegative function

$$\gamma(z,t) := \varphi(z) - \varphi(t) - (z-t)\psi(t) \quad \text{for } z,t > 0.$$

Boucheron et al. [26, p. 525] establish that

$$\gamma(l_n(z), l_k(t)) \le \gamma(1, l_k(t)) + \gamma(z, l_k(t)) \quad \text{for } z,t > 0. \tag{4.9.4}$$

The generalized Klein inequality, Proposition 4.9.1, can be applied (with due diligence) to extend (4.9.4) to matrices. In particular,

$$\bar{\operatorname{tr}}\, \mathbf{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) = \bar{\operatorname{tr}}\, \mathbf{\Gamma}(l_n(\boldsymbol{Z}), l_k(\boldsymbol{T})) \le \bar{\operatorname{tr}}[\mathbf{\Gamma}(\mathbf{I}, l_k(\boldsymbol{T})) + \mathbf{\Gamma}(\boldsymbol{Z}, l_k(\boldsymbol{T}))] = \bar{\operatorname{tr}}[\mathbf{\Gamma}(\mathbf{I}, \boldsymbol{T}_k) + \mathbf{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k)].$$

Observe that the right-hand side of this inequality is integrable. Indeed, all of the quantities involving $\boldsymbol{T}_k$ are uniformly bounded because the eigenvalues of $\boldsymbol{T}_k$ fall in the range $[k^{-1}, k]$ and the functions φ and ψ are continuous on this interval. The terms involving $\boldsymbol{Z}$ may not be bounded, but they are integrable because $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. We may now apply the Dominated Convergence Theorem to take the limit:

$$\mathbb{E}\, \bar{\operatorname{tr}}\, \mathbf{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) \to \mathbb{E}\, \bar{\operatorname{tr}}\, \mathbf{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k) \quad \text{as } n \to \infty, \tag{4.9.5}$$

where we rely again on the sure limit $\boldsymbol{Z}_n \to \boldsymbol{Z}$ as $n \to \infty$.

Boucheron et al. also establish that

$$\gamma(z, l_k(t)) \le \gamma(z, 1) + \gamma(z,t) \quad \text{for } z,t > 0.$$

The generalized Klein inequality, Proposition 4.9.1, ensures that

$$\bar{\operatorname{tr}}\, \mathbf{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k) \le \bar{\operatorname{tr}}[\mathbf{\Gamma}(\boldsymbol{Z}, \mathbf{I}) + \mathbf{\Gamma}(\boldsymbol{Z}, \boldsymbol{T})].$$

We may assume that the second term on the right-hand side is integrable or else the desired inequality (4.9.2) would be vacuous. The first term is integrable because $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$

are integrable. Therefore, we may apply the Dominated Convergence Theorem:

$$\mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z},\boldsymbol{T}_k) \to \mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z},\boldsymbol{T}) \quad \text{as } k \to \infty, \tag{4.9.6}$$

where we rely again on the sure limit $\boldsymbol{T}_k \to \boldsymbol{T}$ as $k \to \infty$.

In summary, the limits (4.9.5) and (4.9.6) provide that $\mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z}_n,\boldsymbol{T}_k) \to \mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z},\boldsymbol{T})$ as $k, n \to \infty$. In view of the limit (4.9.3), we have completed the proof of (4.9.2). □

4.10 Generalized Subadditivity of Matrix φ-Entropy

In this section, we establish Theorem 4.2.10, which states the subadditivity of the generalized matrix φ-entropy for every function in the Φ_∞ class. The generalized result depends on the supremum representation of the generalized matrix φ-entropy, which we establish in Section 4.10.1. We use the supremum representation to derive a generalized Jensen-type inequality in Section 4.10.2 and finally prove Theorem 4.2.10 in Section 4.10.3.

4.10.1 Representation of the Generalized φ-Entropy as a Supremum

The following variational representation of the generalized φ-entropy is the extension of Lemma 4.4.1 and is the key behind the generalized subadditivity theorem.

Lemma 4.10.1. *Fix a function $\varphi \in \Phi_\infty$, and let $\psi = \varphi'$. Suppose that $\boldsymbol{Z}$ is a $d \times d$ positive-semidefinite matrix. Then*

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}) = \sup_{\boldsymbol{T}\in\mathbb{H}_+^d} \bar{\operatorname{tr}}\big[(\psi(\boldsymbol{T}) - \psi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T})(\boldsymbol{Z}-\boldsymbol{T}) + \varphi(\boldsymbol{T}) - \varphi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T})\big]. \tag{4.10.1}$$

In particular, the generalized φ-entropy $H_\varphi(\boldsymbol{Z}|\mathfrak{A})$ can be written in the dual form

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}) = \sup_{\boldsymbol{T}\in\mathbb{H}_+^d} \bar{\operatorname{tr}}\,\big(\boldsymbol{\Upsilon}_1(\boldsymbol{T})\cdot\boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\big), \tag{4.10.2}$$

where $\boldsymbol{\Upsilon}_i : \mathbb{H}_+^d \to \mathbb{H}^d$ for $i = 1, 2$.

We proceed to establish Lemma 4.10.1 in the ensuing sections.

4.10.1.1 Generalized Convexity Lemma

To establish the variational formula, we require the following convexity result, which is an extentsion of Lemma 4.4.2.

Lemma 4.10.2. *Fix a function $\varphi \in \Phi_\infty$, and let $\psi = \varphi'$. Then*

$$\langle \boldsymbol{K}, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\rangle \geq \langle \mathbb{E}_{\mathfrak{A}}\, \boldsymbol{K}, \mathsf{D}\psi(\mathbb{E}_{\mathfrak{A}}\, \boldsymbol{Y})(\mathbb{E}_{\mathfrak{A}}\, \boldsymbol{K})\rangle, \quad \textit{for } \boldsymbol{K}, \boldsymbol{Y} \in \mathbb{H}_+^d.$$

The proof of Lemma 4.10.2 depends on the following proposition [46].

Proposition 4.10.3. *For any m matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_m \in \mathbb{M}^d$, and any $*$-subalgebra $\mathfrak{A}$ of $\mathbb{M}^d$, there exists a sequence $\{\mathsf{C}_k\}_{k\in\mathbb{N}}$ of operators of the form*

$$\mathsf{C}_k(\boldsymbol{A}) = \sum\nolimits_{j=1}^{N_k} p_{k,j} \boldsymbol{U}_{k,j} \boldsymbol{A} \boldsymbol{U}_{k,j}^*, \tag{4.10.3}$$

where $\boldsymbol{U}_{k,j}$ are unitary, $p_{k,j} > 0$ and $\sum_{j=1}^{N_k} p_{k,j} = 1$, such that,

$$\mathbb{E}_{\mathfrak{A}}(\boldsymbol{A}_j) = \lim_{k\to\infty} \mathsf{C}_k(\boldsymbol{A}_j) \quad \textit{for each } j = 1, \dots, m.$$

Each operator C_k maps a matrix to a convex combination of its unitary conjugations. Proposition 4.10.3 states that for any number of fixed matrices, one can produce sequences of matrices from the operators $\{\mathsf{C}_k\}_{k\in\mathbb{N}}$ such that each sequence converges to the conditional expectation of that matrix with respect to the $*$-subalgebra. This synchronized approximation becomes very useful in proving convexity of multivariate functions, as we shall see in the following proof Lemma 4.10.2.

Proof of Lemma 4.10.2. First, we verify that the function $(\boldsymbol{K}, \boldsymbol{Y}) \mapsto \langle \boldsymbol{K}, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\rangle$ is invariant under unitary conjugation. Suppose $\boldsymbol{Y}$ has the spectral decomposition $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$, then for any unitary $\boldsymbol{V}$,

$$\begin{aligned}
\big\langle \boldsymbol{VKV}^*, \mathsf{D}\psi(\boldsymbol{VYV}^*)(\boldsymbol{VKV}^*)\big\rangle &= \big\langle \boldsymbol{VKV}^*, \boldsymbol{VU}\big[\psi^{[1]}(\boldsymbol{\Lambda}) \odot (\boldsymbol{U}^*\boldsymbol{V}^*\boldsymbol{VKV}^*\boldsymbol{VU})\big]\boldsymbol{U}^*\boldsymbol{V}^*\big\rangle \\
&= \big\langle \boldsymbol{VKV}^*, \boldsymbol{VU}\big[\psi^{[1]}(\boldsymbol{\Lambda}) \odot (\boldsymbol{U}^*\boldsymbol{KU})\big]\boldsymbol{U}^*\boldsymbol{V}^*\big\rangle \\
&= \big\langle \boldsymbol{K}, \boldsymbol{U}\big[\psi^{[1]}(\boldsymbol{\Lambda}) \odot (\boldsymbol{U}^*\boldsymbol{KU})\big]\boldsymbol{U}^*\big\rangle \\
&= \big\langle \boldsymbol{K}, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\big\rangle.
\end{aligned} \tag{4.10.4}$$

The first relation is the Daleckiĭ–Kreĭn Formula, Proposition 4.6.3, and we recall $\psi^{[1]}$ is the first divided difference of ψ. As $\boldsymbol{V}$ is unitary, the second relation follows by equating $\boldsymbol{V}^*\boldsymbol{V}$ to $\mathbf{I}$. The third relation is because the normalized trace inner product is invariant under unitary conjugation. We apply the Daleckiĭ–Kreĭn Formula again in the fourth relation.

Proposition 4.10.3 allows us to approximate $\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{K}$ and $\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{Y}$ simultaneously with a sequence of operators $\{\mathsf{C}_k\}_{k\in\mathbb{N}}$ that map each matrix into a convex combination of its unitary conjugations.

$$\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{K} = \lim_{k\to\infty} \mathsf{C}_k(\boldsymbol{K}) \quad \text{and} \quad \mathbb{E}_{\mathfrak{A}}\,\boldsymbol{Y} = \lim_{k\to\infty} \mathsf{C}_k(\boldsymbol{Y}).$$

Recall that we already establish the convexity of the function $(\boldsymbol{K},\boldsymbol{Y}) \mapsto \langle \boldsymbol{K}, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\rangle$ by Lemma 4.4.2, we

$$\begin{aligned}\big\langle \mathsf{C}_k(\boldsymbol{K}), \mathsf{D}\psi(\mathsf{C}_k(\boldsymbol{Y}))(\mathsf{C}_k(\boldsymbol{K}))\big\rangle &\le \sum\nolimits_{j=1}^{N_k} p_{k,j}\cdot \big\langle \boldsymbol{U}_{kj,}\boldsymbol{K}\boldsymbol{U}_{kj,}^*, \mathsf{D}\psi(\boldsymbol{U}_{kj,}\boldsymbol{Y}\boldsymbol{U}_{kj,}^*)(\boldsymbol{U}_{kj,}\boldsymbol{K}\boldsymbol{U}_{kj,}^*)\big\rangle \\ &= \sum\nolimits_{j=1}^{N_k} \big\langle \boldsymbol{K}, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\big\rangle, \end{aligned} \tag{4.10.5}$$

where the second equation is due to (4.10.4).

Next, take k to infinity on the left-hand side of (4.10.5) and we established the desired inequality:

$$\langle \mathbb{E}_{\mathfrak{A}}\,\boldsymbol{K}, \mathsf{D}\psi(\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{Y})(\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{K})\rangle \le \langle \boldsymbol{K}, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\rangle.$$

□

4.10.1.2 Proof of Lemma 4.10.1

The argument parallels the proof of Lemma 4.4.1. Since $\boldsymbol{Z}$ is deterministic, there is no regularity issues and the argument is much simpler. The case when φ is a positive affine function is trivial. Thus, we prove the case when φ is not affine. When we substitute $\boldsymbol{T} = \boldsymbol{Z}$ into the argument of the supremum in (4.4.1), the right-hand side equals $H_\varphi(\boldsymbol{Z}|\mathfrak{A})$ and attains the supremum. Thus, we just need to verify the inequality

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}) \ge \bar{\operatorname{tr}}\big[(\psi(\boldsymbol{T}) - \psi(\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{T}))(\boldsymbol{Z}-\boldsymbol{T}) + \varphi(\boldsymbol{T}) - \varphi(\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{T})\big],$$

for any matrix $\boldsymbol{T} \in \mathbb{H}_+^d$.

We use an interpolation argument. For any $\boldsymbol{T} \in \mathbb{H}_+^d$, we define the family of positive-semidefinite matrices

$$\boldsymbol{T}_s := (1-s)\cdot \boldsymbol{Z} + s\cdot \boldsymbol{T} \quad \text{for } s\in[0,1],$$

which interpolates between $\boldsymbol{Z}$ and $\boldsymbol{T}$. We then define a real-valued function $F(s)$:

$$F(s) := \bar{\operatorname{tr}}\big[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T}_s))\cdot(\boldsymbol{Z}-\boldsymbol{T}_s)\big] + \bar{\operatorname{tr}}[\varphi(\boldsymbol{T}_s) - \varphi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T}_s)]. \tag{4.10.6}$$

Following the same arguments as in the proof of Theorem 4.2.5, we differentiate the function F to obtain:

$$\begin{aligned} F'(s) = -s\cdot \bar{\operatorname{tr}}\left[\mathsf{D}\psi(\boldsymbol{T}_s)(\boldsymbol{T}-\boldsymbol{Z})\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + s\cdot \bar{\operatorname{tr}}\left[\mathsf{D}\psi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T}_s)(\mathbb{E}_{\mathfrak{A}}(\boldsymbol{T}-\boldsymbol{Z}))\cdot(\mathbb{E}_{\mathfrak{A}}(\boldsymbol{T}-\boldsymbol{Z}))\right] \\ - \bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + \bar{\operatorname{tr}}\left[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right]. \end{aligned} \tag{4.10.7}$$

The last two terms of (4.10.7) cancel, and we can rewrite the first two terms using the trace inner product:

$$F'(s) = s\cdot\left[\langle \mathbb{E}_{\mathfrak{A}}(\boldsymbol{Z}-\boldsymbol{T}), \mathsf{D}\psi(\mathbb{E}_{\mathfrak{A}}\boldsymbol{T}_s)(\mathbb{E}_{\mathfrak{A}}(\boldsymbol{Z}-\boldsymbol{T}))\rangle - \langle(\boldsymbol{Z}-\boldsymbol{T}), \mathsf{D}\psi(\boldsymbol{T}_s)(\boldsymbol{Z}-\boldsymbol{T})\rangle\right].$$

Invoke Lemma 4.10.2 and we conclude that $F'(s)\le 0$ for $s\in[0,1]$.

