Springer Series in Statistics

Zhidong Bai
Jack W. Silverstein

Spectral Analysis of Large Dimensional Random Matrices

Second Edition

Springer

Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

For other titles published in this series, go to
http://www.springer.com/series/692

Zhidong Bai
Jack W. Silverstein

Spectral Analysis of Large Dimensional Random Matrices

Second Edition

Springer

Zhidong Bai
School of Mathematics and Statistics
KLAS MOE
Northeast Normal University
5268 Renmin Street
Changchun, Jilin 130024
China
baizd@nenu.edu.cn
&
Department of Statistics and Applied Probability
National University of Singapore
6 Science Drive 2
Singapore 117546
Singapore
stabaizd@nus.edu.sg

Jack W. Silverstein
Department of Mathematics
Box 8205
North Carolina State University
Raleigh, NC 27695-8205
jack@unity.ncsu.edu

ISSN 0172-7397
ISBN 978-1-4419-0660-1 e-ISBN 978-1-4419-0661-8
DOI 10.1007/978-1-4419-0661-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009942423

This book is dedicated to:

Professor Calyampudi Radhakrishna Rao's 90th Birthday
Professor Ulf Grenander's 87th Birthday
Professor Yongquan Yin's 80th Birthday

and to

My wife, Xicun Dan, my sons
Li and Steve Gang, and grandsons
Yongji, and Yonglin

— Zhidong Bai

My children, Hila and Idan

— Jack W. Silverstein

Preface to the Second Edition

The ongoing developments being made in large dimensional data analysis continue to generate great interest in random matrix theory in both theoretical investigations and applications in many disciplines. This has doubtlessly contributed to the significant demand for this monograph, resulting in its first printing being sold out. The authors have received many requests to publish a second edition of the book.

Since the publication of the first edition in 2006, many new results have been reported in the literature. However, due to limitations in space, we cannot include all new achievements in the second edition. In accordance with the needs of statistics and signal processing, we have added a new chapter on the limiting behavior of eigenvectors of large dimensional sample covariance matrices. To illustrate the application of RMT to wireless communications and statistical finance, we have added a chapter on these areas. Certain new developments are commented on throughout the book. Some typos and errors found in the first edition have been corrected.

The authors would like to express their appreciation to Ms. Lü Hong for her help in the preparation of the second edition. They would also like to thank Professors Ying-Chang Liang, Zhaoben Fang, Baoxue Zhang, and Shurong Zheng, and Mr. Jiang Hu, for their valuable comments and suggestions. They also thank the copy editor, Mr. Hal Heinglein, for his careful reading, corrections, and helpful suggestions. The first author would like to acknowledge the support from grants NSFC 10871036, NUS R-155-000-079-112, and R-155-000-096-720.

Changchun, China, and Singapore — Zhidong Bai
Cary, North Carolina, USA — Jack W. Silverstein
March 2009

Preface to the First Edition

This monograph is an introductory book on the theory of random matrices (RMT). The theory dates back to the early development of quantum mechanics in the 1940s and 1950s. In an attempt to explain the complex organizational structure of heavy nuclei, E. Wigner, Professor of Mathematical Physics at Princeton University, argued that one should not compute energy levels from Schrödinger's equation. Instead, one should imagine the complex nuclei system as a black box described by $n \times n$ Hamiltonian matrices with elements drawn from a probability distribution with only mild constraints dictated by symmetry considerations. Under these assumptions and a mild condition imposed on the probability measure in the space of matrices, one finds the joint probability density of the n eigenvalues. Based on this consideration, Wigner established the well-known semicircular law. Since then, RMT has been developed into a big research area in mathematical physics and probability. Its rapid development can be seen from the following statistics from the Mathscinet database under keyword Random Matrix on 10 June 2005 (Table 0.1).

Table 0.1 Publication numbers on RMT in 10 year periods since 1955

1955–1964	1965–1974	1975–1984	1985–1994	1995–2004
23	138	249	635	1205

Modern developments in computer science and computing facilities motivate ever widening applications of RMT to many areas.

In statistics, classical limit theorems have been found to be seriously inadequate in aiding in the analysis of very high dimensional data.

In the biological sciences, a DNA sequence can be as long as several billion strands. In financial research, the number of different stocks can be as large as tens of thousands.

In wireless communications, the number of users can be several million.

All of these areas are challenging classical statistics. Based on these needs, the number of researchers on RMT is gradually increasing. The purpose of this monograph is to introduce the basic results and methodologies developed in RMT. We assume readers of this book are graduate students and beginning researchers who are interested in RMT. Thus, we are trying to provide the most advanced results with proofs using standard methods as detailed as we can.

After more than a half century, many different methodologies of RMT have been developed in the literature. Due to the limitation of our knowledge and length of the book, it is impossible to introduce all the procedures and results. What we shall introduce in this book are those results obtained either under moment restrictions using the moment convergence theorem or the Stieltjes transform.

In an attempt at complementing the material presented in this book, we have listed some recent publications on RMT that we have not introduced.

The authors would like to express their appreciation to Professors Chen Mufa, Lin Qun, and Shi Ningzhong, and Ms. Lü Hong for their encouragement and help in the preparation of the manuscript. They would also like to thank Professors Zhang Baoxue, Lee Sungchul, Zheng Shurong, Zhou Wang, and Hu Guorong for their valuable comments and suggestions.

Changchun, China
Cary, North Carolina, USA

Zhidong Bai
Jack W. Silverstein
June 2005

Contents

Preface to the Second Edition vii
Preface to the First Edition ix

1 Introduction 1
1.1 Large Dimensional Data Analysis 1
1.2 Random Matrix Theory 4
1.2.1 Spectral Analysis of Large Dimensional Random Matrices 4
1.2.2 Limits of Extreme Eigenvalues 6
1.2.3 Convergence Rate of the ESD 6
1.2.4 Circular Law 7
1.2.5 CLT of Linear Spectral Statistics 8
1.2.6 Limiting Distributions of Extreme Eigenvalues and Spacings 9
1.3 Methodologies 9
1.3.1 Moment Method 9
1.3.2 Stieltjes Transform 10
1.3.3 Orthogonal Polynomial Decomposition 11
1.3.4 Free Probability 13

2 Wigner Matrices and Semicircular Law 15
2.1 Semicircular Law by the Moment Method 16
2.1.1 Moments of the Semicircular Law 16
2.1.2 Some Lemmas in Combinatorics 16
2.1.3 Semicircular Law for the iid Case 20
2.2 Generalizations to the Non-iid Case 26
2.2.1 Proof of Theorem 2.9 26
2.3 Semicircular Law by the Stieltjes Transform 31
2.3.1 Stieltjes Transform of the Semicircular Law 31
2.3.2 Proof of Theorem 2.9 33

3 Sample Covariance Matrices and the Marčenko-Pastur Law 39
3.1 M-P Law for the iid Case . . . 40
3.1.1 Moments of the M-P Law . . . 40
3.1.2 Some Lemmas on Graph Theory and Combinatorics . 41
3.1.3 M-P Law for the iid Case . . . 47
3.2 Generalization to the Non-iid Case . . . 51
3.3 Proof of Theorem 3.10 by the Stieltjes Transform . . . 52
3.3.1 Stieltjes Transform of the M-P Law . . . 52
3.3.2 Proof of Theorem 3.10 . . . 53

4 Product of Two Random Matrices . . . 59
4.1 Main Results . . . 60
4.2 Some Graph Theory and Combinatorial Results . . . 61
4.3 Proof of Theorem 4.1 . . . 68
4.3.1 Truncation of the ESD of $\mathbf{T}_n$ . . . 68
4.3.2 Truncation, Centralization, and Rescaling of the X-variables . . . 70
4.3.3 Completing the Proof . . . 71
4.4 LSD of the F-Matrix . . . 75
4.4.1 Generating Function for the LSD of $\mathbf{S}_n\mathbf{T}_n$ . . . 75
4.4.2 Completing the Proof of Theorem 4.10 . . . 77
4.5 Proof of Theorem 4.3 . . . 80
4.5.1 Truncation and Centralization . . . 80
4.5.2 Proof by the Stieltjes Transform . . . 82

5 Limits of Extreme Eigenvalues . . . 91
5.1 Limit of Extreme Eigenvalues of the Wigner Matrix . . . 92
5.1.1 Sufficiency of Conditions of Theorem 5.1 . . . 93
5.1.2 Necessity of Conditions of Theorem 5.1 . . . 101
5.2 Limits of Extreme Eigenvalues of the Sample Covariance Matrix . . . 105
5.2.1 Proof of Theorem 5.10 . . . 106
5.2.2 Proof of Theorem 5.11 . . . 113
5.2.3 Necessity of the Conditions . . . 113
5.3 Miscellanies . . . 114
5.3.1 Spectral Radius of a Nonsymmetric Matrix . . . 114
5.3.2 TW Law for the Wigner Matrix . . . 115
5.3.3 TW Law for a Sample Covariance Matrix . . . 117

6 Spectrum Separation . . . 119
6.1 What Is Spectrum Separation? . . . 119
6.1.1 Mathematical Tools . . . 126
6.2 Proof of (1) . . . 128
6.2.1 Truncation and Some Simple Facts . . . 128
6.2.2 A Preliminary Convergence Rate . . . 129

6.2.3 Convergence of $s_n - \mathrm{E}s_n$ 139
6.2.4 Convergence of the Expected Value 144
6.2.5 Completing the Proof 148
6.3 Proof of (2) 149
6.4 Proof of (3) 151
6.4.1 Convergence of a Random Quadratic Form 151
6.4.2 spread of eigenvaluesSpread of Eigenvalues 154
6.4.3 Dependence on y 157
6.4.4 Completing the Proof of (3) 160

7 Semicircular Law for Hadamard Products 165
7.1 Sparse Matrix and Hadamard Product 165
7.2 Truncation and Normalization 168
7.2.1 Truncation and Centralization 169
7.3 Proof of Theorem 7.1 by the Moment Approach 172

8 Convergence Rates of ESD 181
8.1 Convergence Rates of the Expected ESD of Wigner Matrices . 181
8.1.1 Lemmas on Truncation, Centralization, and Rescaling . 182
8.1.2 Proof of Theorem 8.2 185
8.1.3 Some Lemmas on Preliminary Calculation 189
8.2 Further Extensions 194
8.3 Convergence Rates of the Expected ESD of Sample Covariance Matrices 195
8.3.1 Assumptions and Results 195
8.3.2 Truncation and Centralization 197
8.3.3 Proof of Theorem 8.10 198
8.4 Some Elementary Calculus 204
8.4.1 Increment of M-P Density 204
8.4.2 Integral of Tail Probability 206
8.4.3 Bounds of Stieltjes Transforms of the M-P Law 207
8.4.4 Bounds for $\tilde{b}_n$ 209
8.4.5 Integrals of Squared Absolute Values of Stieltjes Transforms 212
8.4.6 Higher Central Moments of Stieltjes Transforms 213
8.4.7 Integral of δ 217
8.5 Rates of Convergence in Probability and Almost Surely 219

9 CLT for Linear Spectral Statistics 223
9.1 Motivation and Strategy 223
9.2 CLT of LSS for the Wigner Matrix 227
9.2.1 Strategy of the Proof 229
9.2.2 Truncation and Renormalization 231
9.2.3 Mean Function of M_n 232
9.2.4 Proof of the Nonrandom Part of (9.2.13) for $j = l, r$ . . 238

9.3 Convergence of the Process $M_n - \mathrm{E}M_n$ 239
9.3.1 Finite-Dimensional Convergence of $M_n - \mathrm{E}M_n$ 239
9.3.2 Limit of S_1 242
9.3.3 Completion of the Proof of (9.2.13) for $j = l, r$ 250
9.3.4 Tightness of the Process $M_n(z) - \mathrm{E}M_n(z)$ 251
9.4 Computation of the Mean and Covariance Function of $G(f)$. 252
9.4.1 Mean Function 252
9.4.2 Covariance Function 254
9.5 Application to Linear Spectral Statistics and Related Results 256
9.5.1 Tchebychev Polynomials 256
9.6 Technical Lemmas 257
9.7 CLT of the LSS for Sample Covariance Matrices 259
9.7.1 Truncation 261
9.8 Convergence of Stieltjes Transforms 263
9.9 Convergence of Finite-Dimensional Distributions 269
9.10 Tightness of $M_n^1(z)$ 280
9.11 Convergence of $M_n^2(z)$ 286
9.12 Some Derivations and Calculations 292
9.12.1 Verification of (9.8.8) 292
9.12.2 Verification of (9.8.9) 295
9.12.3 Derivation of Quantities in Example (1.1) 296
9.12.4 Verification of Quantities in Jonsson's Results 298
9.12.5 Verification of (9.7.8) and (9.7.9) 300
9.13 CLT for the F-Matrix 304
9.13.1 CLT for LSS of the F-Matrix 306
9.14 Proof of Theorem 9.14 308
9.14.1 Lemmas 308
9.14.2 Proof of Theorem 9.14 318
9.15 CLT for the LSS of a Large Dimensional Beta-Matrix 325
9.16 Some Examples 326

10 Eigenvectors of Sample Covariance Matrices 331
10.1 Formulation and Conjectures 332
10.1.1 Haar Measure and Haar Matrices 332
10.1.2 Universality 335
10.2 A Necessary Condition for Property 5′ 336
10.3 Moments of $X_p(F^{\mathbf{S}_p})$ 339
10.3.1 Proof of (10.3.1) $\Rightarrow$ (10.3.2) 340
10.3.2 Proof of (b) 341
10.3.3 Proof of (10.3.2) $\Rightarrow$ (10.3.1) 341
10.3.4 Proof of (c) 349
10.4 An Example of Weak Convergence 349
10.4.1 Converting to $D[0, \infty)$ 350
10.4.2 A New Condition for Weak Convergence 357

10.4.3 Completing the Proof 362
10.5 Extension of (10.2.6) to $\mathbf{B}_n = \mathbf{T}^{1/2}\mathbf{S}_p\mathbf{T}^{1/2}$ 366
10.5.1 First-Order Limit 366
10.5.2 CLT of Linear Functionals of $\mathbf{B}_p$ 367
10.6 Proof of Theorem 10.16 368
10.7 Proof of Theorem 10.21 372
10.7.1 An Intermediate Lemma 372
10.7.2 Convergence of the Finite-Dimensional Distributions . . 373
10.7.3 Tightness of $M_n^1(z)$ and Convergence of $M_n^2(z)$ 385
10.8 Proof of Theorem 10.23 388

11 Circular Law 391
11.1 The Problem and Difficulty 391
11.1.1 Failure of Techniques Dealing with Hermitian Matrices 392
11.1.2 Revisiting Stieltjes Transformation 393
11.2 A Theorem Establishing a Partial Answer to the Circular Law 396
11.3 Lemmas on Integral Range Reduction 397
11.4 Characterization of the Circular Law 401
11.5 A Rough Rate on the Convergence of $\nu_n(x, z)$ 409
11.5.1 Truncation and Centralization 409
11.5.2 A Convergence Rate of the Stieltjes Transform of $\nu_n(\cdot, z)$ 411
11.6 Proofs of (11.2.3) and (11.2.4) 420
11.7 Proof of Theorem 11.4 424
11.8 Comments and Extensions 425
11.8.1 Relaxation of Conditions Assumed in Theorem 11.4 . . . 425
11.9 Some Elementary Mathematics 428
11.10New Developments 430

12 Some Applications of RMT 433
12.1 Wireless Communications 433
12.1.1 Channel Models 435
12.1.2 random matrix channelRandom Matrix Channels 436
12.1.3 Linearly Precoded Systems 438
12.1.4 Channel Capacity for MIMO Antenna Systems 442
12.1.5 Limiting Capacity of Random MIMO Channels 450
12.1.6 A General DS-CDMA Model 452
12.2 Application to Finance 454
12.2.1 A Review of Portfolio and Risk Management 455
12.2.2 Enhancement to a Plug-in Portfolio 460

A Some Results in Linear Algebra 469
A.1 Inverse Matrices and Resolvent 469
A.1.1 Inverse Matrix Formula 469
A.1.2 Holing a Matrix 470

A.1.3 Trace of an Inverse Matrix 470
A.1.4 Difference of Traces of a Matrix **A** and Its Major Submatrices 471
A.1.5 Inverse Matrix of Complex Matrices 472
A.2 Inequalities Involving Spectral Distributions 473
A.2.1 Singular-Value Inequalities 473
A.3 Hadamard Product and Odot Product 480
A.4 Extensions of Singular-Value Inequalities 483
A.4.1 Definitions and Properties 484
A.4.2 Graph-Associated Multiple Matrices 485
A.4.3 Fundamental Theorem on Graph-Associated MMs 488
A.5 Perturbation Inequalities 496
A.6 Rank Inequalities 503
A.7 A Norm Inequality 505

B Miscellanies 507
B.1 Moment Convergence Theorem 507
B.2 Stieltjes Transform 514
B.2.1 Preliminary Properties 514
B.2.2 Inequalities of Distance between Distributions in Terms of Their Stieltjes Transforms 517
B.2.3 Lemmas Concerning Levy Distance 521
B.3 Some Lemmas about Integrals of Stieltjes Transforms 523
B.4 A Lemma on the Strong Law of Large Numbers 526
B.5 A Lemma on Quadratic Forms 530

Relevant Literature 533

Index 547

Chapter 1
Introduction

1.1 Large Dimensional Data Analysis

The aim of this book is to investigate the spectral properties of random matrices (RM) when their dimensions tend to infinity. All classical limiting theorems in statistics are under the assumption that the dimension of data is fixed. Then, it is natural to ask why the dimension needs to be considered large and whether there are any differences between the results for a fixed dimension and those for a large dimension.

In the past three or four decades, a significant and constant advancement in the world has been in the rapid development and wide application of computer science. Computing speed and storage capability have increased a thousand folds. This has enabled one to collect, store, and analyze data sets of very high dimension. These computational developments have had a strong impact on every branch of science. For example, Fisher's resampling theory had been silent for more than three decades due to the lack of efficient random number generators until Efron proposed his renowned bootstrap in the late 1970s; the minimum L_1 norm estimation had been ignored for centuries since it was proposed by Laplace until Huber revived it and further extended it to robust estimation in the early 1970s. It is difficult to imagine that these advanced areas in statistics would have received such deep development if there had been no assistance from the present-day computer.

Although modern computer technology helps us in so many respects, it also brings a new and urgent task to the statistician; that is, whether the classical limit theorems (i.e., those assuming a fixed dimension) are still valid for analyzing high dimensional data and how to remedy them if they are not.

Basically, there are two kinds of limiting results in multivariate analysis: those for a fixed dimension (classical limit theorems) and those for a large dimension (large dimensional limit theorems). The problem turns out to be which kind of result is closer to reality. As argued by Huber in [157], some statisticians might say that five samples for each parameter on average are

Z. Bai and J.W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices, Second Edition*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0661-8_1,

enough to use asymptotic results. Now, suppose there are $p = 20$ parameters and we have a sample of size $n = 100$. We may consider the case as $p = 20$ being fixed and n tending to infinity, $p = 2\sqrt{n}$, or $p = 0.2n$. So, we have at least three different options from which to choose for an asymptotic setup. A natural question is then which setup is the best choice among the three. Huber strongly suggested studying the situation of an increasing dimension together with the sample size in linear regression analysis.

This situation occurs in many cases. In parameter estimation for a structured covariance matrix, simulation results show that parameter estimation becomes very poor when the number of parameters is more than four. Also, it is found in linear regression analysis that if the covariates are random (or have measurement errors) and the number of covariates is larger than six, the behavior of the estimates departs far away from the theoretic values unless the sample size is very large. In signal processing, when the number of signals is two or three and the number of sensors is more than 10, the traditional MUSIC (MUltiple SIgnal Classification) approach provides very poor estimation of the number of signals unless the sample size is larger than 1000. Paradoxically, if we use only half of the data set—namely, we use the data set collected by only five sensors—the signal number estimation is almost 100% correct if the sample size is larger than 200. Why would this paradox happen? Now, if the number of sensors (the dimension of data) is p, then one has to estimate p^2 parameters ($\frac{1}{2}p(p+1)$ real parts and $\frac{1}{2}p(p-1)$ imaginary parts of the covariance matrix). Therefore, when p increases, the number of parameters to be estimated increases proportional to p^2 while the number ($2np$) of observations increases proportional to p. This is the underlying reason for this paradox. This suggests that one has to revise the traditional MUSIC method if the sensor number is large.

An interesting problem was discussed by Bai and Saranadasa [27], who theoretically proved that when testing the difference of means of two high dimensional populations, Dempster's [91] nonexact test is more powerful than Hotelling's T^2 test even when the T^2 statistic is well defined.

It is well known that statistical efficiency will be significantly reduced when the dimension of data or number of parameters becomes large. Thus, several techniques for dimension reduction have been developed in multivariate statistical analysis. As an example, let us consider a problem in principal component analysis. If the data dimension is 10, one may select three principal components so that more than 80% of the information is reserved in the principal components. However, if the data dimension is 1000 and 300 principal components are selected, one would still have to face a high dimensional problem. If one only chooses three principal components, he would have lost 90% or even more of the information carried in the original data set. Now, let us consider another example.

Example 1.1. Let X_{ij} be iid standard normal variables. Write

$$S_n = \left(\frac{1}{n} \sum_{k=1}^{n} X_{ik} X_{jk} \right)_{i,j=1}^{p},$$

which can be considered as a sample covariance matrix with n samples of a p-dimensional mean-zero random vector with population matrix I. An important statistic in multivariate analysis is

$$T_n = \log(\det S_n) = \sum_{j=1}^{p} \log(\lambda_{n,j}),$$

where $\lambda_{n,j}$, $j = 1, \cdots, p$, are the eigenvalues of S_n. When p is fixed, $\lambda_{n,j} \to 1$ almost surely as $n \to \infty$ and thus $T_n \xrightarrow{\text{a.s.}} 0$.

Further, by taking a Taylor expansion on $\log(1 + x)$, one can show that

$$\sqrt{n/p}\, T_n \xrightarrow{\mathscr{D}} N(0, 2),$$

for any fixed p. This suggests the possibility that T_n is asymptotically normal, provided that $p = O(n)$. However, this is not the case. Let us see what happens when $p/n \to y \in (0, 1)$ as $n \to \infty$. Using results on the limiting spectral distribution of $\{S_n\}$ (see Chapter 3), we will show that with probability 1

$$\frac{1}{p} T_n \to \int_{a(y)}^{b(y)} \frac{\log x}{2\pi x y} \sqrt{(b(y) - x)(x - a(y))}\, dx = \frac{y - 1}{y} \log(1 - y) - 1 \equiv d(y) < 0 \tag{1.1.1}$$

where $a(y) = (1 - \sqrt{y})^2$, $b(y) = (1 + \sqrt{y})^2$. This shows that almost surely

$$\sqrt{n/p}\, T_n \sim d(y) \sqrt{np} \to -\infty.$$

Thus, any test that assumes asymptotic normality of T_n will result in a serious error.