4.10.2 Generalized Conditional Jensen Inequality

The supremum representation in Lemma 4.10.1 leads directly to the following Jensen inequality, which is an extension of our previous Lemma 4.4.3.

Lemma 4.10.4. *Fix a function $\varphi\in\Phi_\infty$. Suppose $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are two commuting $*$-subalgebra in $\mathbb{M}^d$. Then, for any matrix $\boldsymbol{Z}\in\mathbb{H}^d_+$*

$$H_\varphi\big(\mathbb{E}_{\mathfrak{A}_1}\boldsymbol{Z}|\mathfrak{A}_2\big) \le H_\varphi(\boldsymbol{Z}|\mathfrak{A}_2).$$

Proof. Lemma 4.10.1 allows us to represent the generalized φ-entropy as a supremum:

$$H_\varphi(\mathbb{E}_{\mathfrak{A}_1}\boldsymbol{Z}|\mathfrak{A}_2) = \sup_{\boldsymbol{T}\in\mathbb{H}^d_+} \bar{\operatorname{tr}}\big(\boldsymbol{\Upsilon}_1(\boldsymbol{T})\cdot\mathbb{E}_{\mathfrak{A}_1}\boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\big).$$

Notice that the supremum is achieved when $\boldsymbol{T} = \mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z} \in \mathfrak{A}_1$, thus instead of taking the supremum of $\mathbb{H}_+^d$, we can take the supremum over a smaller set $\mathfrak{A}_1 \cap \mathbb{H}_+^d$:

$$H_\varphi\big(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z} | \mathfrak{A}_2\big) = \sup_{\boldsymbol{T} \in \mathfrak{A}_1 \cap \mathbb{H}_+^d} \bar{\operatorname{tr}}\big(\boldsymbol{\Upsilon}_1(\boldsymbol{T}) \cdot \mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\big).$$

This representation is preferable as we can make the following calculations:

$$\begin{aligned} H_\varphi\big(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z} | \mathfrak{A}_2\big) &= \sup_{\boldsymbol{T} \in \mathfrak{A}_1 \cap \mathbb{H}_+^d} \bar{\operatorname{tr}}\, \mathbb{E}_{\mathfrak{A}_1} \big(\boldsymbol{\Upsilon}_1(\boldsymbol{T}) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\big) \\ &= \sup_{\boldsymbol{T} \in \mathfrak{A}_1 \cap \mathbb{H}_+^d} \bar{\operatorname{tr}}\big(\boldsymbol{\Upsilon}_1(\boldsymbol{T}) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\big) \\ &\leq \sup_{\boldsymbol{T} \in \mathbb{H}_+^d} \bar{\operatorname{tr}}\big(\boldsymbol{\Upsilon}_1(\boldsymbol{T}) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\big) \\ &= H_\varphi(\boldsymbol{Z} | \mathfrak{A}_2). \end{aligned}$$

When $\boldsymbol{T}$ is an element of $\mathfrak{A}_1$, both $\boldsymbol{\Upsilon}_1(\boldsymbol{T})$ and $\boldsymbol{\Upsilon}_2(\boldsymbol{T})$ are in $\mathfrak{A}_1$ and the property (4.10.1) leads to the first relation. The second relation is because $\bar{\operatorname{tr}}(\mathbb{E}_{\mathfrak{A}} \boldsymbol{A}) = \bar{\operatorname{tr}} \boldsymbol{A}$ for each matrix $\boldsymbol{A}$ in $\mathbb{H}_+^d$. The third relation is the supremum potentially increases we enlarge the range of the supermen. We apply Lemma 4.10.1 in the last relation to conclude the proof. □

4.10.3 Proof of Theorem 4.2.10

We are ready to established the subadditivity of the generalized φ-entropy, which is a direct consequence of the generalized conditional Jensen inequality, Lemma 4.10.4. The argument again parallels the proof of Theorem 4.2.5. In this argument, we write $\mathbb{E}$ to abbreviate the expectation $\mathbb{E}_{\mathfrak{A}_1,\ldots,\mathfrak{A}_n}$ taken over all the $*$-subalgebras.

First, we separate the φ-entropy into two parts by adding and subtracting terms:

$$\begin{aligned} H_\varphi(\boldsymbol{Z} | \mathfrak{A}_1, \ldots, \mathfrak{A}_n) &= \bar{\operatorname{tr}}\left[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z}) + \varphi(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z}) - \varphi(\mathbb{E}\, \boldsymbol{Z})\right] \\ &= \bar{\operatorname{tr}}\left[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z})\right] + \bar{\operatorname{tr}}\left[\varphi(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z}) - \varphi(\mathbb{E}\, \boldsymbol{Z})\right]. \end{aligned} \tag{4.10.8}$$

Identify the first term of (4.10.8) as the entropy $H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1)$. Because the $*$-subalgebras $\{\mathfrak{A}_1, \ldots, \mathfrak{A}_n\}$ commute, we can rewrite the total expectation

$$\mathbb{E}\, \boldsymbol{Z} = \mathbb{E}_{\mathfrak{A}_2,\ldots,\mathfrak{A}_n} \big(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z}\big).$$

As a result, we identify the second term in (4.10.8) as the entropy $H_\varphi(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z}|\mathfrak{A}_2, \dots, \mathfrak{A}_n)$. Thus

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1, \dots, \mathfrak{A}_n) = H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1) + H_\varphi(\mathbb{E}_{\mathfrak{A}_1} \boldsymbol{Z}|\mathfrak{A}_2, \dots, \mathfrak{A}_n).$$

Apply the generalized conditional Jensen inequality, and we have

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1, \dots, \mathfrak{A}_n) \leq H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1) + H_\varphi(\boldsymbol{Z}|\mathfrak{A}_2, \dots, \mathfrak{A}_n). \tag{4.10.9}$$

The first term on the right-hand side of (4.10.9) coincides with the first summand on the right-hand side of the subadditivity inequality (4.2.6). We argue that the second term on the right-hand side of (4.10.9) contains the remaining summands. Repeat the previous argument inductively to the term $H_\varphi(\boldsymbol{Z}|\mathfrak{A}_2, \dots, \mathfrak{A}_n)$, we obtain

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1, \dots, \mathfrak{A}_n) \leq H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1) + H_\varphi(\boldsymbol{Z}|\mathfrak{A}_2) + H_\varphi(\boldsymbol{Z}|\mathfrak{A}_3, \dots, \mathfrak{A}_n).$$

Continuing in this fashion, we arrive at the generalized subadditivity inequality (4.2.6):

$$H_\varphi(\boldsymbol{Z}|\mathfrak{A}_1, \dots, \mathfrak{A}_n) \leq \sum\nolimits_{i=1}^n H_\varphi(\boldsymbol{Z}|\mathfrak{A}_i).$$

Chapter 5

Solving Ptychography with a Convex Relaxation

Preface

This chapter is adapted from the work [100] that appears in the New Journal of Physics. This project is a collaboration between the candidate and Roarke Horstmeyer. Other authors include Brendan Ames, who was a postdoctoral researcher at Caltech during the project, a fellow graduate student Xiaoze Ou, the candidate's advisor Joel A. Tropp, and Horstmeyer's advisor Changhuei Yang.

The first novelty of this work consists of the application of a recent convex formulation of the phase retrieval problem to the setting of ptychography and the implementation of positive-semidefinite (PSD) programming algorithms to solve the problem. Simulation results provide concrete evidence that our convex implementation, the Convex Lifted Ptychography (CLP) solver, achieves better recovery results compared with the existing state-of-the-art algorithms that are based on alternating projection (AP). However, the computational complexity of PSD algorithms scales poorly when the size of the problem increases, rendering the complex implementation uncompetitive compared with (AP) algorithms. The second novelty of our work is a more efficient algorithm, called Low-Rank Ptychography (LRP), which is a trade-off between the superior performance of the convex implementation CLP and the computational efficiency of AP algorithms. The LRP algorithm is non-convex and based on low-rank matrix factorization and avoids the expensive eigenvalue decomposition that is required at each iterative step of the CLP solver. The computational complexity is scalable to large problems. Simulations establish that the LRP algorithm performs better than AP algorithms and it also approximates the performance of the CLP solver. We also implement our new algorithms on real experiment data to compare their performances with the AP algorithms.

From the mathematical and algorithmic perspectives, some open problems include: first, quantifying the exact performance guarantee of the CLP solver; second, establishing the convergence conditions for the non-convex approach of solving SPD based on low-rank matrix factorization. Theoretical analysis of the convex solver for the phase retrieval problem in the literature relies on many randomization assumptions, such as constructing the sampling matrix from certain random distributions. Our implementation is deterministic and highly structured, which poses analytical challenges and does not fit into existing performance analysis. Deriving a theoretical performance guarantee will aide a better understanding of the CLP solver. The LRP algorithm relies on a low-rank matrix factorization approach to solve PSD programs, which exhibits good convergence in practice but the exact convergence conditions are not formulated. It is an ongoing topic that interests many researchers.

From the application perspective, there are many directions to explore using our efficient LRP solver. One example is to apply the LRP solver to other type of test objects. In our project, we work with two-dimensional signals. It is plausible that one can extend the current formulation to conduct ptychography on three-dimensional objects, which allows extraction of more enriched information.

5.1 Introduction

Over the past two decades, ptychography [161, 162] has surpassed all other imaging techniques in its ability to produce high-resolution, wide field-of-view measurements of microscopic and nanoscopic phenomena. Whether in the X-ray regime at third-generation synchrotron sources [190, 222, 64, 201], in the electron microscope for atomic scale phenomena [104], or in the optical regime for biological specimens [141], ptychography has shown an unparalleled ability to acquire hundreds of megapixels of sample information near the diffraction limit. The standard ptychography principle is simple: a series of diffraction patterns are recorded from a sample as it is scanned through a focused beam. These intensity-only measurements are then computationally converted into a reconstruction of the complex sample (i.e., its amplitude and phase), which contains more pixels than a single recorded diffraction pattern.

A recently introduced imaging procedure, termed Fourier ptychography (FP), uses a similar principle to create gigapixel optical images with a conventional microscope [249]. The only required hardware modification is an LED array, which illuminates a stationary sample from different directions as the microscope captures a sequence of images. As in standard ptychography, FP must also recover the sample's phase as it merges together the captured image sequence into a high-resolution output. Standard and Fourier ptychographic data are connected via a linear transformation [102], which allows both setups to use nearly identical image reconstruction algorithms.

Standard and Fourier ptychography both avoid the need for a large, well-corrected lens to image at the diffraction-limit. Instead, they shift the majority of resolution-limiting factors into the computational realm. Unfortunately, an accurate and reliable solver does not yet exist. As a coherent diffractive imaging technique [49], ptychography must reconstruct the phase of the scattered field from measured intensities, which is an ill-posed problem. To date, most ptychography algorithms solve the phase retrieval problem by applying known constraints in an iterative manner. We categorize this class of algorithm as an "alternating projection" (AP) strategy. The simplest example of an AP strategy is the Gerchburg-Saxton (i.e., error reduction) algorithm [78]. Our AP category also includes the iterative projection and gradient search techniques reviewed by Fienup [74] and Marchesini [143], which map to analogous procedures in ptychography [81]. All AP strategies, including several related variants [66, 223, 244], often converge to incorrect local minima or can stagnate [73]. Few guarantees exist regarding convergence, let alone convergence to a reasonable solution. Despite these shortcomings, many authors have pushed beyond the basic algorithms [71] to account for unknown system parameters [140, 139], improve outcomes by careful initialization [144], perform multiplexed acquisition [225], and attempt three-dimensional imaging [82, 226].

In this article, we formulate a convex program for the ptychography problem, which allows us to use efficient computational methods to obtain a reliable image reconstruction. Convex optimization has recently matured into a powerful computational tool that now solves a variety of challenging problems [32]. However, many sub-disciplines of imaging, especially those involving phase retrieval, have been slow to reap its transformative benefits. Several prior works [12, 72, 8, 202, 42] have connected convex optimization with abstract phase retrieval problems, but this is the first work that applies convex optimization to the quickly growing field of high-resolution ptychography.

While it is possible in some experiments to improve reconstruction performance using prior sample knowledge or an appropriate heuristic, we consider here the general case of unaided recovery, which is especially relevant in biological imaging. Under these circumstances, we will show that our convex optimization approach to ptychographic reconstruction has many advantages over AP. Our formulation has no local minima, so we can always obtain a solution with minimum cost. The methodology is significantly more noise-tolerant than AP, and the results are more reproducible. There are also opportunities to establish theoretical guarantees using machinery from convex analysis.