These examples show that the classical limit theorems are no longer suitable for dealing with high dimensional data analysis. Statisticians must seek out special limiting theorems to deal with large dimensional statistical problems. Thus, the theory of random matrices (RMT) might be one possible method for dealing with large dimensional data analysis and hence has received more attention among statisticians in recent years. For the same reason, the importance of RMT has found applications in many research areas, such as signal processing, network security, image processing, genetic statistics, stock market analysis, and other finance or economic problems.

1.2 Random Matrix Theory

RMT traces back to the development of quantum mechanics (QM) in the 1940s and early 1950s. In QM, the energy levels of a system are described by eigenvalues of a Hermitian operator $\mathbf{A}$ on a Hilbert space, called the Hamiltonian. To avoid working with an infinite dimensional operator, it is common to approximate the system by discretization, amounting to a truncation, keeping only the part of the Hilbert space that is important to the problem under consideration. Hence, the limiting behavior of large dimensional random matrices has attracted special interest among those working in QM, and many laws were discovered during that time. For a more detailed review on applications of RMT in QM and other related areas, the reader is referred to the book *Random Matrices* by Mehta [212].

Since the late 1950s, research on the limiting spectral analysis of large dimensional random matrices has attracted considerable interest among mathematicians, probabilists, and statisticians. One pioneering work is the semicircular law for a Gaussian (or Wigner) matrix (see Chapter 2 for the definition), due to Wigner [296, 295]. He proved that the expected spectral distribution of a large dimensional Wigner matrix tends to the so-called semicircular law. This work was generalized by Arnold [8, 7] and Grenander [136] in various aspects. Bai and Yin [37] proved that the spectral distribution of a sample covariance matrix (suitably normalized) tends to the semicircular law when the dimension is relatively smaller than the sample size. Following the work of Marčenko and Pastur [201] and Pastur [230, 229], the asymptotic theory of spectral analysis of large dimensional sample covariance matrices was developed by many researchers, including Bai, Yin, and Krishnaiah [41], Grenander and Silverstein [137], Jonsson [169], Wachter [291, 290], Yin [300], and Yin and Krishnaiah [304]. Also, Yin, Bai, and Krishnaiah [301, 302], Silverstein [260], Wachter [290], Yin [300], and Yin and Krishnaiah [304] investigated the limiting spectral distribution of the multivariate F-matrix, or more generally of products of random matrices. In the early 1980s, major contributions on the existence of the limiting spectral distribution (LSD) and their explicit forms for certain classes of random matrices were made. In recent years, research on RMT has turned toward second-order limiting theorems, such as the central limit theorem for linear spectral statistics, the limiting distributions of spectral spacings, and extreme eigenvalues.

1.2.1 Spectral Analysis of Large Dimensional Random Matrices

Suppose $\mathbf{A}$ is an $m \times m$ matrix with eigenvalues λ_j, $j = 1, 2, \cdots, m$. If all these eigenvalues are real (e.g., if $\mathbf{A}$ is Hermitian), we can define a one-dimensional

distribution function

$$F^{\mathbf{A}}(x) = \frac{1}{m}\#\{j \le m : \lambda_j \le x\} \tag{1.2.1}$$

called the empirical spectral distribution (ESD) of the matrix $\mathbf{A}$. Here $\#E$ denotes the cardinality of the set E. If the eigenvalues λ_j's are not all real, we can define a two-dimensional empirical spectral distribution of the matrix $\mathbf{A}$:

$$F^{\mathbf{A}}(x,y) = \frac{1}{m}\#\{j \le m : \Re(\lambda_j) \le x,\, \Im(\lambda_j) \le y\}. \tag{1.2.2}$$

One of the main problems in RMT is to investigate the convergence of the sequence of empirical spectral distributions $\{F^{\mathbf{A}_n}\}$ for a given sequence of random matrices $\{\mathbf{A}_n\}$. The limit distribution F (possibly defective; that is, total mass is less than 1 when some eigenvalues tend to $\pm\infty$), which is usually nonrandom, is called the *limiting spectral distribution* (LSD) of the sequence $\{\mathbf{A}_n\}$.

We are especially interested in sequences of random matrices with dimension (number of columns) tending to infinity, which refers to *the theory of large dimensional random matrices*.

The importance of ESD is due to the fact that many important statistics in multivariate analysis can be expressed as functionals of the ESD of some RM. We now give a few examples.

Example 1.2. Let $\mathbf{A}$ be an $n \times n$ positive definite matrix. Then

$$\det(\mathbf{A}) = \prod_{j=1}^{n} \lambda_j = \exp\left(n \int_0^\infty \log x F^{\mathbf{A}}(dx)\right).$$

Example 1.3. Let the covariance matrix of a population have the form $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_q + \sigma^2\mathbf{I}$, where the dimension of $\boldsymbol{\Sigma}$ is p and the rank of $\boldsymbol{\Sigma}_q$ is $q(< p)$. Suppose $\mathbf{S}$ is the sample covariance matrix based on n iid samples drawn from the population. Denote the eigenvalues of $\mathbf{S}$ by $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p$. Then the test statistic for the hypothesis $H_0 : \operatorname{rank}(\boldsymbol{\Sigma}_q) = q$ against $H_1 : \operatorname{rank}(\boldsymbol{\Sigma}_q) > q$ is given by

$$T = \frac{1}{p-q}\sum_{j=q+1}^{p} \sigma_j^2 - \left(\frac{1}{p-q}\sum_{j=q+1}^{p} \sigma_j\right)^2$$

$$= \frac{p}{p-q}\int_0^{\sigma_q} x^2 F^{\mathbf{S}}(dx) - \left(\frac{p}{p-q}\int_0^{\sigma_q} x F^{\mathbf{S}}(dx)\right)^2.$$

1.2.2 Limits of Extreme Eigenvalues

In applications of the asymptotic theorems of spectral analysis of large dimensional random matrices, two important problems arise after the LSD is found. The first is the bound on extreme eigenvalues; the second is the convergence rate of the ESD with respect to sample size. For the first problem, the literature is extensive. The first success was due to Geman [118], who proved that the largest eigenvalue of a sample covariance matrix converges almost surely to a limit under a growth condition on all the moments of the underlying distribution. Yin, Bai, and Krishnaiah [301] proved the same result under the existence of the fourth moment, and Bai, Silverstein, and Yin [33] proved that the existence of the fourth moment is also necessary for the existence of the limit. Bai and Yin [38] found the necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. By the symmetry between the largest and smallest eigenvalues of a Wigner matrix, the necessary and sufficient conditions for almost sure convergence of the smallest eigenvalue of a Wigner matrix were also found.

Compared to almost sure convergence of the largest eigenvalue of a sample covariance matrix, a relatively harder problem is to find the limit of the smallest eigenvalue of a large dimensional sample covariance matrix. The first attempt was made in Yin, Bai, and Krishnaiah [302], in which it was proved that the almost sure limit of the smallest eigenvalue of a Wishart matrix has a positive lower bound when the ratio of the dimension to the degrees of freedom is less than 1/2. Silverstein [262] modified the work to allow a ratio less than 1. Silverstein [263] further proved that, with probability 1, the smallest eigenvalue of a Wishart matrix tends to the lower bound of the LSD when the ratio of the dimension to the degrees of freedom is less than 1. However, Silverstein's approach strongly relies on the normality assumption on the underlying distribution and thus cannot be extended to the general case. The most current contribution was made in Bai and Yin [36], in which it is proved that, under the existence of the fourth moment of the underlying distribution, the smallest eigenvalue (when $p \leq n$) or the $p-n+1$st smallest eigenvalue (when $p > n$) tends to $a(y) = \sigma^2(1-\sqrt{y})^2$, where $y = \lim(p/n) \in (0, \infty)$. Compared to the case of the largest eigenvalues of a sample covariance matrix, the existence of the fourth moment seems to be necessary also for the problem of the smallest eigenvalue. However, this problem has not yet been solved.

1.2.3 Convergence Rate of the ESD

The second problem, the convergence rate of the spectral distributions of large dimensional random matrices, is of practical interest. Indeed, when the LSD is used in estimating functionals of eigenvalues of a random matrix, it is

important to understand the reliability of performing the substitution. This problem had been open for decades. In finding the limits of both the LSD and the extreme eigenvalues of symmetric random matrices, a very useful and powerful method is the moment method, which does not give any information about the rate of the convergence of the ESD to the LSD. The first success was made in Bai [16, 17], in which a Berry-Esseen type inequality of the difference of two distributions was established in terms of their Stieltjes transforms. Applying this inequality, a convergence rate for the expected ESD of a large Wigner matrix was proved to be $O(n^{-1/4})$ and that for the sample covariance matrix was shown to be $O(n^{-1/4})$ if the ratio of the dimension to the degrees of freedom is far from 1 and $O(n^{-5/48})$ if the ratio is close to 1. Some further developments can be found in Bai et al. [23, 24, 25], Bai et al. [26], Götze et al. [132], and Götze and Tikhomirov [133, 134].

1.2.4 Circular Law

The most perplexing problem is the so-called circular law, which conjectures that the spectral distribution of a nonsymmetric random matrix, after suitable normalization, tends to the uniform distribution over the unit disk in the complex plane. The difficulty exists in that two of the most important tools used for symmetric matrices do not apply for nonsymmetric matrices. Furthermore, certain truncation and centralization techniques cannot be used. The first known result was given in Mehta [212] (1967 edition) and in an unpublished paper of Silverstein (1984) that was reported in Hwang [159]. They considered the case where the entries of the matrix are iid standard complex normal. Their method uses the explicit expression of the joint density of the complex eigenvalues of the random matrix that was found by Ginibre [120]. The first attempt to prove this conjecture under some general conditions was made in Girko [123, 124]. However, his proofs contain serious mathematical gaps and have been considered questionable in the literature. Recently, Edelman [98] found the conditional joint distribution of complex eigenvalues of a random matrix whose entries are real normal $N(0, 1)$ when the number of its real eigenvalues is given and proved that the expected spectral distribution of the real Gaussian matrix tends to the circular law. Under the existence of the $4 + \varepsilon$ moment and the existence of a density, Bai [14] proved the strong version of the circular law. Recent work has eliminated the density requirement and weakened the moment condition. Further details are given in Chapter 11. Some consequent achievements can be found in Pan and Zhou [227] and Tao and Vu [273].

1.2.5 CLT of Linear Spectral Statistics

As mentioned above, functionals of the ESD of RMs are important in multivariate inference. Indeed, a parameter θ of the population can sometimes be expressed as

$$\theta = \int f(x)dF(x).$$

To make statistical inference on θ, one may use the integral

$$\hat{\theta} = \int f(x)dF_n(x),$$

which we call *linear spectral statistics* (LSS), as an estimator of θ, where $F_n(x)$ is the ESD of the RM computed from the data set. Further, one may want to know the limiting distribution of $\hat{\theta}$ through suitable normalization. In Bai and Silverstein [30], the normalization was found to be n by showing the limiting distribution of the linear functional

$$X_n(f) = n \int f(t)d(F_n(t) - F(t))$$

to be Gaussian under certain assumptions.

The first work in this direction was done by Jonsson [169], in which $f(t) = t^r$ and F_n is the ESD of a normalized standard Wishart matrix. Further work was done by Johansson [165], Bai and Silverstein [30], Bai and Yao [35], Sinai and Soshnikov [269], Anderson and Zeitouni [2], and Chatterjee [77], among others.