Furthermore, there are many efficient algorithms for our convex formulation of the ptychography problem. To obtain solutions at scale, we apply a factorization method due to Burer and Monteiro [35, 36]. This method avoids the limitations of earlier convex algorithms for abstract phase retrieval, whose storage and complexity scale cubically in the number of reconstructed pixels [42]. Moreover, recent results establish that this factorization technique converges globally under certain conditions [61], offering a promising theoretical guarantee. The end result is a new, noise-tolerant algorithm for ptychographic reconstruction that is efficient enough to process the multi-gigapixel images that future applications will demand.

Here is an outline for the chapter. First, we develop a linear algebraic framework to illustrate the ptychographic image formation process. Second, we manipulate this framework to pose its sample recovery problem as a convex program. This initial algorithm, termed “convex lifted ptychography" (CLP), supports a-priori knowledge of noise statistics to significantly increase the accuracy of image reconstruction in the presence of noise. Third, we build upon research in low-rank semidefinite programming [35, 36] to develop a second non-convex algorithm, called “low-rank ptychography" (LRP), which improves on the computational profile of CLP. Finally, we explore the performance of LRP in both simulation and experiment to demonstrate how it

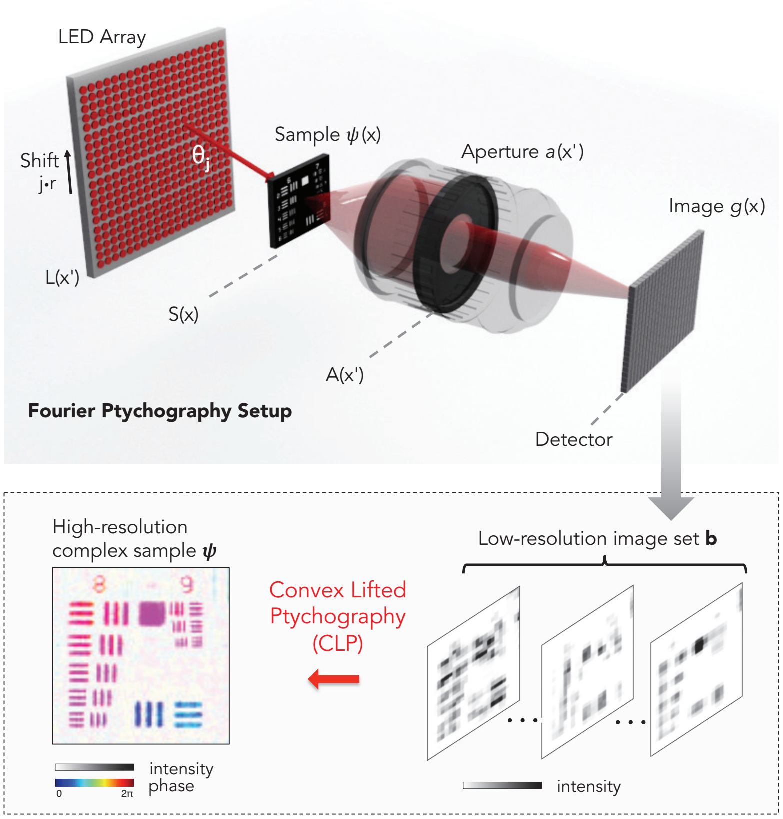


Figure 5.1.1: Diagram of the Fourier ptychography setup (top), where we use an LED array to illuminate a sample from different directions and acquire an image set **b** (bottom). This chapter introduces a convex phase retrieval algorithm to transform this image set into a high-resolution complex sample estimate ψ. Included image set and reconstructed resolution target are experimental results.

may be of great use in reducing the image capture time of Fourier ptychography.

5.2 Fundamentals

In this section, we outline the data capture process of Fourier ptychography (FP, see figure 5.1.1). At the end of this section, we discuss how a simple exchange of variables yields a nearly equivalent mathematical description of “standard" (i.e., diffraction imaging-based) ptychography data, which our proposed algorithm may also process. Since this exchange is straightforward, we choose to focus our attention on the FP problem for the majority of the manuscript. We encourage the interested reader to re-derive our algorithm for the standard ptychography arrangement. In addition, while the following analysis considers a two-dimensional experimental geometry for simplicity, extension to three dimensions is direct.

We assume that a distant plane $L(x')$ contains q different quasi-monochromatic optical sources (central wavelength λ) evenly distributed along x' with a spacing r. We assume each optical source acts as an effective point emitter that illuminates a sample $\psi(x)$ at a plane $S(x)$ a large distance l away from $L(x')$. Under this assumption, the jth source illuminates the sample with a spatially coherent plane wave at angle $\theta_j = \tan^{-1}(jr/l)$, where $-q/2 \leq j \leq q/2$. Additionally assuming the sample $\psi(x)$ is thin, we may express the optical field exiting the thin sample as the product,

$$s(x,j) = \psi(x)e^{ikxp_j}, \tag{5.2.1}$$

where the wavenumber $k = 2\pi/\lambda$ and $p_j = \sin\theta_j$ describes the off-axis angle of the jth optical source. The jth illuminated sample field $s(x,j)$ then enters an imaging system with a low numerical aperture (NA). Neglecting scaling factors and a quadratic phase factor for simplicity, Fourier optics gives the field at the imaging system pupil plane, $A(x')$, as $\mathcal{F}[s(x,j)] = \hat{\psi}(x' - p_j)$. Here, $\mathcal{F}$ represents the Fourier transform between conjugate variables x and x', $\hat{\psi}$ is the Fourier transform of ψ, and we have applied the Fourier shift property. The shifted spectrum field $\hat{\psi}(x' - p_j)$ is then modulated by the imaging system's aperture function $a(x')$, which acts as a low-pass filter. It is now useful to consider the spectrum $\hat{\psi}$ discretized into n pixels with a maximum spatial frequency k. We denote the bandpass cutoff of the aperture function a as $k{\cdot}m/n$, where m is an integer less than n. The modulation of $\hat{\psi}$ by a results in a field characterized by m discrete samples, which propagates to the camera imaging plane and is critically sampled by an m-pixel digital detector. This forms a reduced-resolution image, g:

$$g(x, j) = \left| \mathcal{F}\left[a(x')\hat{\psi}(x' - p_j) \right] \right|^2 . \tag{5.2.2}$$

$g(x, j)$ is an $(m \times q)$ Fourier ptychography data matrix. Its jth column contains a low-resolution image of the sample intensity while it is under illumination from the jth optical source.

The goal of Fourier ptychographic post-processing is to reconstruct a high-resolution (n-pixel) complex spectrum $\hat{\psi}(x')$, from the multiple low-resolution (m-pixel) intensity measurements contained within the data matrix g. Once $\hat{\psi}$ is found, an inverse-Fourier transform will yield the desired complex sample reconstruction, ψ. As noted above, most current ptychography setups solve this inverse problem using alternating projections (AP): after initializing a complex sample estimate, ψ_0, iterative constraints help force ψ_0 to obey all known physical conditions. First, its amplitude is forced to obey the measured intensity set from the detector plane (i.e., the values in g). Second, its spectrum $\hat{\psi}_0$ is forced to lie within a known support in the plane that is Fourier conjugate to the detector. Different projection operators and update rules are available, but are closely related [74, 143, 81]. While these projection strategies are known to converge when each constraint set is convex, the intensity constraint applied at the detector plane is not convex [13], leading to erroneous solutions [197] and possible stagnation [73].

The Fourier ptychography setup in figure 5.1.1 may be converted into a standard ptychography experiment by interchanging the sample plane S and the aperture plane A. This results in a standard ptychographic data matrix taking the form of equation 5.2.2 but now with a sample spectrum described in real space as ψ, which is filtered by the Fourier transform of the aperture function, $\hat{a}$. This corresponds to illuminating a thin sample ψ (centered at position p) with an illumination probe field, $\hat{a}$. These two simple functional transformations lead to a linear relationship between standard and Fourier ptychographic data [102]. To apply the algorithmic tools outlined next to standard ptychography, simply adhere to the following protocol wherever either variable appears: 1) replace the sample spectrum $\hat{\psi}$ with the sample function ψ, and 2) replace the aperture function a with the shape of the focused probe field that illuminates the sample, $\hat{a}$, in standard ptychography setups.

5.3 Results: Convex Lifted Ptychography (CLP)

5.3.1 The CLP solver

We begin the process of solving equation 5.2.2 as a convex program by expressing it in matrix form. First, we represent the unknown sample spectrum $\hat{\psi}$ as an $(n \times 1)$ vector. Again, n is the known sample resolution before it is reduced by the finite bandpass of the lens aperture. Second, the jth detected image becomes an $(m \times 1)$ vector $\mathbf{g}_j$, where again m is the number of pixels in each low-resolution image. The ratio n/m defines the ptychographic resolution improvement factor. It is equivalent to the largest angle of incidence from an off-axis optical source, divided by the acceptance angle of the imaging lens. Third, we express each lens aperture function $a(x + p_j)$ as an $(n \times 1)$ discrete aperture vector $\mathbf{a}_j$, which modulates the unknown sample spectrum $\hat{\psi}$.

To rewrite equation 5.2.2 as a matrix product, we define $\{\mathbf{A}_j\}_{j=1}^{q}$ to be the sequence of $(m \times n)$ rectangular matrices that contain a deterministic aperture function $\mathbf{a}_j$ along a diagonal. For an aberration-free rectangular aperture, each matrix $\mathbf{A}_j$ has a diagonal of ones originating at $(0, p'_j)$ and terminating at $(m, p'_j + m - 1)$, where p'_j is now a discretized version of our shift variable p_j. Finally, we introduce an $m \times m$ discrete Fourier transform matrix $\mathbf{F}^{(m)}$ to express the transformation of the low-pass filtered sample spectrum through our fixed imaging system for each low-resolution image $\mathbf{g}_j$:

$$\mathbf{g}_j = \left|\mathbf{F}^{(m)}\mathbf{A}_j\hat{\psi}\right|^2, \quad 1 \leq j \leq q. \tag{5.3.1}$$

Ptychography acquires a series of q images, $\{\mathbf{g}_j\}_{j=1}^{q}$. We combine this image set into a single vector by “stacking" all images in equation 5.3.1:

$$\mathbf{b} = \left|\mathbf{F}\mathbf{A}\hat{\psi}\right|^2 = \left|\mathbf{D}\hat{\psi}\right|^2. \tag{5.3.2}$$

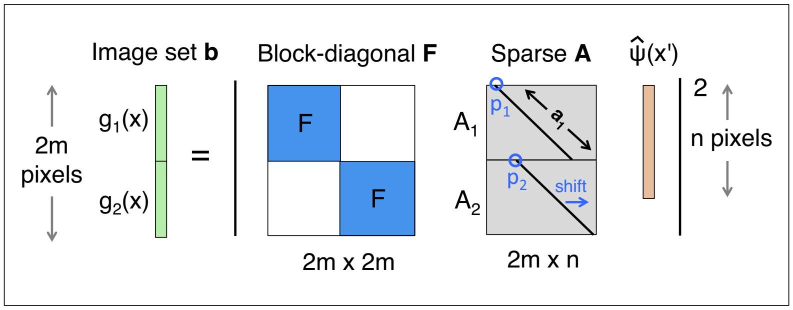


Figure 5.3.1: A set of images captured by Fourier ptychography stack together into a long data vector, **b**. Each associated matrix transform is similarly stacked and combined to form our final measurement matrix, **D** = **FA**. Here, we show stacking of just two images for simplicity. Typically, over 200 images are stacked.

Here, $\mathbf{b}$ is $\{\mathbf{g}\}$ expressed as a $(q \cdot m \times 1)$ stacked image vector (see figure 5.3.1). In addition, we define $\mathbf{D} = \mathbf{FA}$, where $\mathbf{F}$ is a $(q \cdot m \times q \cdot m)$ block diagonal matrix containing q copies of the low-resolution DFT matrices $\mathbf{F}^{(m)}$ in its diagonal blocks, and $\mathbf{A}$ has size $(q \cdot m \times n)$ and is formed by vertically stacking each aperture matrix $\mathbf{A}_j$:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(m)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{F}^{(m)} \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_q \end{pmatrix}. \tag{5.3.3}$$

We denote the transpose of the ith row of $\mathbf{D}$ as $\mathbf{d}_i$, which is a column vector. The set $\{\mathbf{d}_i\}$ forms our measurement vectors. The measured intensity in the ith pixel is the square of the inner product between $\mathbf{d}_i$ and the spectrum $\hat{\psi}$: $b_i = |\left\langle \mathbf{d}_i, \hat{\psi} \right\rangle|^2$. Next, we "lift" the solution $\hat{\psi}$ out of the quadratic relationship in equation 5.3.2. As suggested in [8], we may instead express it in the space of $(n \times n)$ positive-semidefinite matrices:

$$b_i = \operatorname{tr}\left(\hat{\psi}^* \mathbf{d}_i \mathbf{d}_i^* \hat{\psi}\right) = \operatorname{tr}\left(\mathbf{d}_i \mathbf{d}_i^* \hat{\psi} \hat{\psi}^*\right) = \operatorname{tr}\left(\mathbf{D}_i \mathbf{X}\right), \tag{5.3.4}$$

where $\mathbf{D}_i = \mathbf{d}_i \mathbf{d}_i^*$ is a rank-1 measurement matrix constructed from the ith measurement vector $\mathbf{d}_i$, $\mathbf{X} = \hat{\psi}\hat{\psi}^*$ is an $(n \times n)$ rank-1 outer product, and $1 \leq i \leq q \cdot m$. Equation 5.3.4 states that our quadratic image measurements $\{b_i\}_{i=1}^{q \cdot m}$ are linear transforms of $\hat{\psi}$ in a higher dimensional space. We may combine these $q \cdot m$ linear transforms into a single linear operator $\mathcal{A}$ to summarize the relationship between the stacked image vector $\mathbf{b}$ and the matrix $\mathbf{X}$ as, $\mathcal{A}(\mathbf{X}) = \mathbf{b}$.