It would seem natural to pursue the properties of linear functionals by way of proving results on the process $G_n(t) = \alpha_n(F_n(t) - F(t))$ when viewed as a random element in $D[0, \infty)$, the metric space of functions with discontinuities of the first kind, along with the Skorohod metric. Unfortunately, this is impossible. The work done in Bai and Silverstein [30] shows that $G_n(t)$ cannot converge weakly to any nontrivial process for any choice of α_n. This fact appears to occur in other random matrix ensembles. When F_n is the empirical distribution of the angles of eigenvalues of an $n \times n$ Haar matrix, Diaconis and Evans [94] proved that all finite dimensional distributions of $G_n(t)$ converge in distribution to independent Gaussian variables when $\alpha_n = n/\sqrt{\log n}$. This shows that with $\alpha_n = n/\sqrt{\log n}$, the process G_n cannot be tight in $D[0, \infty)$.

The result of Bai and Silverstein [30] has been applied in several areas, especially in wireless communications, where sample covariance matrices are used to model transmission between groups of antennas. See, for example, Tulino and Verdu [283] and Kamath and Hughes [170].

1.2.6 Limiting Distributions of Extreme Eigenvalues and Spacings

The first work on the limiting distributions of extreme eigenvalues was done by Tracy and Widom [278], who found the expression for the largest eigenvalue of a Gaussian matrix when suitably normalized. Further, Johnstone [168] found the limiting distribution of the largest eigenvalue of the large Wishart matrix. In El Karoui [101], the Tracy-Widom law of the largest eigenvalue is established for the complex Wishart matrix when the population covariance matrix differs from the identity.

When the majority of the population eigenvalues are 1 and some are larger than 1, Johnstone proposed the *spiked eigenvalues model* in [168]. Then, Baik et al. [43] and Baik and Silverstein [44] investigated the strong limit of spiked eigenvalues. Bai and Yao [34] investigated the CLT of spiked eigenvalues. A special case of the CLT when the underlying distribution is complex Gaussian was considered in Baik et al. [43], and the real Gaussian case was considered in Paul [231].

The work on spectrum spacing has a long history that dates back to Mehta [213]. Most of the work in these two directions assumes the Gaussian (or generalized) distributions.

1.3 Methodologies

The eigenvalues of a matrix can be regarded as continuous functions of entries of the matrix. But these functions have no closed form when the dimension of the matrix is larger than 4. So special methods are needed to understand them. There are three important methods employed in this area: the moment method, Stieltjes transform, and orthogonal polynomial decomposition of the exact density of eigenvalues. Of course, the third method needs the assumption of the existence and special forms of the densities of the underlying distributions in the RM.

1.3.1 Moment Method

In the following, $\{F_n\}$ will denote a sequence of distribution functions, and the k-th moment of the distribution F_n is denoted by

$$\beta_{n,k} = \beta_k(F_n) := \int x^k dF_n(x). \tag{1.3.1}$$

The moment method is based on the moment convergence theorem (MCT); see Lemmas B.1, B.2, and B.3.

Let $\mathbf{A}$ be an $n \times n$ Hermitian matrix, and denote its eigenvalues by $\lambda_1 \leq \cdots \leq \lambda_n$. The ESD, $F^{\mathbf{A}}$, of $\mathbf{A}$ is defined as in (1.2.1) with m replaced by n. Then, the k-th moment of $F^{\mathbf{A}}$ can be written as

$$\beta_{n,k}(\mathbf{A}) = \int_{-\infty}^{\infty} x^k F^{\mathbf{A}}(dx) = \frac{1}{n}\mathrm{tr}(\mathbf{A}^k). \tag{1.3.2}$$

This expression plays a fundamental role in RMT. By MCT, the problem of showing that the ESD of a sequence of random matrices $\{\mathbf{A}_n\}$ (strongly or weakly or in another sense) tends to a limit reduces to showing that, for each fixed k, the sequence $\{\frac{1}{n}\mathrm{tr}(\mathbf{A}^k)\}$ tends to a limit β_k in the corresponding sense and then verifying the Carleman condition (B.1.4),

$$\sum_{k=1}^{\infty} \beta_{2k}^{-1/2k} = \infty.$$

Note that in most cases the LSD has finite support, and hence the characteristic function of the LSD is analytic and the necessary condition for the MCT holds automatically. Most results in finding the LSD or proving the existence of the LSD were obtained by estimating the mean, variance, or higher moments of $\frac{1}{n}\mathrm{tr}(\mathbf{A}^k)$.

1.3.2 Stieltjes Transform

The definition and simple properties of the Stieltjes transform can be found in Appendix B, Section B.2. Here, we just illustrate how it can be used in RMT. Let $\mathbf{A}$ be an $n \times n$ Hermitian matrix and F_n be its ESD. Then, the Stieltjes transform of F_n is given by

$$s_n(z) = \int \frac{1}{x - z} dF_n(x) = \frac{1}{n}\mathrm{tr}(\mathbf{A} - z\mathbf{I})^{-1}.$$

Using the inverse matrix formula (see Theorem A.4), we get

$$s_n(z) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{a_{kk} - z - \boldsymbol{\alpha}_k^*(\mathbf{A}_k - z\mathbf{I})^{-1}\boldsymbol{\alpha}_k}$$

where $\mathbf{A}_k$ is the $(n-1) \times (n-1)$ matrix obtained from $\mathbf{A}$ with the k-th row and column removed and $\boldsymbol{\alpha}_k$ is the k-th column vector of $\mathbf{A}$ with the k-th element removed.

If the denominator $a_{kk} - z - \boldsymbol{\alpha}_k^*(\mathbf{A}_k - z\mathbf{I})^{-1}\boldsymbol{\alpha}_k$ can be proven to be equal to $g(z, s_n(z)) + o(1)$ for some function g, then the LSD F exists and its Stieltjes

transform of F is the solution to the equation

$$s = 1/g(z, s).$$

Its applications will be discussed in more detail later.

1.3.3 Orthogonal Polynomial Decomposition

Assume that the matrix $\mathbf{A}$ has a density $p_n(\mathbf{A}) = H(\lambda_1, \cdots, \lambda_n)$. It is known that the joint density function of the eigenvalues will be of the form

$$p_n(\lambda_1, \cdots, \lambda_n) = cJ(\lambda_1, \cdots, \lambda_n)H(\lambda_1, \cdots, \lambda_n),$$

where J comes from the integral of the Jacobian of the transform from the matrix space to its eigenvalue-eigenvector space. Generally, it is assumed that H has the form $H(\lambda_1, \cdots, \lambda_n) = \prod_{k=1}^{n} g(\lambda_k)$ and J has the form $\prod_{i<j}(\lambda_i - \lambda_j)^{\beta} \prod_{k=1}^{n} h_n(\lambda_k)$. For example, $\beta = 1$ and $h_n = 1$ for a real Gaussian matrix, $\beta = 2$, $h_n = 1$ for a complex Gaussian matrix, $\beta = 4$, $h_n = 1$ for a quaternion Gaussian matrix, and $\beta = 1$ and $h_n(x) = x^{n-p}$ for a real Wishart matrix with $n \geq p$.

Examples considered in the literature are the following

(1) Real Gaussian matrix (symmetric; i.e., $\mathbf{A}' = \mathbf{A}$):

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{4\sigma^2}\mathrm{tr}(\mathbf{A}^2)\right).$$

In this case, the diagonal entries of $\mathbf{A}$ are iid real $N(0, 2\sigma^2)$ and entries above diagonal are iid real $N(0, \sigma^2)$.

(2) Complex Gaussian matrix (Hermitian; i.e., $\mathbf{A}^* = \mathbf{A}$):

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{A}^2)\right).$$

In this case, the diagonal entries of $\mathbf{A}$ are iid real $N(0, \sigma^2)$ and entries above diagonal are iid complex $N(0, \sigma^2)$ (whose real and imaginary parts are iid $N(0, \sigma^2/2)$).

(3) Real Wishart matrix of order $p \times n$:

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{A}'\mathbf{A})\right).$$

In this case, the entries of $\mathbf{A}$ are iid real $N(0, \sigma^2)$.

(4) Complex Wishart matrix of order $p \times n$:

$$p_n(\mathbf{A}) = c\exp\left(-\frac{1}{\sigma^2}\mathrm{tr}(\mathbf{A}^*\mathbf{A})\right).$$

In this case, the entries of $\mathbf{A}$ are iid complex $N(0,\sigma^2)$.

For generalized densities, there are the following.

(1) Symmetric matrix:
$$p_n(\mathbf{A}) = c\exp(-\mathrm{tr}V(\mathbf{A})).$$

(2) Hermitian matrix:
$$p_n(\mathbf{A}) = c\exp(-\mathrm{tr}V(\mathbf{A})).$$
In the two cases above, V is assumed to be a polynomial of even degree with a positive leading coefficient.

(3) Real covariance matrix of dimension p and degrees of freedom n:
$$p_n(\mathbf{A}) = c\exp(-\mathrm{tr}V(\mathbf{A}'\mathbf{A})).$$

(4) Complex covariance matrix of dimension p and degrees of freedom n:
$$p_n(\mathbf{A}) = c\exp(-\mathrm{tr}V(\mathbf{A}^*\mathbf{A})).$$
In the two cases above, V is assumed to be a polynomial with a positive leading coefficient.

Note that the factor $\prod_{i<j}(\lambda_i-\lambda_j)$ is the determinant of the Vandermonde matrix generated by $\lambda_1,\cdots,\lambda_n$. Therefore, we may rewrite the density of the eigenvalues of the matrices as

$$\begin{aligned}
&p_n(\lambda_1,\cdots,\lambda_n)\\
&= c\prod_{k=1}^{n} h_n(\lambda_k)g(\lambda_k)\det\begin{pmatrix} 1 & 1 & \cdots & 1\\ \lambda_1 & \lambda_2 & \cdots & \lambda_n\\ \vdots & \vdots & \cdots & \vdots\\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1}\end{pmatrix}^{\beta}\\
&= c\prod_{k=1}^{n} h_n(\lambda_k)g(\lambda_k)\det\begin{pmatrix} 1 & 1 & \cdots & 1\\ m_1(\lambda_1) & m_1(\lambda_2) & \cdots & m_1(\lambda_n)\\ \vdots & \vdots & \cdots & \vdots\\ m_{n-1}(\lambda_1) & m_{n-1}(\lambda_2) & \cdots & m_{n-1}(\lambda_n)\end{pmatrix}^{\beta},
\end{aligned}$$

where m_k is any polynomial of degree k and having leading coefficient 1. For ease of finding the marginal densities of several eigenvalues, one may choose the m functions as orthogonal polynomials with respective $[g(x)h_n(x)]^{2/\beta}$. Then, through mathematical analysis, one can draw various conclusions from the expression above.

Note that the moment method and Stieltjes transform method can be done under moment assumptions. This book will primarily concentrate on

results without assuming density conditions. Readers who are interested in the method of orthogonal polynomials are referred to Deift [88].

1.3.4 Free Probability

Free probability is a mathematical theory that studies noncommutative random variables. The "freeness" property is the analogue of the classical notion of independence, and it is connected with free products. This theory was initiated by Dan Voiculescu around 1986 in order to attack the free group factors isomorphism problem, an important unsolved problem in the theory of operator algebras. Typically the random variables lies in a unital algebra A such as a C^* algebra or a von Neumann algebra. The algebra comes equipped with a noncommutative expectation, a linear functional $\varphi : A \to C$ such that $\varphi(1) = 1$. Unital subalgebras $A_1, \cdots, A_n$ are then said to be free if the expectation of the product $a_1 \cdots a_n$ is zero whenever each a_j has zero expectation, lies in an A_k, and no adjacent a_j's come from the same subalgebra A_k. Random variables are free if they generate free unital subalgebras.

An interesting aspect and active research direction of free probability lies in its applications to RMT. The functional φ stands for the normalized expected trace of a random matrix. For any $n \times n$ Hermitian random matrix $\mathbf{A}_n$ and a given integer k, $\varphi(\mathbf{A}_n^k) = \frac{1}{n}\mathrm{tr}(\mathrm{E}\mathbf{A}_n^k)$. If $\lim_n \varphi(\mathbf{A}_n^k) = \alpha_k$, for all k, then instead of referring to the collection of numbers α_k, it is better to use some random variable A (if it exists) to characterize the α_k's as moments of A. By setting $\varphi(A^k) = \alpha_k$, one may say that $\mathbf{A}_n \to A$ in distribution. A general definition is given as follows.

Definition 1.4. Consider $n \times n$ random matrices $A_n^{(1)}, \cdots, A_n^{(m)}$ and variables $A_1, \cdots, A_m$. We say that

$$(A_n^{(1)}, \cdots, A_n^{(m)}) \to (A_1, \cdots, A_m) \text{ in distribution}$$

if

$$\lim_{n\to\infty} \varphi(A_n^{(i_1)} \cdots A_n^{(i_k)}) = \varphi(A_{i_1} \cdots A_{i_k})$$

for all choices of k, $1 \le i_1, \cdots, i_k \le m$.

When $m = 1$, the definition of convergence in distribution is to say that if the normalized expected trace of $\mathbf{A}_n^k$ tends to the k-th moment of A, then we define $\mathbf{A}_n$ tending to A. For example, let $\mathbf{A}_n$ be the normalized Wigner matrix (see Chapter 2). Then A is the semicircular law. Now, suppose we have two independent sequences of normalized Wigner matrices, $\{\mathbf{A}_n\}$ and $\{\mathbf{B}_n\}$. How do we characterize their limits? If individually, then $\mathbf{A}_n \to s_a$ and $\mathbf{B}_n \to s_b$, and both s_a and s_b are semicircular laws. The problem is how to consider the joint limit of the sequences of pairs $(\mathbf{A}_n, \mathbf{B}_n)$. Or equivalently,

what is the relationship of s_a and s_b? According to free probability, we have the following definition.

Definition 1.5. The matrices $\mathbf{A}_1, \cdots, \mathbf{A}_m$ are called free if

$$\varphi([p_1(\mathbf{A}_{i_1}) \cdots p_k(\mathbf{A}_{i_k})]) = 0$$

whenever

- $p_1, \cdots, p_k$ are polynomials in one variable,
- $i_1 \neq i_2 \neq i_3 \neq \cdots \neq i_k$ (only neighboring elements are required to be distinct),
- $\varphi(p_j(\mathbf{A}_{i_j})) = 0$ for all $j = 1, \cdots, k$.

Note that the definition of freeness can be considered as a way of organizing the information about all joint moments of free variables in a systematic and conceptual way. Indeed, the definition above allows one to calculate mixed moments of free variables in terms of moments of the single variables. For example, if a, b are free, then the definition of freeness requires that $\varphi[(a - \varphi(a)1)(b - \varphi(b)1)] = 0$, which implies that $\varphi(ab) = \varphi(a)\varphi(b)$. In the same way, $\varphi[(a - \varphi(a)1)(b - \varphi(b)1)(a - \varphi(a)1)(b - \varphi(b)1)] = 0$ leads finally to $\varphi(abab) = \varphi(aa)\varphi(b)\varphi(b) + \varphi(a)\varphi(a)\varphi(bb) - \varphi(a)\varphi(b)\varphi(a)\varphi(b)$. Analogously, all mixed moments can (at least in principle) be calculated by reducing them to alternating products of centered variables as in the definition of freeness. Thus the statements s_a, s_b are free, and each of them being semicircular determines all joint moments in s_a and s_b. This shows that s_a and s_b are not ordinary random variables but take values on some noncommutative algebra.

To apply the theory of free probability to RMT, we need to extend the definition of free to asymptotic freeness; that is, replacing the state functional φ by ϕ, where

$$\phi(\mathbf{A}) = \lim_{n \to \infty} \frac{1}{n} \mathrm{trE}(\mathbf{A}_n).$$

Since normalized traces of powers of a Hermitian matrix are the moments of the ESD of the matrix, free probability reveals important information on their LSD. It is shown that freeness of random matrices corresponds to independence and to distributions being invariant under orthogonal transformations. Formulas have been derived that express the LSD of sums and products of free random matrices in terms of their individual LSDs.

For an excellent introduction to free probability, see Biane [52] and Nica and Speicher [221].

Chapter 2
Wigner Matrices and Semicircular Law

A Wigner matrix is a symmetric (or Hermitian in the complex case) random matrix. Wigner matrices play an important role in nuclear physics and mathematical physics. The reader is referred to Mehta [212] for applications of Wigner matrices to these areas. Here we mention that they also have a strong statistical meaning. Consider the limit of a normalized Wishart matrix. Suppose that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are iid samples drawn from a p-dimensional multivariate normal population $N(\boldsymbol{\mu}, \mathbf{I}_p)$. Then, the sample covariance matrix is defined as

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})',$$

where $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. When n tends to infinity, $\mathbf{S}_n \to \mathbf{I}_p$ and $\sqrt{n}\,(\mathbf{S}_n - \mathbf{I}_p) \to \sqrt{p}\mathbf{W}_p$. It can be seen that the entries above the main diagonal of $\sqrt{p}\mathbf{W}_p$ are iid $N(0,1)$ and the entries on the diagonal are iid $N(0,2)$. This matrix is called the (standard) Gaussian matrix or Wigner matrix.

A generalized definition of Wigner matrix only requires the matrix to be a Hermitian random matrix whose entries on or above the diagonal are independent. The study of spectral analysis of the large dimensional Wigner matrix dates back to Wigner's [295] famous **semicircular law**. He proved that the expected ESD of an $n \times n$ standard Gaussian matrix, normalized by $1/\sqrt{n}$, tends to the semicircular law F whose density is given by

$$F'(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2}, & \text{if } |x| \le 2, \\ 0, & \text{otherwise.} \end{cases} \tag{2.0.1}$$

This work has been extended in various aspects. Grenander [136] proved that $\|F^{\mathbf{W}_n} - F\| \to 0$ in probability. Further, this result was improved as in the sense of "almost sure" by Arnold [8, 7]. Later on, this result was further generalized, and it will be introduced in the following sections.

Z. Bai and J.W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices, Second Edition*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0661-8_2,

2.1 Semicircular Law by the Moment Method

In order to apply the moment method (see Appendix B, Section B.1) to prove the convergence of the ESD of Wigner matrices to the semicircular distribution, we calculate the moments of the semicircular distribution and show that they satisfy the Carleman condition. In the remainder of this section, we will show the convergence of the ESD of the Wigner matrix by the moment method.

2.1.1 Moments of the Semicircular Law

Let β_k denote the k-th moment of the semicircular law. We have the following lemma.

Lemma 2.1. *For $k = 0, 1, 2, \cdots$, we have*

$$\beta_{2k} = \frac{1}{k+1}\binom{2k}{k},$$
$$\beta_{2k+1} = 0.$$

Proof. Since the semicircular distribution is symmetric about 0, thus we have $\beta_{2k+1} = 0$. Also, we have

$$\begin{aligned}
\beta_{2k} &= \frac{1}{2\pi}\int_{-2}^{2} x^{2k}\sqrt{4-x^2}dx \\
&= \frac{1}{\pi}\int_{0}^{2} x^{2k}\sqrt{4-x^2}dx \\
&= \frac{2^{2k+1}}{\pi}\int_{0}^{1} y^{k-1/2}(1-y)^{1/2}dy \quad (\text{by setting } \ x = 2\sqrt{y}) \\
&= \frac{2^{2k+1}}{\pi}\frac{\Gamma(k+1/2)\Gamma(3/2)}{\Gamma(k+2)} = \frac{1}{k+1}\binom{2k}{k}.
\end{aligned}$$

2.1.2 Some Lemmas in Combinatorics

In order to calculate the limits of moments of the ESD of a Wigner matrix, we need some information from combinatorics. This is because the mean and variance of each empirical moment will be expressed as a sum of expectations of products of matrix entries, and we need to be able to systematically count the number of significant terms. To this end, we introduce some concepts from graph theory and establish some lemmas.

A graph is a triple (E, V, F), where E is the set of edges, V is the set of vertices, and F is a function, $F: E \mapsto V \times V$. If $F(e) = (v_1, v_2)$, the vertices v_1, v_2 are called the ends of the edge e, v_1 is the initial of e, and v_2 is the terminal of e. If $v_1 = v_2$, edge e is a loop. If two edges have the same set of ends, they are said to be coincident.

Let $\mathbf{i} = (i_1, \cdots, i_k)$ be a vector valued on $\{1, \cdots, n\}^k$. With the vector $\mathbf{i}$, we define a Γ-graph as follows. *Draw a horizontal line and plot the numbers $i_1, \cdots, i_k$ on it. Consider the distinct numbers as vertices, and draw k edges e_j from i_j to i_{j+1}, $j = 1, \cdots, k$, where $i_{k+1} = i_1$ by convention. Denote the number of distinct i_j's by t.* Such a graph is called a $\Gamma(k, t)$-graph. An example of $\Gamma(6, 4)$ is shown in Fig. 2.1.

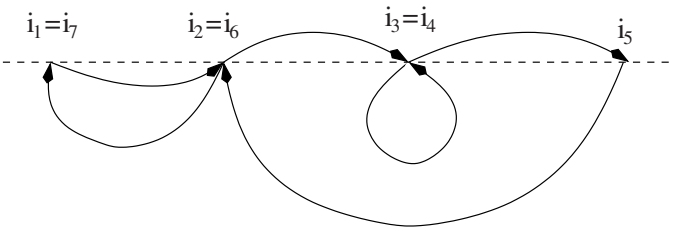

Fig. 2.1 A Γ-graph

By definition, a $\Gamma(k, t)$-graph starts from vertex i_1, and the k edges consecutively connect one after another and finally return to vertex i_1. That is, a $\Gamma(k, t)$-graph forms a cycle.

Two $\Gamma(k, t)$-graphs are said to be isomorphic if one can be converted to the other by a permutation of $(1, \cdots, n)$. By this definition, all Γ-graphs are classified into isomorphism classes.

We shall call the $\Gamma(k, t)$-graph canonical if it has the following properties:

1. Its vertex set is $V = \{1, \cdots, t\}$.
2. Its edge set is $E = \{e_1, \cdots, e_k\}$.
3. There is a function g from $\{1, 2, \cdots, k\}$ onto $\{1, 2, \cdots, t\}$ satisfying $g(1) = 1$ and $g(i) \leq \max\{g(1), \cdots, g(i-1)\} + 1$ for $1 < i \leq k$.
4. $F(e_i) = (g(i), g(i+1))$, for $i = 1, \cdots, k$, with convention $g(k+1) = g(1) = 1$.

It is easy to see that each isomorphism class contains one and only one canonical Γ-graph that is associated with a function g, and a general graph in this class can be defined by $F(e_j) = (i_{g(j)}, i_{g(j+1)})$. Therefore, we have the following lemma.

Lemma 2.2. *Each isomorphism class contains $n(n-1)\cdots(n-t+1)$ $\Gamma(k, t)$ graphs.*

The canonical $\Gamma(k,t)$-graphs can be classified into three categories.

Category 1 (denoted by $\Gamma_1(k)$): A canonical graph $\Gamma(k,t)$ is said to belong to category 1 if each edge is coincident with exactly one other edge of opposite direction and the graph of noncoincident edges forms a tree (i.e., a connected graph without cycles). It is obvious that there is no $\Gamma_1(k)$ if k is odd.