One can now pose the phase retrieval problem in ptychography as a rank minimization procedure:

$$\begin{aligned} \text{minimize} \quad & \text{rank}(\mathbf{X}) \\ \text{subject to} \quad & \mathcal{A}(\mathbf{X}) = \mathbf{b}, \\ & \mathbf{X} \succeq 0, \end{aligned} \tag{5.3.5}$$

where $\mathbf{X} \succeq 0$ denotes $\mathbf{X}$ is positive-semidefinite. This rank minimization problem is not convex and is a computational challenge. Instead, adapting ideas from [72], we form a convex relaxation of equation 5.3.5 by replacing the rank of matrix $\mathbf{X}$ with its trace. This creates a convex semidefinite program:

$$\begin{aligned} \text{minimize} \quad & \text{tr}(\mathbf{X}) \\ \text{subject to} \quad & \mathcal{A}(\mathbf{X}) = \mathbf{b}, \\ & \mathbf{X} \succeq 0. \end{aligned} \tag{5.3.6}$$

Several recent results establish that the relaxation in equation 5.3.6 is equivalent to equation 5.3.5 under certain conditions on the operator $\mathcal{A}$ [186, 44]. Although not necessarily equivalent in general, this relaxation consistently offers us highly accurate experimental performance. To account for the presence of noise, we may reform equation 5.3.6 such that the measured intensities in $\mathbf{b}$ are no longer strictly enforced constraints, but instead appear in the objective function:

$$\begin{aligned} \text{minimize} \quad & \alpha \ \text{tr}(\mathbf{X}) + \frac{1}{2}\|\mathcal{A}(\mathbf{X}) - \mathbf{b}\| \\ \text{subject to} \quad & \mathbf{X} \succeq 0. \end{aligned} \tag{5.3.7}$$

Here, α is a scalar regularization variable that directly trades off goodness for complexity of fit. Its optimal value depends upon the assumed noise level. Equation 5.3.7 forms our final convex problem to recover a resolution-improved complex sample ψ from a set of obliquely illuminated images in $\mathbf{b}$. Many standard tools are available to solve this semidefinite program (see Appendix A). Its solution defines our Convex Lifted Ptychography (CLP) approach.

In practice, CLP returns a low-rank matrix $\mathbf{X}$, with a rapidly decaying spectrum,

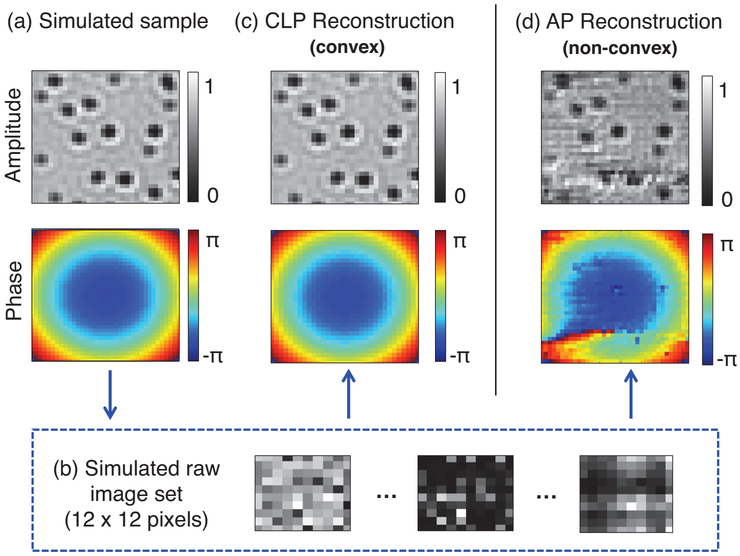


Figure 5.3.2: Simulation of the CLP algorithm. (a) An $n = 36 \times 36$ pixel complex sample (simulated) consisting of absorptive microspheres modulated with an independent quadratic phase envelope. (b) Sequence of low-resolution simulated intensity measurements ($m = 12 \times 12$ pixels each), serving as algorithm input. (c)-(d) Example CLP and AP reconstructions, where CLP is successful but AP converges to an incorrect local minimum. Here we use $q = 8^2$ images to achieve a resolution gain of 3 along each spatial dimension and simultaneously acquire phase.

as the optimal solution of equation 5.3.7. The trace term in the CLP objective function is primarily responsible for enforcing the low-rank structure of $\mathbf{X}$. While this trace term also appears like an alternative method to minimize the unknown signal energy, we caution that a fair interpretation must consider its effect in a lifted ($n \times n$) solution space. We obtain our final complex image estimate ψ by first performing a singular value decomposition of $\mathbf{X}$. Given low-noise imaging conditions and spatially coherent illumination, we set ψ to the Fourier transform of the largest resulting singular vector. Viewed as an autocorrelation matrix, we may also find useful statistical measurements within the remaining smaller singular vectors of $\mathbf{X}$. We note that one may also identify $\mathbf{X}$ as the discrete mutual intensity matrix of a partially coherent optical field: $\mathbf{X} = \left\langle \hat{\psi}\hat{\psi}^* \right\rangle$, where $\langle\rangle$ denotes an ensemble average [169]. Under this interpretation, equation 5.3.7 becomes an alternative solver for the stationary mixed states of a ptychography setup [224].

Without any further modification, three points distinguish equation 5.3.7 from existing AP-based ptychography solvers. First, the convex solver has a larger search space. If AP is used to iteratively update an n-pixel estimate, equation 5.3.7 must solve for an $n \times n$ positive-semidefinite matrix. Second, this boost in the solution space dimension guarantees the convex program may find its global optimum with tractable computation. This allows CLP to avoid AP's frequent convergence to local minima (i.e., failure to approach the true image). Unlike prior solvers for the ptychography problem, no local minima exist in the CLP approach. However, CLP cannot yet claim a single global minimum, since it is not necessarily a strictly convex program. Finally, equation 5.3.7 implicitly considers the presence of noise by offering a parameter (α) to tune with an assumed noise level. AP-based solvers lack this parameter and can be easily led into incorrect local minima by even low noise levels, which we demonstrate next.

5.3.2 CLP simulations and noise performance

We simulate Fourier ptychography following the setup in figure 5.1.1. We capture multiple two-dimensional images in (x, y) from a three-dimensional optical geometry. The simulated FP setup contains a detector with $m = 12^2$ pixels that are each 4 μm wide, a 0.1 numerical aperture (NA) lens at plane $A(x', y')$ (6° collection angle, unity magnification), and an array of spatially coherent optical sources at plane $L(x', y')$ (632 nm center wavelength, 10 nm spectral bandwidth). The array is designed to offer an illumination NA of 0.2 ($\theta_{max} = 11.5°$ maximum illumination angle). Together, the lens and illumination NAs define the reconstructed resolution of our complex sample as $n = 36^2$ pixels, increasing the pixel count of one raw image by a factor $n/m = 9$.

Figure 5.3.2(b) shows example simulated raw images from a sample of absorptive microspheres modulated by a quadratic phase envelope. Within each raw image, the set of microspheres is not clearly resolved. Here, we simulate the capture of $q = 8^2$ low resolution images, each uniquely illuminated from one of $q = 8^2$ optical sources in the square array. We then input this image set into both the standard AP algorithm (i.e., the PIE strategy) [71], as well as CLP in equation 5.3.7, to recover a high resolution (36×36 pixel) complex sample. Here, we select the PIE strategy as our comparison benchmark for two reasons. First, it is one of the most widely used ptychography algorithms. Second, similar to CLP, its structure implicitly assumes a Gaussian noise model [81]. Even in the noiseless case, 5 iterations of nonlinear AP introduces unpre-

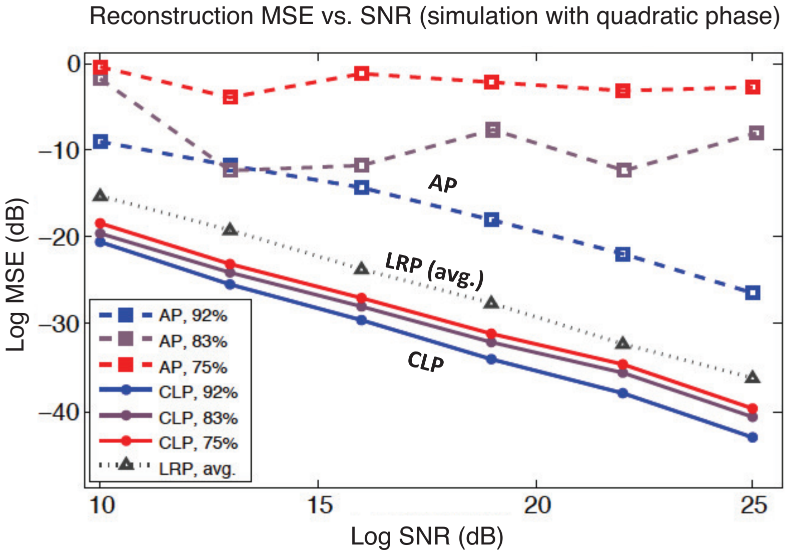


Figure 5.3.3: Reconstruction MSE versus signal to noise ratio (SNR) of CLP and AP (log scale, dB). Each curve represents reconstruction with a different number of captured images, q, corresponding to a different percentage of spectrum overlap (ol, noted in legend). Each point is an average over 5 independent algorithm runs with unique additive noise. Also included is the average performance of our LRP algorithm over the same 3 spectrum overlap settings (see Section 4).

dictable artifacts to both the recovered amplitude and phase (figure 5.3.2(d)), while CLP offers near perfect recovery (figure 5.3.2(c)). A constant phase offset is subtracted from both reconstructions for fair comparison, and we selected $\alpha = .001$.

Next, we quantify AP and CLP performance. We repeat the reconstructions in figure 5.3.2, again setting $\alpha = .001$ in equation 5.3.7 while varying two relevant parameters: the number of captured images q, and their signal-to-noise ratio (SNR). We define the SNR as $\text{SNR} = 10\log_{10}(\langle|\psi|^2\rangle \,/\, \langle|N^2|\rangle)$, where $\langle|\psi|^2\rangle$ is the mean sample intensity and $\langle|N^2|\rangle$ is the mean intensity of uniform Gaussian noise added to each simulated raw image. To account for the unknown constant phase offset in all phase retrieval reconstructions, we follow [140] and define our reconstruction mean-squared error as $\text{MSE} = \sum_x |\psi(x) - \rho s(x)|^2 \,/\, \sum_x |\psi(x)|^2$, where $\rho = \sum_x \psi(x)s^*(x)/\sum_x |s(x)|^2$ is a constant phase factor shifting our reconstructed phase to optimally match the known phase of the ground truth sample.

Figure 5.3.3 plots MSE as a function of SNR for this large set of CLP and AP reconstructions. Each of the algorithms' 3 independent curves simulates reconstruction using a different number of captured images, q. We summarize q by defining a Fourier spectrum overlap percentage: $ol = 1 - (n - m)/qm$. Each of the 6 points within one curve simulates a different level of additive measurement noise. Each point is an average over 5 independent trials. Since AP tends not to converge in the presence of noise, we represent each AP trial with the reconstruction that offers the lowest MSE across all iteration steps (up to 20 iterations). All CLP reconstructions improve linearly as SNR increases, while AP performance fluctuates unpredictably. For both algorithms, performance improves with increased spectrum overlap ol, and reconstruction fidelity quickly deteriorates and then effectively fails when ol drops below ~60%.

5.4 Results: Factorization for Low-Rank Ptychography (LRP)

Posing the inverse problem of ptychography as a semidefinite program (equation 5.3.7) is a good first step towards a more tractable solver. However, the constraint that $\mathbf{X}$ remain positive-semidefinite is computationally demanding: each iteration typically requires a full eigenvalue decomposition of $\mathbf{X}$. As the size of $\mathbf{X}$ scales with n^2, processable image sizes are limited to an order of 10^4 pixels on current desktop machines. This scaling limit does not prevent large-scale CLP processing of ptychography data. It is common practice to segment each detected image into as few as 10^3 pixels, process each segment separately, and then "tile" the resulting reconstructions back together into a final full-resolution solution [249]. CLP may also parallelize its computation with this strategy.