Category 2 ($\Gamma_2(k,t)$) consists of all those canonical $\Gamma(k,t)$-graphs that have at least one single edge; i.e., an edge not coincident with any other edges.

Category 3 ($\Gamma_3(k,t)$) consists of all other canonical $\Gamma(k,t)$-graphs. If we classify the k edges into coincidence classes, a $\Gamma_3(k,t)$-graph contains either a coincidence class of at least three edges or a cycle of noncoincident edges. In both cases, $t \le (k+1)/2$. Then, in fact we have proved the following lemma.

Lemma 2.3. *In a $\Gamma_3(k,t)$-graph, $t \le (k+1)/2$.*

Now, we begin to count the number of $\Gamma_1(k)$-graphs for $k = 2m$. We have the following lemma.

Lemma 2.4. *The number of $\Gamma_1(2m)$-graphs is $\frac{1}{m+1}\binom{2m}{m}$.*

Proof. Suppose G is a graph of $\Gamma_1(2m)$. We define a function $H : E \to \{-1,1\}$; $H(e) = +1$ if e is single up to itself (called an innovation) and $= -1$ otherwise (called a Type 3 (T_3) edge, the edge that coincides with an innovation that is single up to it). Corresponding to the graph G, we call the sequence $(H(e_1),\cdots,H(e_k)) = (a_1 = 1, a_2, \cdots, a_{2m-1}, a_{2m} = -1)$ the characteristic sequence of the graph G. By definition, all partial sums of the characteristic sequence are nonnegative; i.e., for all $1 \le \ell \le 2m$,

$$a_1 + a_2 + \cdots + a_\ell \ge 0. \tag{2.1.1}$$

We show that there is a one-to-one correspondence between $\Gamma_1(2m)$-graphs and the characteristic sequences. That is, we need to show that any sequence of ± 1 satisfying (2.1.1) corresponds to a $\Gamma_1(2m)$-graph. Suppose $(a_1,\cdots,a_{2m})$ is a given sequence satisfying (2.1.1). We construct a $\Gamma_1(2m)$-graph with the given sequence as its characteristic sequence.

By (2.1.1), $a_1 = 1$ and $F(e_1) = (1,2)$; i.e., $g(1) = 1$, $g(2) = 2$. Suppose $g(1), g(2), \cdots, g(s)$ $(2 \le s < 2m)$ have been defined with the following properties:

(i) For each $i \le s$, we have $g(i) \le \max\{g(1),\cdots,g(i-1)\} + 1$.
(ii) If we define $(g(i), g(i+1))$, $i = 1,\cdots,s-1$, as edges, then from $g(1) = 1$ to $g(s)$ there is a path of single innovations if $g(s) \ne 1$. All other edges not on the path must coincide with another edge of opposite direction. If $g(s) = 1$, then each edge coincides with another edge of opposite direction.
(iii) $H(g(i), g(i+1)) = a_i$ for all $i < s$.

Now, we define $g(s+1)$ in the following way:

Case 1. If $a_s = 1$, define $g(s+1) = \max\{g(1), \cdots, g(s)\} + 1$. Obviously, the edge $(g(s), g(s+1))$ is a single innovation that, combining the original path of single innovations, forms the new path of single innovations from $g(1) = 1$ to $g(s+1)$ if $g(s) \neq 1$. If $g(s) = 1$, then $g(s+1) \neq 1$ and the edge $(g(s), g(s+1))$ forms the new path of single innovations. Also, all other edges coincide with an edge of opposite directions. That is, conditions (i)–(iii) are satisfied.

Case 2. If $a_s = -1$, then $g(s) \neq 1$ for otherwise condition (2.1.1) will be violated. Hence, there is an $i < s$ such that $(g(i), g(s))$ is a single innovation (the last edge of a path of single innovations). Then, define $g(s+1) = g(i)$. If $g(i) = 1$, then the new graph has no single edges. If $g(i) \neq 1$, the original path of single innovations has at least two single innovations. Then, the new path of single innovations is obtained by cutting the last edge from the original path of single innovations. Also, conditions (i)–(iii) are satisfied.

By induction, the functions $g(1), \cdots, g(2m)$ are well defined, and hence a $\Gamma_1(2m)$ with characteristic sequence $(a_1, \cdots, a_{2m})$ is defined.

Therefore, to count the number of $\Gamma_1(2m)$-graphs is equivalent to counting the number of characteristic sequences of isomorphism classes.

Arbitrarily arrange m ones and m minus ones. The total number of possibilities is obviously $\binom{2m}{m}$. We shall use the symmetrization principle to count the number of noncharacteristic sequences. Write the sequence of ± 1s as $(a_1, \cdots, a_{2m})$ and $S_0 = 0$ and $S_i = S_{i-1} + a_i$, for $i = 1, 2, \cdots, 2m$. Plot the graph of $(i, S(i))$ on the plane. The graph should start from $(0, 0)$ and return to $(2m, 0)$. If for all i, $S_i \geq 0$ (that is, the figure is totally above or on the horizontal axis), then $(a_1, \cdots, a_{2m})$ is a characteristic sequence. Otherwise, if $(a_1, \cdots, a_{2m})$ is not a characteristic sequence, then there must be an $i \geq 1$ such that $S_i = -1$. Then we turn over the rear part after i along the line $S = -1$ and we get a new graph $(0, 0)$ to $(2m, -2)$, as shown in Fig. 2.2.

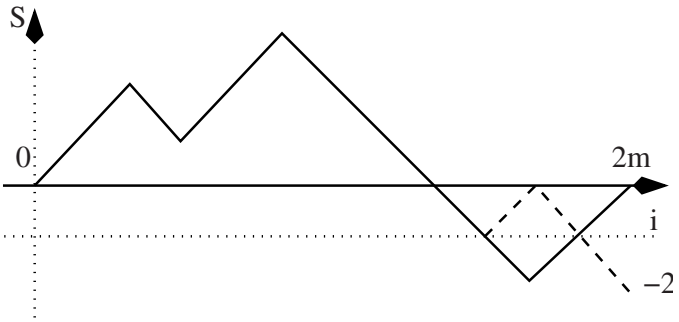

Fig. 2.2 Symmetrization principle

This is equivalent to defining $b_j = a_j$ for $j \leq i$ and $b_j = -a_j$ for $j > i$. Then, the sequence $(b_1, \cdots, b_{2m})$ contains $m - 1$ ones and $m + 1$ minus ones.

Conversely, for any sequence of $m-1$ ones and $m+1$ minus ones, there must be a smallest integer $i < 2m$ such that $b_1 + \cdots + b_i = -1$. Then the sequence $(b_1, \cdots, b_k, -b_{k+1}, \cdots, -b_{2m})$ contains m ones and m minus ones which is a noncharacteristic sequence. The number of b-sequences is $\binom{2m}{m-1}$. Thus, the number of characteristic sequences is

$$\binom{2m}{m} - \binom{2m}{m-1} = \frac{1}{m+1}\binom{2m}{m}.$$

The proof of the lemma is complete.

2.1.3 Semicircular Law for the iid Case

In this subsection, we will show the semicircular law for the iid case; that is, we shall prove the following theorem. For brevity of notation, we shall use $\mathbf{X}_n$ for an $n \times n$ Wigner matrix and save the notation $\mathbf{W}_n$ for the normalized Wigner matrix, i.e., $\frac{1}{\sqrt{n}}\mathbf{X}_n$.

Theorem 2.5. *Suppose that $\mathbf{X}_n$ is an $n \times n$ Hermitian matrix whose diagonal entries are iid real random variables and those above the diagonal are iid complex random variables with variance $\sigma^2 = 1$. Then, with probability 1, the ESD of $\mathbf{W}_n = \frac{1}{\sqrt{n}}\mathbf{X}_n$ tends to the semicircular law.*

Before applying the MCT to the proof of Theorem 2.5, we first remove the diagonal entries of $\mathbf{X}_n$, truncate the off-diagonal entries of the matrix, and renormalize them, without changing the LSD. We will proceed with the proof by taking the following steps.

Step 1. Removing the Diagonal Elements
Let $\widetilde{\mathbf{W}}_n$ be the matrix obtained from $\mathbf{W}_n$ by replacing the diagonal elements with zero. We shall show that the two matrices are asymptotically equivalent; i.e., their LSDs are the same if one of them exists.

Let $N_n = \#\{|x_{ii}| \geq \sqrt[4]{n}\}$. Replace the diagonal elements of $\mathbf{W}_n$ by $\frac{1}{\sqrt{n}}x_{ii}I(|x_{ii}| < \sqrt[4]{n})$, and denote the resulting matrix by $\widehat{\mathbf{W}}_n$. Then, by Corollary A.41, we have

$$L^3(F^{\widehat{\mathbf{W}}_n}, F^{\widetilde{\mathbf{W}}_n}) \leq \frac{1}{n}\mathrm{tr}[(\widetilde{\mathbf{W}}_n - \widehat{\mathbf{W}}_n)^2] \leq \frac{1}{n^2}\sum_{i=1}^{n}|x_{ii}|^2 I(|x_{ii}| < \sqrt[4]{n}) \leq \frac{1}{\sqrt{n}}.$$

On the other hand, by Theorem A.43, we have

$$\left\| F^{\mathbf{W}_n} - F^{\widetilde{\mathbf{W}}_n} \right\| \leq \frac{N_n}{n}.$$

Therefore, to complete the proof of our assertion, it suffices to show that $N_n/n \to 0$ almost surely. Write $p_n = \mathrm{P}(|x_{11}| \geq \sqrt[4]{n}) \to 0$. By Bernstein's inequality,[1] we have, for any $\varepsilon > 0$,

$$\begin{aligned}\mathrm{P}(N_n \geq \varepsilon n) &= \mathrm{P}\left(\sum_{i=1}^{n}(I(|x_{ii}| \geq \sqrt[4]{n}) - p_n) \geq (\varepsilon - p_n)n\right) \\ &\leq 2\exp(-(\varepsilon - p_n)^2 n^2/2[np_n + (\varepsilon - p_n)n]) \leq 2e^{-bn},\end{aligned}$$

for some positive constant $b > 0$. This completes the proof of our assertion.

In the following subsections, we shall assume that the diagonal elements of $\mathbf{W}_n$ are all zero.

Step 2. Truncation
For any fixed positive constant C, truncate the variables at C and write $x_{ij(C)} = x_{ij}I(|x_{ij}| \leq C)$. Define a truncated Wigner matrix $\mathbf{W}_{n(C)}$ whose diagonal elements are zero and off-diagonal elements are $\frac{1}{\sqrt{n}}x_{ij(C)}$. Then, we have the following truncation lemma.

Lemma 2.6. *Suppose that the assumptions of Theorem 2.5 are true. Truncate the off-diagonal elements of $\mathbf{X}_n$ at C, and denote the resulting matrix by $\mathbf{X}_{n(C)}$. Write $\mathbf{W}_{n(C)} = \frac{1}{\sqrt{n}}\mathbf{X}_{n(C)}$. Then, for any fixed constant C,*

$$\limsup_n L^3(F^{\mathbf{W}_n}, F^{\mathbf{W}_{n(C)}}) \leq \mathrm{E}\big(|x_{11}|^2 I(|x_{11}| > C)\big), \quad \text{a.s.} \tag{2.1.2}$$

Proof. By Corollary A.41 and the law of large numbers, we have

$$\begin{aligned}L^3(F^{\mathbf{W}_n}, F^{\mathbf{W}_{n(C)}}) &\leq \frac{2}{n^2}\left(\sum_{1\leq i<j\leq n}|x_{ij}|^2 I(|x_{11}| > C)\right) \\ &\to \mathrm{E}\big(|x_{11}|^2 I(|x_{11}| > C)\big).\end{aligned}$$

This completes the proof of the lemma.