5.4.1 The LRP solver

While such tiling parallelization offers significant speedup, a simple observation helps avoid the poor scaling of CLP altogether: the desired solution of the ptychography problem in equation 5.3.5 is low-rank. Instead of solving for an $n \times n$ matrix $\mathbf{X}$, it is thus natural to adopt a low-rank ansatz and factorize the matrix $\mathbf{X}$ as $\mathbf{X} = \mathbf{R}\mathbf{R}^T$, where our decision variable $\mathbf{R}$ is now an $n \times r$ rectangular matrix containing complex entries, with $r < n$ [35, 36]. Inserting this factorization into our optimization problem in equation 5.3.6 and writing the constraints in terms of the measurement matrix

$\mathbf{D}_i = \mathbf{d}_i\mathbf{d}_i^T$ creates the non-convex program,

$$\begin{aligned} &\text{minimize} && \mathrm{tr}(\mathbf{R}\mathbf{R}^T) \\ &\text{subject to} && \mathrm{tr}(\mathbf{D}_i\mathbf{R}\mathbf{R}^T) = b_i \quad \text{for all } i. \end{aligned} \tag{5.4.1}$$

Besides removing the positive semidefinite constraint in equation 5.3.6, the factored form of equation 5.4.1 presents two more key adjustments to our original convex formulation. First, using the relationship $\mathrm{tr}(\mathbf{R}\mathbf{R}^T) = \|\mathbf{R}\|_F^2$, where F denotes a Frobenius norm, it is direct to rewrite the objective function and each constraint in equation 5.4.1 with just one $n \times r$ decision matrix, $\mathbf{R}$. Now instead of storing an $n \times n$ matrix like CLP, LRP must only store an $n \times r$ matrix. Since most practical applications of ptychography require coherent optics, the desired solution rank r will typically be close to 1, thus significantly relaxing storage requirements (i.e., coherent light satisfies $\mathbf{X} = \hat{\psi}\hat{\psi}^*$, so we expect $\mathbf{R}$ as a column vector and $\mathbf{R}\mathbf{R}^T$ a rank-1 outer product). Fixing r at a small value, LRP memory usage now scales linearly instead of quadratically with the number of reconstructed pixels, n.

Second, the feasible set of equation 5.4.1 is no longer convex. We thus must shift our solution strategy away from a simple semidefinite program. Prior work in [35, 36] suggests that an efficient and practically successful method of solving equation 5.4.1 is to minimize the following augmented Lagrangian function:

$$L(\mathbf{R}, \mathbf{y}, \sigma) = \mathrm{tr}(\mathbf{R}\mathbf{R}^T) - \sum\nolimits_i y_i \cdot \big(\mathrm{tr}(\mathbf{D}_i\mathbf{R}\mathbf{R}^T) - b_i\big) + \frac{\sigma}{2} \cdot \sum\nolimits_i \big(\mathrm{tr}(\mathbf{D}_i\mathbf{R}\mathbf{R}^T) - b_i\big)^2, \tag{5.4.2}$$

where $\mathbf{R} \in \mathbb{C}^{n\times r}$ is the unknown decision variable and the two variables $y \in \mathbb{R}^{q\cdot m}$ and $\sigma \in \mathbb{R}^+$ are new parameters to help guide our algorithm to its final reconstruction. The first term in equation 5.4.2 is the objective function from equation 5.4.1, indirectly encouraging a low-rank factorized product. This tracks our original assumption of a rank-1 solution within a “lifted" solution space. The second term contains the known equality constraints in equation 5.4.1 (i.e., the measured intensities), each assigned a weight y_i. The third term is a penalized fitting error that we abbreviate with label v. It is weighted by one penalty parameter σ, mimicking the role of a Lagrangian multiplier.

With an appropriate fixed selection of y_i's and σ, the minimization of $L(\mathbf{R}, \mathbf{y}, \sigma)$

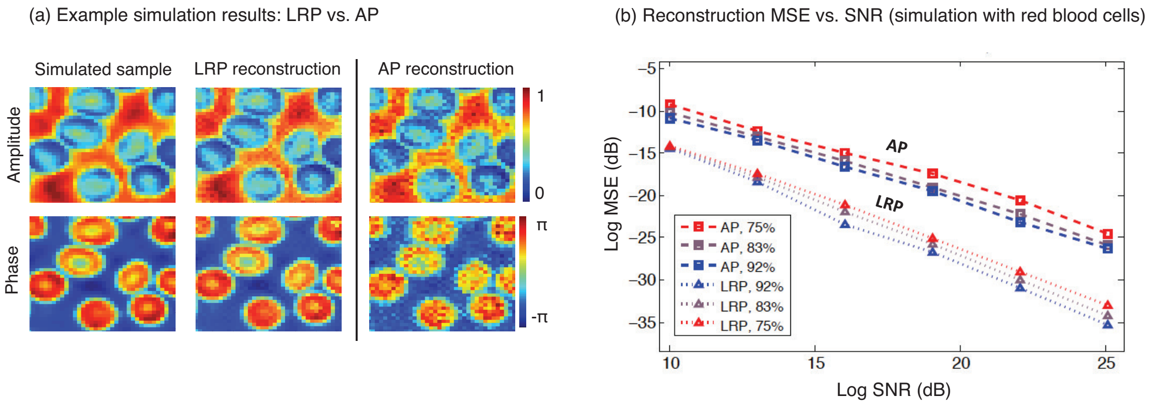
with respect to $\mathbf{R}$ identifies our desired optimum of equation 5.4.1. Specifically, if a local minimum of L is identified each iteration (which is nearly always the case in practice), then the minimization sequence accumulation point is a guaranteed solution [36]. As an unconstrained function, the minimum of L is quickly found via standard tools (e.g., a quasi-Newton approach such as the LBFGS algorithm [199]), as previously demonstrated across a wide range of applications and experiments [35].

The goal of our low-rank ptychography (LRP) algorithm thus reduces to the following task: determine a suitable set of (y_i, σ) such that we may minimize equation 5.4.2 with respect to $\mathbf{R}$, which leads to our desired solution. We use the iterative algorithm suggested in [35] to sequentially minimize L with respect to $\mathbf{R}^k$ at iteration k, and then update a new parameter set (y^{k+1}, σ^{k+1}) at iteration $k+1$. We update parameters (y^{k+1}, σ^{k+1}) to ensure their associated term's contribution to the summation forming L is relatively small. This suggests $\mathbf{R}^{k+1}$ is proceeding to a more feasible solution. The relative permissible size of the second and third terms in L are controlled by two important parameters, $\eta < 1$ and $\gamma > 1$: if the third term v sufficiently decreases such that $v^{k+1} \leq \eta v^k$, then we hold its multiplier σ fixed and update the equality constraint multipliers, y_i. Otherwise, we increase σ by a factor γ such that $\sigma^{k+1} = \gamma\sigma^k$. A detailed discussion of these algorithmic steps is in [35, 36].

We initialize the LRP algorithm with an estimate of the unknown high-resolution complex sample function ψ_0, contained within a low-rank matrix $\mathbf{R}^0$. We terminate the algorithm either if it reaches a sufficient number of iterations, or if the minimizer fulfills some convergence criterion. We form $\mathbf{R}^0$ using a spectral method, which can help increase solver accuracy and decrease computation time [41]. Specifically, we select the r columns of $\mathbf{R}^0$ as the leading r eigenvectors of $\mathbf{D}^*\text{diag}[\mathbf{b}]\mathbf{D}$, where $\mathbf{D}$ is the measurement matrix in equation 5.3.2. While this spectral approach works quite well in practice, a random initialization of $\mathbf{R}^0$ also often produces an accurate reconstruction.

5.4.2 LRP simulations and noise performance

Following the same procedure used to simulate the CLP algorithm, we test the MSE performance of the LRP algorithm as a function of SNR in figure 5.4.1. We again add different amounts of uncorrelated Gaussian noise to each simulated raw image


(a) Example simulation results: LRP vs. AP

(b) Reconstruction MSE vs. SNR (simulation with red blood cells)

Figure 5.4.1: Simulation of the LRP and AP algorithms using the same parameters as for figures 5.3.2–5.3.3, but now with a different “red blood cell" sample. (a) Using 8^2 simulated intensity measurements as input (SNR=19, 12^2 pixels each), both algorithms successfully recover each cell's phase, but AP is less accurate. (b) MSE versus SNR plot with varying amounts of noise added to the same data set. The MSE for LRP is ~5-10 dB lower than for AP across all noise levels and aperture overlap settings (each point from 5 independent trials).

set and compare the LRP reconstruction with ground truth. This simulated sample is the experimentally obtained amplitude and phase of a human blood smear. It is qualitatively similar to the sample used in figure 5.3.2. Unlike with the simulations in figures 5.3.2–5.3.3, the AP algorithm no longer malfunctions at lower spectrum overlap percentages (i.e., lower values of ol). Despite this apparent success, the MSE of the LRP minimizer is still ~5-10 dB better than the MSE of the AP minimizer, across all levels of SNR. This reduced LRP reconstruction error follows without any parameter optimization or explicit noise modeling.

In these simulations, we somewhat arbitrarily fix η and γ at 0.5 and 1.5, respectively, and set the desired rank of the solution, r, to 1. In practice, these free variables offer significant freedom to tune the response of LRP to noise. For example, similar to the noise parameter α in equation 5.3.7, the multiplier σ (controlled via γ) in equation 5.4.2 helps trade off complexity for goodness of fit by re-weighting the quadratic fitting error term.

In addition to reducing required memory, the LRP algorithm also improves upon the computational cost of CLP. For an n-pixel sample reconstruction, the per iteration cost of the CLP algorithm is currently $O(n^3)$, using big-O notation. The

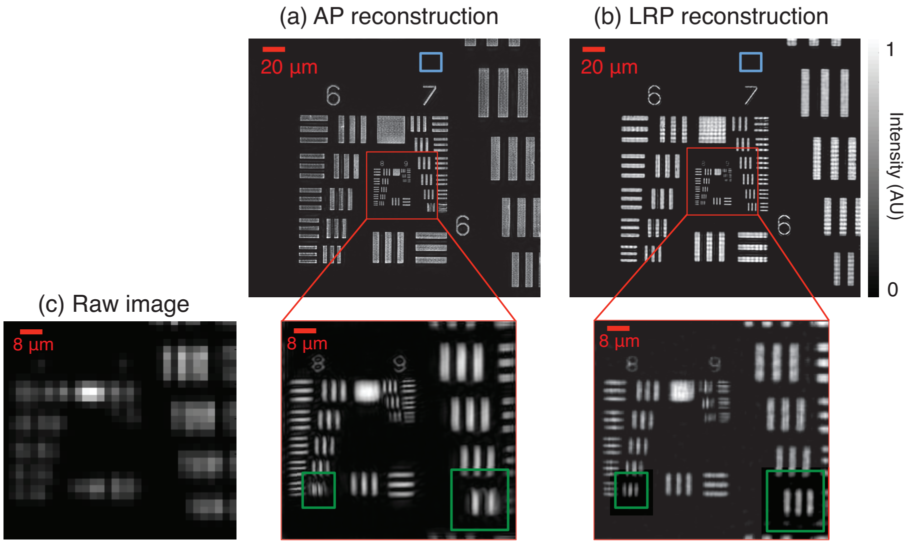


Figure 5.4.2: Experimental reconstruction of a USAF target, where the number of resolved pixels is increased by a factor of 25. We test two different ptychography algorithms: (a) AP and (b) LRP. Here we only show reconstructed intensity. LRP avoids artifacts (e.g., boxed in green) commonly encountered in the AP approach. Cited variances are measured in blue boxes (top). (c) Same cropped region of one low-resolution raw image, for comparison.

positive-semidefinite constraint in equation 5.3.7, which requires a full eigenvalue decomposition, defines this behavior limit. The per-iteration cost of the LRP algorithm, on the other hand, is $O(n \log n)$. This large per-iteration cost reduction is the primary source of LRP speedup. For example, LRP required ~21 seconds to complete an average simulation of the example in figure 5.3.2, while CLP required ~170 minutes and AP required 1 second on the same desktop machine.

5.5 Results: Experiment

We experimentally verify how LRP improves the accuracy and noise stability of ptychographic reconstruction using a Fourier ptychographic (FP) microscope. Our experimental procedure closely follows the protocol in [249]. While we demonstrate at optical wavelengths, it is straightforward to acquire a Fourier ptychographic data set in an X-ray or electron microscope (e.g., with a tilting source [90]). Alternatively, two trivial changes within equation 5.4.1 directly prepares standard ptychographic data for LRP processing (see end of section 2). Given its removal of local minima and improved treatment of noise, we expect our strategy will benefit both experimental arrangements.

In this section, we first quantitatively verify that LRP accurately measures high resolution and sample phase. Compared with AP reconstructions, our LRP algorithm generates fewer undesirable artifacts in experiment. Second, we will compare the AP and LRP reconstructions of a biological sample, which establishes the improved noise stability of our new algorithm.

5.5.1 Quantitative performance

Our FP microscope consists of a 15×15 array of surface-mounted LEDs (model SMD 3528, center wavelength λ=632 nm, 4 mm LED pitch, 150 μm active area diameter), which serve as our quasi-coherent optical sources. The array is placed l=80 mm beneath the sample plane, and each LED has an approximate 20 nm spectral bandwidth. Prior work establishes that the impact of non-ideal source coherence is gradual [102]. While negligible in these experiments, we may eventually account for source statistics in the multi-rank structure of the LRP optimizer $\mathbf{R}$.

To quantitatively verify resolution improvement, we turn on each of the 15×15 LEDs beneath a U.S. Air Force (USAF) resolution calibration target. A 2X Olympus microscope objective (apochromatic Plan APO 0.08 NA) transfers each resulting optical field to a CCD detector (Kodak KAI-29050, 5.5 μm pixels), creating 225 low resolution images. Using this 0.08 NA microscope objective (5° collection angle) and a 0.35 illumination NA ($\theta_{max} = 20°$ illumination angle), our FP microscope offers a total complex field resolution gain of $n/m = 25$. Each image spectrum overlaps by $ol \approx 70\%$ in area with each neighboring image spectrum.