Note that the right-hand side of (2.1.2) can be made arbitrarily small by making C large. Therefore, in the proof of Theorem 2.5, we can assume that the entries of the matrix $\mathbf{X}_n$ are uniformly bounded.

Step 3. Centralization
Applying Theorem A.43, we have

$$\left\|F^{\mathbf{W}_{n(C)}} - F^{\mathbf{W}_{n(C)} - a\mathbf{1}\mathbf{1}'}\right\| \leq \frac{1}{n}, \tag{2.1.3}$$

[1] Bernstein's inequality states that if $X_1, \cdots, X_n$ are independent random variables with mean zero and uniformly bounded by b, then, for any $\varepsilon > 0$,
$P(|S_n| \geq \varepsilon) \leq 2\exp(-\varepsilon^2/[2(B_n^2 + b\varepsilon)])$, where $S_n = X_1 + \cdots + X_n$ and $B_n^2 = \mathrm{E}S_n^2$.

where $a = \frac{1}{\sqrt{n}}\Re(\mathrm{E}(x_{12(C)}))$. Furthermore, by Corollary A.41, we have

$$L(F^{\mathbf{W}_{n(C)}-\Re(\mathrm{E}(\mathbf{W}_{n(C)}))}, F^{\mathbf{W}_{n(C)}-a\mathbf{1}\mathbf{1}'}) \leq \frac{|\Re(\mathrm{E}(x_{12(C)}))|^2}{n} \to 0. \tag{2.1.4}$$

This shows that we can assume that the real parts of the mean values of the off-diagonal elements are 0. In the following, we proceed to remove the imaginary part of the mean values of the off-diagonal elements.

Before we treat the imaginary part, we introduce a lemma about eigenvalues of a skew-symmetric matrix.

Lemma 2.7. *Let $\mathbf{A}_n$ be an $n \times n$ skew-symmetric matrix whose elements above the diagonal are 1 and those below the diagonal are -1. Then, the eigenvalues of $\mathbf{A}_n$ are $\lambda_k = i\cot(\pi(2k-1)/2n)$, $k = 1, 2, \cdots, n$. The eigenvector associated with λ_k is $\mathbf{u}_k = \frac{1}{\sqrt{n}}(1, \rho_k, \cdots, \rho_k^{n-1})'$, where $\rho_k = (\lambda_k - 1)/(\lambda_k + 1) = \exp(-i\pi(2k-1)/n)$.*

Proof. We first compute the characteristic polynomial of $\mathbf{A}_n$.

$$D_n = |\lambda\mathbf{I} - \mathbf{A}_n| = \begin{vmatrix} \lambda & -1 & -1 & \cdots & -1 \\ 1 & \lambda & -1 & \cdots & -1 \\ 1 & 1 & \lambda & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & \lambda \end{vmatrix}$$
$$= \begin{vmatrix} \lambda - 1 & -(1+\lambda) & 0 & \cdots & 0 \\ 0 & \lambda - 1 & -(1+\lambda) & \cdots & 0 \\ 0 & 0 & \lambda - 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & \lambda \end{vmatrix}.$$

Expanding the above along the first row, we get the following recursive formula

$$D_n = (\lambda - 1)D_{n-1} + (1+\lambda)^{n-1},$$

with the initial value $D_1 = \lambda$. The solution is

$$\begin{aligned} D_n &= \lambda(\lambda-1)^{n-1} + (\lambda+1)(\lambda-1)^{n-2} + \cdots + (\lambda+1)^{n-1} \\ &= \frac{1}{2}\left((\lambda-1)^n + (\lambda+1)^n\right). \end{aligned}$$

Setting $D_n = 0$, we get

$$\frac{\lambda+1}{\lambda-1} = e^{i\pi(2k-1)/n}, \quad k = 1, 2, \cdots, n, \tag{2.1.5}$$

which implies that $\lambda = i\cot(\pi(2k-1)/2n)$.

Comparing the two sides of the equation $\mathbf{A}_n\mathbf{u}_k = \lambda_k\mathbf{u}_k$, we obtain

$$-u_{k,1} - \cdots - u_{k,\ell-1} + u_{k,\ell+1} + \cdots + u_{k,n} = \lambda_k u_{k,\ell}$$

for $\ell = 1, 2, \cdots, n$. Thus, subtracting the equations for $\ell + 1$ from that for ℓ, we get

$$u_{k,\ell} + u_{k,\ell+1} = \lambda_k (u_{k,\ell} - u_{k,\ell+1}),$$

which implies that

$$\frac{u_{k,\ell+1}}{u_{k,\ell}} = \frac{\lambda_k - 1}{\lambda_k + 1} = e^{-i\pi(2k-1)/n} := \rho_k.$$

Therefore, one can choose $u_{k,\ell} = \rho_k^{\ell-1}/\sqrt{n}$.

The proof of the lemma is complete.

Write $b = \mathrm{E}\Im(x_{12(C)})$. Then, $\mathrm{E}\Im(\mathbf{W}_{n(C)}) = ib\mathbf{A}_n$. By Lemma 2.7, the eigenvalues of the matrix $i\Im(\mathrm{E}(\mathbf{W}_{n(C)})) = ib\mathbf{A}_n$ are $ib\lambda_k = -n^{-1/2}b\cot(\pi(2k-1)/2n)$, $k = 1, \cdots, n$. If the spectral decomposition of $\mathbf{A}_n$ is $\mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^*$, then we rewrite $i\Im(\mathrm{E}(\mathbf{W}_{n(C)})) = \mathbf{B}_1 + \mathbf{B}_2$, where $\mathbf{B}_j = -\frac{1}{\sqrt{n}}b\mathbf{U}_n\mathbf{D}_{nj}\mathbf{U}_n^*$, $j = 1, 2$, where $\mathbf{U}_n$ is a unitary matrix, $\mathbf{D}_n$ =diag$[\lambda_1, \cdots, \lambda_n]$, and

$$\mathbf{D}_{n1} = \mathbf{D}_n - \mathbf{D}_{n2} = \mathrm{diag}[0, \cdots, 0, \lambda_{[n^{3/4}]}, \lambda_{[n^{3/4}]+1}, \cdots, \lambda_{n-[n^{3/4}]}, 0, \cdots, 0].$$

For any $n \times n$ Hermitian matrix $\mathbf{C}$, by Corollary A.41, we have

$$\begin{aligned} L^3(F^{\mathbf{C}}, F^{\mathbf{C}-\mathbf{B}_1}) &\le \frac{1}{n^2} \sum_{n^{3/4} \le k \le n - n^{3/4}} \cot^2(\pi(2k-1)/2n) \\ &< \frac{2}{n \sin^2(n^{-1/4}\pi)} \to 0 \end{aligned} \tag{2.1.6}$$

and, by Theorem A.43,

$$\|F^{\mathbf{C}} - F^{\mathbf{C}-\mathbf{B}_2}\| \le \frac{2n^{3/4}}{n} \to 0. \tag{2.1.7}$$

Summing up estimations (2.1.3)–(2.1.7), we established the following centralization lemma.

Lemma 2.8. *Under the conditions assumed in Lemma 2.6, we have*

$$L(F^{\mathbf{W}_{n(C)}}, F^{\mathbf{W}_{n(C)} - \mathrm{E}(\mathbf{W}_{n(C)})}) = o(1). \tag{2.1.8}$$

Step 4. Rescaling

Write $\sigma^2(C) = \mathrm{Var}(x_{11(C)})$, and define $\widetilde{\mathbf{W}}_n = \sigma^{-1}(C)\big(\mathbf{W}_{n(C)} - \mathrm{E}(\mathbf{W}_{n(C)})\big)$. Note that the off-diagonal entries of $\sqrt{n}\widetilde{\mathbf{W}}_n$ are $\widehat{x}_{kj} = \sigma^{-1}(C)\big(x_{kj(C)} - \mathrm{E}(x_{kj(C)})\big)$.

Applying Corollary A.41, we obtain

$$L^3(F^{\widetilde{\mathbf{W}}_n}, F^{\mathbf{W}_{n(C)}-\mathrm{E}(\mathbf{W}_{n(C)})}) \le \frac{2(\sigma(C)-1)^2}{n^2\sigma^2(C)} \sum_{1\le i<j\le n} |x_{kj(C)} - \mathrm{E}(x_{kj(C)})|^2$$
$$\to (\sigma(C)-1)^2, \quad \text{a.s.} \tag{2.1.9}$$

Note that $(\sigma(C)-1)^2$ can be made arbitrarily small if C is large. Combining (2.1.9) with Lemmas 2.6 and 2.8, to prove the semicircular law, we may assume that the entries of $\mathbf{X}$ are bounded by C, having mean zero and variance 1. Also, we may assume the diagonal elements are zero.

Step 5. Proof of the Semicircular Law

We will prove Theorem 2.5 by the moment method. For simplicity, we still use $\mathbf{W}_n$ and x_{ij} to denote the Wigner matrix and basic variables after truncation, centralization, and rescaling.

The semicircular distribution satisfies the Riesz condition. Therefore it is enough to show that the moments of the spectral distribution converge to the corresponding moments of the semicircular distribution almost surely. The k-th moment of the ESD of $\mathbf{W}_n$ is

$$\begin{aligned}\beta_k(\mathbf{W}_n) = \beta_k(F^{\mathbf{W}_n}) &= \int x^k \, dF^{\mathbf{W}_n}(x) \\ &= \frac{1}{n}\sum_{i=1}^{n} \lambda_i^k = \frac{1}{n}\mathrm{tr}(\mathbf{W}_n^k) = \frac{1}{n^{1+\frac{k}{2}}}\mathrm{tr}(\mathbf{X}_n^k) \\ &= \frac{1}{n^{1+\frac{k}{2}}}\sum_{\mathbf{i}} X(\mathbf{i}),\end{aligned} \tag{2.1.10}$$

where λ_i's are the eigenvalues of the matrix $\mathbf{W}_n$, $X(\mathbf{i}) = x_{i_1 i_2} x_{i_2 i_3} \cdots x_{i_k i_1}$, $\mathbf{i} = (i_1, \cdots, i_k)$, and the summation $\sum_{\mathbf{i}}$ runs over all possibilities that $\mathbf{i} \in \{1, \cdots, n\}^k$.

By applying the moment convergence theorem, we complete the proof of the semicircular law for the iid case by showing the following:

(1) $E[\beta_k(\mathbf{W}_n)]$ converges to the k-th moment β_k of the semicircular distribution, which are $\beta_{2m-1} = 0$ and $\beta_{2m} = (2m)!/m!(m+1)!$ given in Lemma 2.1.
(2) For each fixed k, $\sum_n \mathrm{Var}[\beta_k(\mathbf{W}_n)] < \infty$.

The Proof of (1); i.e., $\mathrm{E}[\beta_k(\mathbf{W}_n)] \to \beta_k$.

We have

$$\mathrm{E}[\beta_k(\mathbf{W}_n)] = \frac{1}{n^{1+k/2}} \sum \mathrm{E}X(\mathbf{i}).$$

For each vector $\mathbf{i}$, construct a graph $G(\mathbf{i})$ as in Subsection 2.1.2. To specify the graph, we rewrite $X(\mathbf{i}) = X(G(\mathbf{i}))$. The summation is taken over all sequences $\mathbf{i} = (i_1, i_2, \cdots, i_k) \in \{1, 2, \cdots, n\}^k$.