For reconstruction, we select $n = 25 \cdot m$ and use the same aperture parameters with AP and LRP to create the high-resolution images in figure 5.4.2. For computational efficiency, we segment each low-resolution image into 3×3 tiles (n=480^2 per tile) and process the tiles in parallel, as performed in [249]. We determine the optimal number of AP and LRP algorithm iterations as 6 and 15, respectively, and fixed this for each tile (and all subsequent reconstructions). We typically initialize LRP with the following parameters: γ=1.5, η=0.3, y^0=10, and σ^0=10. We determine the microscope aperture function with an iterative procedure [168] before each experiment and fix it for each algorithm trial.

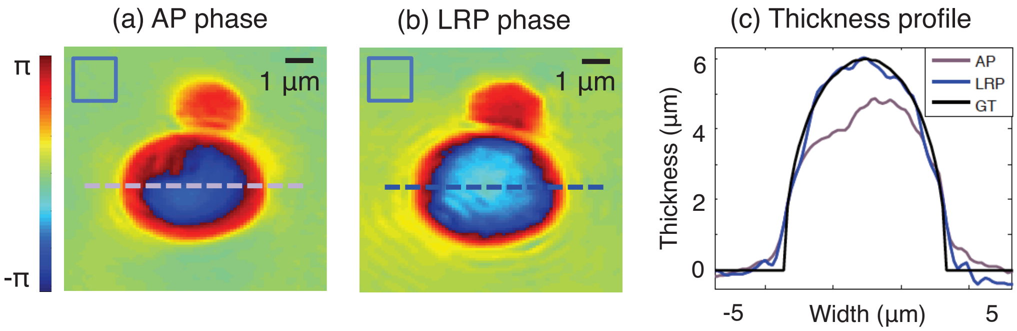


Figure 5.5.1: Experimental measurement of the quantitative optical phase emerging from two polystyrene microspheres. Both (a) AP and (b) LRP reconstruct phase maps that appear qualitatively similar, although the AP phase map flattens at the sphere's center. Variances measured in blue boxes. (c) Plot of microsphere thickness from a trace through the center of the large sphere (dashed line) demonstrates close agreement between LRP and ground truth (GT).

Both ∼1 megapixel reconstructions achieve their maximum expected resolving power (Group 9, Element 3: 1.56 μm line pair spacing). This is approximately 5 times sharper than the smallest resolved feature in one raw image (Group 7, Element 2 in Fig 5.4.2(c)). Our new LRP algorithm avoids certain artifacts that are commonly observed during the nonlinear descent of AP (boxed in green). Both reconstructions slowly fluctuate in background areas that we expect to be uniformly bright or dark. These fluctuations are caused in part by experimental noise, an imperfect aperture function estimate, and possible misalignments in the LED shift values, p_j. In a representative background area marked by a 40^2 pixel blue box in figure 5.4.2, AP and LRP exhibit normalized background amplitude variances of $\sigma_A^2 = 5.4 \times 10^{-4}$ and $\sigma_L^2 = 5.0 \times 10^{-4}$, respectively. Accounting for experimental uncertainty in the aperture function a and shifts p_j (e.g., following [101, 168]) may reduce this error in both algorithms.

To verify that our LRP solver reconstructs quantitatively accurate phase, we next image a monolayer of polystyrene microspheres (index of refraction $n_m = 1.587$) immersed in oil ($n_o = 1.515$, both indexes for $\lambda = 632$ nm light). To demonstrate the LRP algorithm easily generalizes to any ptychographic arrangement, we perform this experiment on a new “high-NA" FP microscope. The high-NA setup uses a larger 0.5 NA microscope objective lens with a 30° collection angle (20X Olympus 0.5 NA UPLFLN). For sample illumination, we now arrange 28 LEDs into 3 concentric rings of 8, 8, and 12 evenly spaced light sources (ring radii=16, 32, and 40 mm, respectively). We place this new light source array 40 mm beneath the sample to create a 0.7 illumination NA with a $\theta_{max} = 45°$ illumination angle. The synthesized numerical aperture of this new FP microscope, computed as the sum of the illumination NA and objective lens NA, is $\mathrm{NA}_s = 1.2$. With a greater-than-unity synthetic NA, our reconstructions can offer oil-immersion quality resolution (~385 nm smallest resolvable feature spacing [167]), without requiring any immersion medium between the sample and objective lens.

Using the same data and parameters for AP and LRP input, we obtain the high-resolution phase reconstructions of two adjacent microspheres in figure 5.5.1 (3 μm and 6 μm diameters). For this reconstruction, we set m=160^2 and n=320^2. We have subtracted a constant phase offset from the LRP solution in (b) to allow for direct comparison to the AP solution in (a). The two reconstructions appear qualitatively similar except at the center of the 6 μm sphere, where the AP phase profile unexpectedly flattens. We highlight this flattening by selecting phase values along each marked dashed line to plot the resulting sample thickness in figure 5.5.1(c). Phase φ and sample thickness t are related via $t = k\Delta\varphi(n_m - n_o)^{-1}$, where k is the average wavenumber and $\Delta\varphi = \varphi - \varphi_0$ is the reconstructed phase minus a constant offset. LRP closely matches the optical thickness of a ground-truth sphere (GT, black curve): the length of the vertical chord connecting the top and bottom arcs of a 6 μm diameter circle. The normalized amplitude variances from a 40^2-pixel background area are $\sigma_A^2 = 9.2 \times 10^{-4}$ and $\sigma_L^2 = 5.8 \times 10^{-4}$, respectively. This again supports our observation that the high resolution reconstructions formed by LRP are more accurate than those formed by AP.

5.5.2 Biological sample reconstruction

Our third imaging example uses the same high-NA FP configuration (collection angle= 30°, $\theta_{max} = 45°$) to resolve a biological phenomenon: the infectious spread of malaria in human blood. The early stages of a *Plasmodium falciparum* infection in erythrocytes (i.e., red blood cells) includes the formation of small parasitic “rings". It is challenging to resolve these parasites under a microscope without using an immersion medium, even after appropriate staining. Oil-immersion is required for an accurate diagnosis

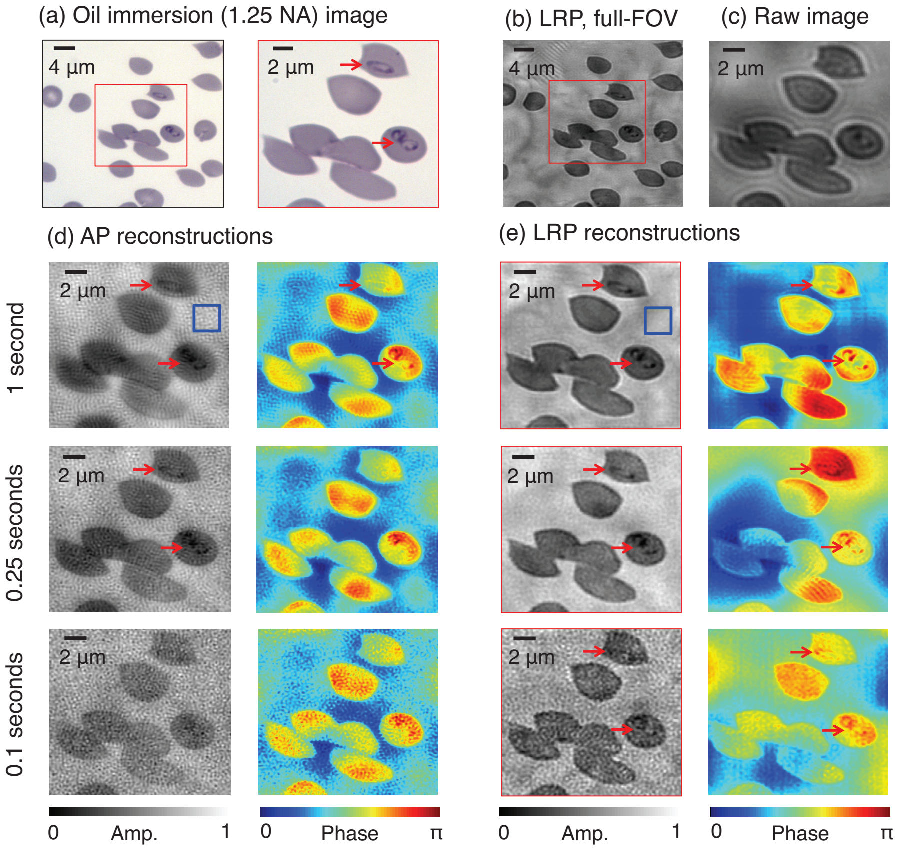


Figure 5.5.2: Experimental reconstruction of malaria-infected human red blood cells. (a) Oil immersion microscope image (1.25 NA) identifies two infected cells of interest (marked with arrows). (b) Example LRP reconstruction (area of interest in red box). (c) One example raw image used for AP and LRP algorithm input. (d) AP-reconstructed amplitude and phase from three different 29-image data sets, using 1 sec (top), 0.25 sec (middle), and 0.1 sec (bottom) exposure times for all images in each set. Variances measured within blue boxes. Increased noise within short-exposure images deteriorates reconstruction quality until both parasites are not resolved. (e) LRP reconstructions using the same three data sets. Both parasites are clearly resolved in the reconstructed phase for all three exposure levels.

of infection [166].

We use FP to resolve *Plasmodium falciparum*-infected cells with a 0.5 NA objective lens and using no oil in figure 5.5.2. We first prepare an infected blood sample following the protocol in [238]: we maintain erythrocyte asexual stage cultures of the P. falciparum strain 3D7 in culture medium, then we smear, fix with methanol, and apply a Hema 3 stain. An example sample region containing two infected cells, imaged with a conventional high-NA oil-immersion microscope (NA = 1.25) under Kohler illumination, is in figure 5.5.2(a).

Next, we capture 28 uniquely illuminated images of these two infected cells using our high-NA FP microscope. Figure 5.5.2(c) contains an example normally illuminated raw image, which does not clearly resolve the parasite infection. Figure 5.5.2(d) presents phase retrieval reconstructions using the standard AP algorithm, where we set m=120^2, n=240^2, run 6 iterations, and again subtract a constant phase offset. We include reconstructions from three data sets: images captured with a 1 second exposure (top), a 0.25 second exposure (middle), and 0.1 second exposure (bottom). A shorter exposure time implies increased noise within each raw image. While the 1 sec exposure-based AP reconstruction resolves each parasite, blurred cell boundaries and non-uniform fluctuations in amplitude suggest an inaccurate AP convergence. However, both parasite infections remain visible within the reconstructed phase. The parasites become challenging to resolve within the phase from 0.25 sec exposure data, and are not resolved within the phase from the 0.1 sec exposure data, due to increased image noise. The normalized background variance of each AP amplitude reconstruction, from a representative 40^2-pixel window (marked blue square), is $\sigma_A^2 = .0020$, .0027, and .0059 for the 1 sec, 0.25 sec, and 0.1 sec exposure reconstructions, respectively.

For comparison, reconstructions using our LRP algorithm are shown in figure 5.5.2(e) (sharpest solutions after 15 iterations). For each reconstructed amplitude, we set the desired solution rank to $r = 3$. We add the 3 modes of the resulting reconstruction in an intensity basis to create the displayed amplitude images. For each reconstructed phase, we set the desired solution matrix rank to $r = 1$ and leave all other parameters unchanged. For all three exposure levels, the amplitude of the cell boundaries remains sharper than in the AP images. Both parasite infections are resolvable in either the reconstructed amplitude or phase, or both, for all three exposure levels. The normalized amplitude variances from the same background window are now $\sigma_L^2 = .0016$ (1 sec), .0022 (0.25 sec), and .0035 (0.1 sec), an average reduction (i.e., improvement) of 26% with respect to the AP results. While not observed within our previous simulations or experiments, the AP reconstructions here offer a generally flatter background phase profile than LRP (i.e., less variation at low spatial frequencies). Without additional filtering or post-processing, the AP algorithm here might offer superior quantitative analysis during e.g. tomographic cell reconstruction, where low-order phase variations must remain accurate. However, it is clear within figure 5.5.2 that LRP better resolves the fine structure of each infection, which is critical during malaria diagnosis. A shorter image exposure time (i.e., up to 10 times shorter) may still enable accurate infection diagnosis when using LRP, as opposed to the standard AP approach.

5.6 Discussion and Conclusion

Through the relaxation in equation 5.3.6, we first transform the traditionally nonlinear phase retrieval process for ptychography into a convex program. We may now use well-established optimization tools to find the ptychography problem's global minimum. Then, we suggest a practically efficient algorithm to solve the resulting semidefinite program with an appropriate factorization. The result is a new ptychographic image recovery algorithm that is robust to noise. We demonstrate its successful performance in three unique experiments, concluding with a practical biological imaging scenario: the identification of malaria infection without using an oil immersion medium and under short-exposure imaging conditions.

Much future work remains to fully explore the specific benefits of our problem reformulation. Besides removing local minima from the recovery process, perhaps the most significant departure from prior phase retrieval solvers is a tunable solution rank, r. As noted earlier, r connects to statistical features of the ptychographic experiment, typically arising from the partial coherence of the illuminating field. Coherence effects are significant at third-generation X-ray synchrotron sources and within electron microscopes. An appropriately selected r may eventually help LRP measure the partial coherence of such sources, as outlined in [224]. The solution rank may also help identify setup vibrations, sample auto-fluorescence, or even 3D sample structure. As in prior work with low-rank matrix optimization, we may also artificially enlarge our solution rank to encourage the transfer of experimental noise into its smaller singular vectors.

Other extensions of LRP include simultaneously solving for unknown aberrations (i.e., the shape of the probe in standard ptychography), systematic setup errors, and inserting additional sample priors such as sparsity. These refinements are currently a critical component of ptychographic recovery in the fields of X-ray and electron microscopy, and will also improve our optical results. Along with algorithm refinement, a detailed comparison between LRP and various other recovery methods, especially under different sources of noise and error, will help identify the experimental conditions under which our strategy is of greatest benefit. What's more, as a particular solution to the general problem of phaseless measurement, our findings can also inform a wide variety of coherent diffractive imaging techniques. Regardless of the specific experimental application, convex analysis will continue to provide useful theoretical guarantees regarding phase retrieval algorithm performance, a crucial feature missing from current nonlinear solvers.

Appendix A. Computational specifics

We performed all processing on a high-end desktop containing two Intel Xeon 2.0 GHz CPUs and two 3GB GeForce GTX GPUs. Code was written in Matlab with built-in GPU acceleration. We solved our CLP semidefinite program using the TFOCS code package [14]. Our LRP algorithm borrows concepts from the LBFGS solver in [199] for one specific minimization step. LRP's total recovery time for the 1 megapixel example in Fig. 5.4.2 was approximately 130 seconds, while AP completed in approximately 15 seconds on the same desktop.

Bibliography

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

[2] G. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge University Press, 2010. No. 118.

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

[4] L. Arnold. On Wigner's semicircle law for the eigenvalues of random matrices. *Probability Theory and Related Fields*, 19(3):191–198, 1971.

[5] Z. Bai and Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The annals of Probability*, 21:1276–1294, 1993.

[6] Z. D. Bai and J. W. Silverstein. *Spectral Analysis of Large-Dimensional Random Matrices*. Springer, New York, NY, 2010.

[7] Z. D. Bai and Y. Q. Yin. Convergence to the semicircle law. *The Annals of Probability*, pages 863–875, 1988.

[8] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

[9] Z. Bao, G. Pan, W. Zhou, et al. Universality for the largest eigenvalue of sample covariance matrices with general population. *The Annals of Statistics*, 43(1):382–421, 2015.

[10] A. D. Barbour and L. H. Y. Chen. *An introduction to Stein's method*, volume 4. World Scientific, 2005.

[11] A. Barvinok. Measure concentration. Available at `http://www.math.lsa.umich.edu/~barvinok/total710.pdf`, 2005.

[12] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm, and fienup variants: a view from convex optimization. *JOSA A*, 19(7):1334–1345, 2002.

[13] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Hybrid projection–reflection method for phase retrieval. *JOSA A*, 20(6):1025–1034, 2003.

[14] S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[15] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[16] S. Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

[17] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[18] R. Bhatia. *Positive Definite Matrices.* Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007.

[19] P. J. Bickel and E. Levina. Some theory for Fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010, 2004.

[20] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008.

[21] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[22] P. Billingsley. *Probability and Measure.* Wiley, New York, NY, 1979.

[23] S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[24] S. G. Bobkov and M. Ledoux. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.*, 156(2):347–365, 1998.

[25] C. Borell. The Brunn-Minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

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

[27] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

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

[29] S. Boucheron, G. Lugosi, and P. Massart. On concentration of self-bounding functions. *Electron. J. Probab.*, 14(64):1884–1899, 2009.

[30] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities.* Oxford University Press, 2013.

[31] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[32] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2009.

[33] L. M. Brègman. Relaxation method for finding a common point of convex sets and its application to optimization problems. *Dokl. Akad. Nauk SSSR*, 171:1019–1022, 1966.

[34] 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.

[35] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[36] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[37] D. Burkholder, B. Davis, and R. Gundy. Integral inequalities for convex functions of operators on martingales. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob*, volume 2, pages 223–240, 1972.

[38] D. Burkholder and R. Gundy. Extrapolation and interpolation of quasi-linear operators on martingales. *Acta mathematica*, 124(1):249–304, 1970.

[39] D. L. Burkholder. Distribution function inequalities for martingales. *the Annals of Probability*, pages 19–42, 1973.

[40] T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.,*, 38(4):2118–2144, 2010.

[41] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[42] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[43] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[44] E. J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[45] E. Carlen. Trace inequalities and quantum entropy: an introductory course. *Contemp. Math.*, 529:73–140, 2010.

[46] 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.

[47] D. Chafaï. Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.*, 44(2):325–363, 2004.

[48] D. Chafaï. Binomial-Poisson entropic inequalities and the $m/m/\infty$ queue. *ESAIM Probab. Stat.*, 10:317–339, 2006.

[49] H. N. Chapman and K. A. Nugent. Coherent lensless x-ray imaging. *Nature Photonics*, 4(12):833–839, 2010.

[50] S. Chatterjee. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138(1-2):305–321, 2007.

[51] S. Chatterjee. *Concentration inequalities with exchangeable pairs*. PhD thesis, Stanford University, Palo Alto, Feb. 2008.

[52] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, page ias001, 2012.

[53] R. Y. Chen, A. A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis via the matrix laplace transform. 2012.

[54] 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.

[55] H.-C. Cheng and M.-H. Hsieh. New characterizations of matrix φ-entropies, Poincaré and Sobolev inequalities and an upper bound to Holevo quantity. *arXiv preprint arXiv:1506.06801*, 2015.

[56] H.-C. Cheng, M.-H. Hsieh, and M. Tomamichel. Exponential decay of matrix φ-entropies on Markov semigroups with applications to dynamical evolutions of quantum ensembles. *arXiv preprint arXiv:1511.02627*, 2015.

[57] D. Christofides and K. Markström. Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures Algorithms*, 32:88–100, 2008.

[58] I. Csiszár. A class of measures of informativity of observation channels. *Period. Math. Hungar.*, 2:191–213, 1972. Collection of articles dedicated to the memory of Alfréd Rényi.

[59] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[60] E. B. Davies and B. Simon. Ultracontractivity and the heat kernel for Schrödinger operators and Dirichlet Laplacians. *J. Funct. Anal.*, 59:335–395, 1984.

[61] C. De Sa, K. Olukotun, and C. Ré. Global convergence of stochastic gradient descent for some nonconvex matrix problems. *arXiv preprint arXiv:1411.1134*, 2014.

[62] P. Deift. Universality for mathematical and physical systems. In *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*, pages 125–152, 2007.

[63] A. Dembo. Information inequalities and concentration of measure. *The Annals of Probability*, pages 927–939, 1997.

[64] M. Dierolf, A. Menzel, P. Thibault, P. Schneider, C. M. Kewish, R. Wepf, O. Bunk, and F. Pfeiffer. Ptychographic x-ray computed tomography at the nanoscale. *Nature*, 467(7314):436–439, 2010.

[65] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[66] V. Elser, I. Rankenburg, and P. Thibault. Searching with iterated maps. *Proceedings of the National Academy of Sciences*, 104(2):418–423, 2007.

[67] L. Erdös. Universality of wigner random matrices: a survey of recent results. *Russian Mathematical Surveys*, 66(3):507, 2011.

[68] L. Erdös, S. Péché, J. A. Ramírez, B. Schlein, and H. T. Yau. Bulk universality for Wigner matrices. *Communications on Pure and Applied Mathematics*, 63(7):895–925, 2010.

[69] L. Erdös, J. Ramírez, B. Schlein, T. Tao, V. H. Vu, and H.-T. Yau. Bulk universality for wigner hermitian matrices with subexponential decay. *Mathematical research letters*, 17(4):667–674, 2010.

[70] E. F. Fama and K. R. French. The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, 18:25–46, 2004.

[71] H. Faulkner and J. Rodenburg. Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm. *Physical review letters*, 93(2):023903, 2004.

[72] M. Fazel. *Matrix rank minimization with applications.* PhD thesis, PhD thesis, Stanford University, 2002.

[73] J. Fienup and C. Wackerman. Phase-retrieval stagnation problems and solutions. *JOSA A*, 3(11):1897–1907, 1986.

[74] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[75] T. Figiel, P. Hitczenko, W. Johnson, G. Schechtman, and J. Zinn. Extremal properties of rademacher functions with applications to the khintchine and rosenthal inequalities. *Transactions of the American Mathematical Society*, 349(3):997–1027, 1997.

[76] D. A. Freedman. *Statistical Models: Theory and Practice.* Cambridge Univ. Press, Cambridge, 2005.

[77] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivar. Anal.*, 98(2):227–255, 2007.

[78] R. Gerchberg. Holography without fringes in the electron microscope. 1972.

[79] J. Ginibre. Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics*, 6(3):440–449, 1965.

[80] A. Gittens and J. A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *arXiv preprint arXiv:1104.4513*, 2011.

[81] P. Godard, M. Allain, V. Chamard, and J. Rodenburg. Noise models for low counting rate coherent diffraction imaging. *Optics express*, 20(23):25914–25934, 2012.

[82] T. Godden, R. Suman, M. Humphry, J. Rodenburg, and A. Maiden. Ptychographic microscope for three-dimensional imaging. *Optics express*, 22(10):12513–12523, 2014.

[83] F. Götze, A. Tikhomirov, et al. The circular law for random matrices. *The Annals of Probability*, 38(4):1444–1491, 2010.

[84] N. Gozlan and C. Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

[85] M. Gromov. Metric structures for Riemannian and non-Riemannian spaces, Based on the 1981 French original, With appendices by M. Katz, P. Pansu and S. Semmes. Translated from the French by Sean Michael Bates. *Progress in Mathematics*, 152, 1999.

[86] L. Gross. Logarithmic Sobolev inequalites. *Amer. J. Math.*, 97:1061–1083, 1975.

[87] A. Guionnet and B. Zegarlinksi. *Lectures on logarithmic Sobolev inequalities.* Springer, 2003.

[88] U. Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 3(70):231–283, 1981.

[89] U. Haagerup and M. Musat. On the best constants in noncommutative khintchine-type inequalities. *Journal of Functional Analysis*, 250(2):588–624, 2007.

[90] S. J. Haigh, H. Sawada, and A. I. Kirkland. Atomic structure imaging beyond conventional resolution limits in the transmission electron microscope. *Physical review letters*, 103(12):126101, 2009.

[91] 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.

[92] F. Hansen. Trace functions with applications in quantum physics. *Journal of Statistical Physics*, 154(3):807–818, 2014.

[93] F. Hansen and Z. Zhang. Characterisation of matrix entropies. *arXiv preprint arXiv:1402.2118*, 2014.

[94] D. L. Hartl, A. G. Clark, and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[95] F. Hiai and D. Petz. *The semicircle law, free random variables and entropy.* American Mathematical Society Providence, 2000.

[96] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[97] A. S. Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.

[98] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge Univ. Press, Cambridge, 1994.

[99] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge university press, 2012.

[100] R. Horstmeyer, R. Y. Chen, X. Ou, B. Ames, J. A. Tropp, and C. Yang. Solving ptychography with a convex relaxation. *New Journal of Physics*, 17(5):053044, 2015.

[101] R. Horstmeyer, X. Ou, J. Chung, G. Zheng, and C. Yang. Overlapped fourier coding for optical aberration removal. *Optics express*, 22(20):24062–24080, 2014.

[102] R. Horstmeyer and C. Yang. A phase space model of fourier ptychographic microscopy. *Optics express*, 22(1):338–358, 2014.

[103] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices. Available at arXiv:1104.1672, 2011.

[104] F. Hüe, J. Rodenburg, A. Maiden, F. Sweeney, and P. Midgley. Wave-front phase retrieval in transmission electron microscopy via ptychography. *Physical Review B*, 82(12):121415, 2010.

[105] R. Ibragimov and S. Sharakhmetov. Short communications: On an exact constant for the Rosenthal inequality. *Theory of Probability & Its Applications*, 42(2):294–302, 1998.

[106] R. Ibragimov and S. Sharakhmetov. The exact constant in the Rosenthal inequality for random variables with mean zero. *Theory of Probability & Its Applications*, 46(1):127–132, 2002.

[107] K. J. Johansson. Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. *Communications in Mathematical Physics*, 215(3):683–705, 2001.

[108] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ, 6th edition, 2002.

[109] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, 2002.

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

[111] M. Junge and Q. Xu. On the best constants in some non-commutative martingale inequalities. *Bulletin of the London Mathematical Society*, 37(02):243–253, 2005.

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

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

[114] M. Junge and Q. Zeng. Noncommutative martingale deviation and poincaré type inequalities with applications. *Probability Theory and Related Fields*, 161(3-4):449–507, 2014.

[115] M. Junge, Q. Zeng, et al. Noncommutative bennett and rosenthal inequalities. *The Annals of Probability*, 41(6):4287–4316, 2013.

[116] N. E. Karoui. Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, pages 663–714, 2007.

[117] N. E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008.

[118] A. Khinchin. Uber dyadische briiche. *Math. Z*, 18:109–116, 1923.

[119] J. Kiefer. On large deviations of the empiric df of vector chance variables and a law of the iterated logarithm. *Pacific journal of mathematics*, 11(2):649–660, 1961.

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

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

[122] R. Latała and K. Oleszkiewicz. Between Sobolev and Poincaré. In *Geometric aspects of functional analysis*, volume 1745 of *Lecture Notes in Math.*, pages 147–168. Springer, Berlin, 2000.

[123] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic), 1995/97.

[124] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 120–216. Springer, Berlin, 1999.

[125] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2001.

[126] M. Ledoux. Concentration, transportation and functional inequalities. *Preprint*, 2002.

[127] M. Ledoux, I. Nourdin, and G. Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2014.

[128] E. Levina and R. Vershynin. Partial estimation of covariance matrices. *Probab. Theory Related Fields*, 2011.

[129] P. Levy. Problemes concrets d'analyse fonctionnelle. Gauthier-Villars (1951). *Mathematical Reviews (MathSciNet): MR41346 Zentralblatt MATH*, 43.

[130] P. Levy. Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*, 1954.

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

[132] E. H. Lieb et al. Some convexity and subadditivity properties of entropy. *Bulletin of the American Mathematical Society*, 81(1):1–13, 1975.

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

[134] J. Littlewood. On a certain bilinear form. *Quart. J. Math. Oxford Ser*, 1:164–174, 1930.

[135] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195(2):491–523, 2005.

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

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

[138] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Ann. Probab.*, 42(3):906–945, 2014.

[139] A. Maiden, M. Humphry, M. Sarahan, B. Kraus, and J. Rodenburg. An annealing algorithm to correct positioning errors in ptychography. *Ultramicroscopy*, 120:64–72, 2012.

[140] A. M. Maiden and J. M. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109(10):1256–1262, 2009.

[141] A. M. Maiden, J. M. Rodenburg, and M. J. Humphry. Optical ptychography: a practical implementation with useful resolution. *Optics letters*, 35(15):2585–2587, 2010.

[142] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483, 1967.

[143] S. Marchesini. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78(1):011301, 2007.

[144] S. Marchesini, A. Schirotzek, C. Yang, H.-t. Wu, and F. Maia. Augmented projections for ptychographic imaging. *Inverse Problems*, 29(11):115009, 2013.

[145] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, London, 1980.

[146] K. Marton. A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory*, 32:445–446, 1986.

[147] K. Marton. A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, 1996.

[148] K. Marton. Measure concentration and strong mixing. *Studia Scientiarum Mathematicarum Hungarica*, 40(1-2):1–2, 2003.

[149] K. Marton. Measure concentration for euclidean distance in the case of dependent random variables. *Annals of probability*, pages 2526–2544, 2004.

[150] K. Marton et al. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

[151] P. Massart. Rates of convergence in the central limit theorem for empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 381–423, 1986.

[152] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000.

[153] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. Probability theory.

[154] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[155] A. Maurer et al. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012.

[156] B. Maurey. Some deviation inequalities. *Geometric & Functional Analysis GAFA*, 1(2):188–197, 1991.

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

[158] S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. Available at arXiv:1112.5448, Jan. 2012.

[159] R. J. Muirhead. *Aspects of Multivariate Statistical Theory.* Wiley, New York, NY, 1982.

[160] S. V. Nagaev. Some refinements of probabilistic andmoment inequalities. *Theory of Probability & Its Applications*, 42(4):707–713, 1998.

[161] P. D. Nelliss, B. C. McCallum, and J. M. Rodenburg. Resolution beyond the 'information limit' in transmission electron microscopy. pages 630–632, 1995.

[162] K. A. Nugent. Coherent methods in the x-ray sciences. *Advances in Physics*, 59(1):1–99, 2010.

[163] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at arXiv:0911.0600, 2009.

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

[165] A. Onatski. The tracy–Widom limit for the largest eigenvalues of singular complex Wishart matrices. *The Annals of Applied Probability*, pages 470–490, 2008.

[166] W. H. Organization. *Malaria microscopy quality assurance manual.* World Health Organization, 2009.

[167] X. Ou, R. Horstmeyer, G. Zheng, and C. Yang. High numerical aperture fourier ptychography: principle, implementation and characterization. *Optics express*, 23(3):3472–3491, 2015.

[168] X. Ou, G. Zheng, and C. Yang. Embedded pupil function recovery for fourier ptychographic microscopy. *Optics express*, 22(5):4960–4972, 2014.

[169] H. M. Ozaktas, S. Yüksel, and M. A. Kutay. Linear algebraic theory of partial coherence: discrete fields and measures of partial coherence. *JOSA A*, 19(8):1563–1571, 2002.

[170] R. Paley and A. Zygmund. On some series of functions,(1). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 337–357. Cambridge Univ Press, 1930.

[171] G. Pan and W. Zhou. Circular law, extreme singular values and potential theory. *Journal of Multivariate Analysis*, 101(3):645–656, 2010.

[172] D. Paulin, L. Mackey, and J. A. Tropp. Deriving matrix concentration inequalities from kernel couplings. Available at arXiv:1305.0612, May 2013.

[173] D. Paulin, L. Mackey, and J. A. Tropp. Efron-stein inequalities for random matrices. *arXiv preprint arXiv:1408.3470*, 2014.

[174] D. Petz. *A survey of certain trace inequalities*, volume 30 of *Banach Center Publ.*, pages 287–298. Polish Acad. Sci., Warsaw, 1994.

[175] I. Pinelis and S. Utev. Estimates of the moments of sums of independent random variables. *Theory of Probability & Its Applications*, 29(3):574–577, 1985.

[176] M. Pinsker. Information and information stability of random variables and processes. *The American Mathematical Monthly*, (73), 1966.

[177] G. Pisier and E. Ricard. The non-commutative Khintchine inequalities for $0 < p < 1$. *Journal of the Institute of Mathematics of Jussieu*, pages 1–21, 10 2015.

[178] G. Pisier and Q. Xu. Non-commutative martingale inequalities. *Communications in mathematical physics*, 189(3):667–698, 1997.

[179] J. Pitrik and D. Virosztek. On the joint convexity of the Bregman divergence of matrices. *Letters in Mathematical Physics*, 105(5):675–692, 2015.

[180] M. Raginsky. Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 419–423. IEEE, 2013.

[181] M. Raginsky. Strong data processing inequalities and *phi*-Sobolev inequalities for discrete channels. *arXiv preprint arXiv:1411.3575*, 2014.

[182] M. Raginsky and I. Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*. Now Publishers Inc., 2014.

[183] N. Randrianantoanina. Non-commutative martingale transforms. *Journal of Functional Analysis*, 194(1):181–212, 2002.

[184] N. Randrianantoanina et al. Conditioned square functions for noncommutative martingales. *The Annals of Probability*, 35(3):1039–1070, 2007.

[185] H. Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[186] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

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

[188] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, volume 1, pages 547–561, Berkeley, Calif., 1961. Univ. California Press.

[189] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Related Fields*, 119(2):163–175, 2001.

[190] J. Rodenburg, A. Hurst, A. Cullis, B. Dobson, F. Pfeiffer, O. Bunk, C. David, K. Jefimovs, and I. Johnson. Hard-x-ray lensless imaging of extended objects. *Physical review letters*, 98(3):034801, 2007.

[191] H. P. Rosenthal. On the subspaces ofl p (p> 2) spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

[192] A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104(485):177–186, 2009.

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

[194] M. Rudelson and R. Vershynin. The least singular value of a random square matrix is $O(n^{-1/2})$. *C. R. Math.*, 346:893–896, 2008.

[195] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

[196] G. Sadeghi and M. Sal Moslehian. Noncommutative martingale concentration inequalities. *Illinois Journal of Mathematics*, 58, 2014.

[197] J. Sanz, T. Huang, and T. Wu. A note on iterative fourier transform phase reconstruction from magnitude. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1251–1254, 1984.

[198] E. Schmidt. Die brunn-minkowskische ungleichung und ihr spiegelbild sowie die isoperimetrische eigenschaft der kugel in der euklidischen und nichteuklidischen geometrie. i. *Mathematische Nachrichten*, 1(2-3):81–157, 1948.

[199] M. Schmidt. The minfunc toolbox for matlab, 2006.

[200] Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(02):149–166, 2000.

[201] D. A. Shapiro, Y.-S. Yu, T. Tyliszczak, J. Cabana, R. Celestre, W. Chao, K. Kaznatcheev, A. D. Kilcoyne, F. Maia, S. Marchesini, et al. Chemical composition mapping with nanometre resolution by soft x-ray microscopy. *Nature Photonics*, 2014.

[202] Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics express*, 19(16):14807–14822, 2011.

[203] R. Speicher. Free probability theory and random matrices. In *Asymptotic Combinatorics with Applications to Mathematical Physics*, pages 53–73. Springer, 2003.

[204] N. Srivastava and R. Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111, 2013.

[205] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, Berkeley, Calif., 1972. Univ. California Press.

[206] D. W. Stroock and B. Zegarlinski. The equivalence of the logarithmic sobolev inequality and the dobrushin-shlosman mixing condition. *Communications in mathematical physics*, 144(2):303–323, 1992.

[207] S. Szarek. On the best constants in the khinchin inequality. *Studia Mathematica*, 2(58):197–208, 1976.

[208] S. J. Szarek. Spaces with large distance to l_∞^n and random matrices. *Amer. J. Math.*, 112:899–942, 1990.

[209] S. J. Szarek. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.

[210] M. Talagrand. An isoperimetric theorem on the cube and the Kintchine–Kahane inequalities. *Proceedings of the American Mathematical Society*, 104(3):905–909, 1988.

[211] M. Talagrand. Small tails for the supremum of a gaussian process. In *Annales de l'IHP Probabilités et statistiques*, volume 24, pages 307–315, 1988.

[212] M. Talagrand. A new isoperimetric inequality for product measure, and the concentration of measure phenomenon. In *Israel Seminar (GAFA)*, volume 1469 of *Lecture Notes in Math*, pages 91–124. Springer-Verlag, 1991.

[213] M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

[214] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Sci. Publ. Math.*, 81:73–205, 1995.

[215] M. Talagrand. A new look at independence. *The Annals of probability*, pages 1–34, 1996.

[216] M. Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

[217] T. Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[218] T. Tao and V. Vu. Random matrices: universality of local eigenvalue statistics. *Acta mathematica*, 206(1):127–204, 2011.

[219] T. Tao and V. Vu. The Wigner–Dyson–Mehta bulk universality conjecture for Wigner matrices. *Electron. J. Probab*, 16(77):2104–2121, 2011.

[220] T. Tao, V. Vu, et al. Random matrices: universality of local spectral statistics of non-Hermitian matrices. *The Annals of Probability*, 43(2):782–874, 2015.

[221] T. Tao, V. Vu, M. Krishnapur, et al. Random matrices: universality of ESDs and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010.

[222] P. Thibault, M. Dierolf, A. Menzel, O. Bunk, C. David, and F. Pfeiffer. High-resolution scanning x-ray diffraction microscopy. *Science*, 321(5887):379–382, 2008.

[223] P. Thibault and M. Guizar-Sicairos. Maximum-likelihood refinement for coherent diffractive imaging. *New Journal of Physics*, 14(6):063004, 2012.

[224] P. Thibault and A. Menzel. Reconstructing state mixtures from diffraction measurements. *Nature*, 494(7435):68–71, 2013.

[225] L. Tian, X. Li, K. Ramchandran, and L. Waller. Multiplexed coded illumination for fourier ptychography with an led array microscope. *Biomedical optics express*, 5(7):2376–2389, 2014.

[226] L. Tian and L. Waller. 3d intensity and phase imaging from light field measurements in an led array microscope. *Optica*, 2(2):104–111, 2015.

[227] N. Tomczak-Jaegermann. The moduli of smoothness and convexity and the Rademacher averages of the trace classes S_p $(1 \leq p < \infty)$. *Studia Mathematica*, 2(50):163–182, 1974.

[228] C. A. Tracy and H. Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994.

[229] C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

[230] J. A. Tropp. The random paving property for uniformly bounded matrices. *STUDIA MATHEMATICA*, 185:1, 2008.

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

[232] J. A. Tropp. From joint convexity of quantum relative entropy to a concavity theorem of lieb. *Proceedings of the American Mathematical Society*, 140(5):1757–1760, 2012.

[233] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[234] J. A. Tropp. The expected norm of a sum of independent random matrices: An elementary approach. *arXiv preprint arXiv:1506.04711*, 2015.

[235] J. A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[236] J. A. Tropp. Second-order matrix concentration inequalities. *arXiv preprint arXiv:1504.05919*, 2015.

[237] B. Tsirelson, I. Ibragimov, and V. Sudakov. Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory*, pages 20–41. Springer, 1976.

[238] D. L. van de Hoef, I. Coppens, T. Holowka, C. B. Mamoun, O. Branch, and A. Rodriguez. Plasmodium falciparum-derived uric acid precipitates induce maturation of dendritic cells. *PloS one*, 8(2):e55584, 2013.

[239] R. Vershynin. *Compressed Sensing: Theory and Applications*, chapter Introduction to the non-asymptotic analysis of random matrices. Cambridge Univ. Press, Cambridge, 2011. To appear. Available at `http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf`.

[240] C. Villani. *Topics in optimal transportation.* Number 58. American Mathematical Soc., 2003.

[241] D. Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1):201–220, 1991.

[242] D. Voiculescu. Lectures on free probability theory. *Lectures on probability theory and statistics (Saint-Flour, 1998)*, 1738:279–349, 2000.

[243] D. Voiculescu. Free entropy. *Bulletin of the London Mathematical Society*, 34(3):257–278, 2002.

[244] Z. Wen, C. Yang, X. Liu, and S. Marchesini. Alternating direction methods for classical and ptychographic phase retrieval. *Inverse Problems*, 28(11):115010, 2012.

[245] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. of Math.*, 67:325–328, 1958.

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

[247] R. Young. On the best possible constants in the khintchine inequality. *Journal of the London Mathematical Society*, 2(3):496–504, 1976.

[248] Z. Zhang. Some operator convex functions of several variables. *Linear Algebra and its Applications*, 463:1–9, 2014.

[249] G. Zheng, R. Horstmeyer, and C. Yang. Wide-field, high-resolution fourier ptychographic microscopy. *Nature photonics*, 7(9):739–745, 2013.