Note that isomorphic graphs correspond to equal terms. Thus, we first group the terms according to isomorphism classes and then split $\mathrm{E}[\beta_k(\mathbf{W}_n)]$ into three sums according to categories. Then

$$\mathrm{E}[\beta_k(\mathbf{W}_n)] = S_1 + S_2 + S_3,$$

where

$$S_j = n^{-1-k/2} \sum_{\Gamma(k,t)\in C_j} \sum_{G(\mathbf{i})\in\Gamma(k,t)} \mathrm{E}[XG(\mathbf{i})],$$

in which the summation $\sum_{\Gamma(k,t)\in C_j}$ is taken on all canonical $\Gamma(k,t)$-graphs in category j and the summation $\sum_{G(\mathbf{i})\in\Gamma(k,t)}$ is taken on all isomorphic graphs for a given canonical graph.

By definition of the categories and by the assumptions on the entries of the random matrices, we have

$$S_2 = 0.$$

Since the random variables are bounded by C, the number of isomorphic graphs is less than n^t by Lemma 2.2, and $t \le (k+1)/2$ by Lemma 2.3, we conclude that

$$|S_3| \le n^{-1-k/2} O(n^t) = o(1).$$

If $k = 2m-1$, then $S_1 = 0$ since there are no terms in S_1. We consider the case where $k = 2m$. Since each edge coincides with an edge of opposite direction, each term in S_1 is $(\mathrm{E}|x_{12}|^2)^m = 1$. So, by Lemma 2.4,

$$\begin{aligned} S_1 &= n^{-1-m} \sum_{\Gamma(2m,t)\in C_1} n(n-1)\cdots(n-m) \\ &= \beta_{2m}\left(1-\frac{1}{m}\right)\cdots\left(1-\frac{m}{n}\right) \to \beta_{2m}. \end{aligned}$$

Assertion (1) is then proved.

The proof of (2). We only need to show that $\mathrm{Var}(\beta_k(\mathbf{W}_n))$ is summable for all fixed k. We have

$$\begin{aligned} &\mathrm{Var}(\beta_k(\mathbf{W}_n)) = E[|\beta_k(\mathbf{W}_n)|^2] - |E[\beta_k(\mathbf{W}_n)]|^2 \\ &= \frac{1}{n^{2+k}} \sum{}^{*} \{\mathrm{E}[X(\mathbf{i})X(\mathbf{j})] - \mathrm{E}[X(\mathbf{i})]\mathrm{E}[X(\mathbf{j})]\}, \end{aligned} \tag{2.1.11}$$

where $\mathbf{i} = (i_1, \cdots, i_k)$, $\mathbf{j} = (j_1, \cdots, j_k)$, and $\sum^*$ is taken over all possibilities for $\mathbf{i}$, $\mathbf{j} \in \{1, \cdots, n\}^k$. Here, the reader should notice that $\beta_k(\mathbf{W}_n)$ is real and hence the second equality in the above is meaningful, although the variables $X(\mathbf{i})$ and $X(\mathbf{j})$ are complex.

Using $\mathbf{i}$ and $\mathbf{j}$, one can construct two graphs $G(\mathbf{i})$ and $G(\mathbf{j})$, as in the proof of (1). If there are no coincident edges between $G(\mathbf{i})$ and $G(\mathbf{j})$, then $X(\mathbf{i})$ is

independent of $X(\mathbf{j})$, and thus the corresponding term in the sum is 0. If the combined graph $G = G(\mathbf{i}) \cup G(\mathbf{j})$ has a single edge, then $\mathrm{E}[X(\mathbf{i})X(\mathbf{j})] = \mathrm{E}[X(\mathbf{i})]\mathrm{E}[X(\mathbf{j})] = 0$, and hence the corresponding term in (2.1.11) is also 0.

Now, suppose that G contains no single edges and the graph of noncoincident edges has a cycle. Then the noncoincident vertices of G are not more than k. If G contains no single edges and the graph of noncoincident edges has no cycles, then there is at least one edge with coincidence multiplicity greater than or equal to 4, and thus the number of noncoincident vertices is not larger than k. Also, each term in (2.1.11) is not larger than $2C^{2k}n^{-2-k}$. Consequently, we can conclude that

$$\mathrm{Var}(\beta_k(\mathbf{W}_n)) \le K_k C^{2k} n^{-2}, \tag{2.1.12}$$

where K_k is a constant that depends on k only. This completes the proof of assertion (2).

The proof of Theorem 2.5 is then complete.

2.2 Generalizations to the Non-iid Case

Sometimes, it is of practical interest to consider the case where, for each n, the entries above or on the diagonal of $\mathbf{W}_n$ are independent complex random variables with mean zero and variance σ^2 (for simplicity we assume $\sigma = 1$ in the following), but may depend on n. For this case, we present the following theorem.

Theorem 2.9. *Suppose that $\mathbf{W}_n = \frac{1}{\sqrt{n}}\mathbf{X}_n$ is a Wigner matrix and the entries above or on the diagonal of $\mathbf{X}_n$ are independent but may be dependent on n and may not necessarily be identically distributed. Assume that all the entries of $\mathbf{X}_n$ are of mean zero and variance 1 and satisfy the condition that, for any constant $\eta > 0$,*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{jk} \mathrm{E}|x_{jk}^{(n)}|^2 I(|x_{jk}^{(n)}| \ge \eta\sqrt{n}) = 0. \tag{2.2.1}$$

Then, the ESD of $\mathbf{W}_n$ converges to the semicircular law almost surely.

Remark 2.10. In Girko's book [121], it is stated that condition (2.2.1) is necessary and sufficient for the conclusion of Theorem 2.9.

2.2.1 Proof of Theorem 2.9

Again, we need to truncate, remove diagonal entries, and renormalize before we use the MCT. Because the entries are not iid, we cannot truncate the

entries at constant positions. Instead, we shall truncate them at $\eta_n\sqrt{n}$ for some sequence $\eta_n \downarrow 0$.

Step 1. Truncation
Note that Corollary A.41 may not be applicable in proving the almost sure asymptotic equivalence between the ESD of the original matrix and that of the truncated one, as was done in the last section. In this case, we shall use the rank inequality (see Theorem A.43) to truncate the variables.

Note that condition (2.2.1) is equivalent to: for any $\eta > 0$,

$$\lim_{n\to\infty} \frac{1}{\eta^2 n^2} \sum_{jk} \mathrm{E}|x_{jk}^{(n)}|^2 I(|x_{jk}^{(n)}| \geq \eta\sqrt{n}) = 0. \tag{2.2.2}$$

Thus, one can select a sequence $\eta_n \downarrow 0$ such that (2.2.2) remains true when η is replaced by η_n. Define $\widetilde{\mathbf{W}}_n = \frac{1}{\sqrt{n}} n(x_{ij}^{(n)} I(|x_{ij}^{(n)}| \leq \eta_n\sqrt{n})$. By using Theorem A.43, one has

$$\begin{aligned} \|F^{\mathbf{W}_n} - F^{\widetilde{\mathbf{W}}_n}\| &\leq \frac{1}{n}\mathrm{rank}(\mathbf{W}_n - \mathbf{W}_{n(\eta_n\sqrt{n})}) \\ &\leq \frac{2}{n} \sum_{1\leq i\leq j\leq n} I(|x_{ij}^{(n)}| \geq \eta_n\sqrt{n}). \end{aligned} \tag{2.2.3}$$

By condition (2.2.2), we have

$$\begin{aligned} &\mathrm{E}\left(\frac{1}{n} \sum_{1\leq i\leq j\leq n} I(|x_{ij}^{(n)}| \geq \eta_n\sqrt{n})\right) \\ &\leq \frac{2}{\eta_n^2 n^2} \sum_{jk} \mathrm{E}|x_{ij}^{(n)}|^2 I(|x_{ij}^{(n)}| \geq \eta_n\sqrt{n}) = o(1), \end{aligned}$$

and

$$\begin{aligned} &\mathrm{Var}\left(\frac{1}{n} \sum_{1\leq i\leq j\leq n} I(|x_{ij}^{(n)}| \geq \eta_n\sqrt{n})\right) \\ &\leq \frac{4}{\eta_n^2 n^3} \sum_{jk} \mathrm{E}|x_{ij}^{(n)}|^2 I(|x_{ij}^{(n)}| \geq \eta_n\sqrt{n}) = o(1/n). \end{aligned}$$

Then, applying Bernstein's inequality, for all small $\varepsilon > 0$ and large n, we have

$$\mathrm{P}\left(\frac{1}{n} \sum_{1\leq i\leq j\leq n} I(|x_{ij}^{(n)}| \geq \eta_n\sqrt{n}) \geq \varepsilon\right) \leq 2e^{-\varepsilon n}, \tag{2.2.4}$$

which is summable. Thus, by (2.2.3) and (2.2.4), to prove that with probability one $F^{\mathbf{W}_n}$ converges to the semicircular law, it suffices to show that with probability one $F^{\widetilde{\mathbf{W}}_n}$ converges to the semicircular law.

Step 2. Removing diagonal elements
Let $\widehat{\mathbf{W}}_n$ be the matrix $\widetilde{\mathbf{W}}_n$ with diagonal elements replaced by 0. Then, by Corollary A.41, we have

$$L^3\left(F^{\widetilde{\mathbf{W}}_n}, F^{\widehat{\mathbf{W}}_n}\right) \le \frac{1}{n^2}\sum_{k=1}^{n} |x_{kk}^{(n)}|^2 I(|x_{kk}^{(n)}| \le \eta_n\sqrt{n}) \le \eta_n^2 \to 0.$$

Step 3. Centralization
By Corollary A.41, it follows that

$$\begin{aligned} &L^3\left(F^{\widehat{\mathbf{W}}_n}, F^{\widehat{\mathbf{W}}_n - \mathrm{E}\widehat{\mathbf{W}}_n}\right) \\ &\le \frac{1}{n^2}\sum_{i\neq j} |\mathrm{E}(x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}))|^2 \\ &\le \frac{1}{n^3\eta_n^2}\sum_{ij} \mathrm{E}|x_{jk}^{(n)}|^2 I(|x_{jk}^{(n)}| \ge \eta_n\sqrt{n}) \to 0. \end{aligned} \tag{2.2.5}$$

Step 4. Rescaling
Write $\widetilde{\mathbf{W}}_n = \frac{1}{\sqrt{n}}\widetilde{\mathbf{X}}_n$, where

$$\widetilde{\mathbf{X}}_n = \left(\frac{x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}) - \mathrm{E}(x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}))}{\sigma_{ij}}(1-\delta_{ij})\right),$$

$\sigma_{ij}^2 = \mathrm{E}|x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}) - \mathrm{E}(x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}))|^2$ and δ_{ij} is Kronecker's delta.

By Corollary A.41, it follows that

$$\begin{aligned} &L^3\left(F^{\widetilde{\mathbf{W}}_n}, F^{\widehat{\mathbf{W}}_n - \mathrm{E}\widehat{\mathbf{W}}_n}\right) \\ &\le \frac{1}{n^2}\sum_{i\neq j}(1-\delta_{ij}^{-1})^2 |x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}) - \mathrm{E}(x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}))|^2. \end{aligned}$$

Note that

$$\begin{aligned} &\mathrm{E}\left(\frac{1}{n^2}\sum_{i\neq j}(1-\delta_{ij}^{-1})^2 |x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}) - \mathrm{E}(x_{ij}^{(n)} I(|x_{ij}^{(n)}| \le \eta_n\sqrt{n}))|^2\right) \\ &\le \frac{1}{n^2\eta_n^2}\sum_{ij}(1-\sigma_{ij})^2 \end{aligned}$$

$$\leq \frac{1}{n^2\eta_n^2}\sum_{ij}(1-\sigma_{ij}^2)$$
$$\leq \frac{1}{n^2\eta_n^2}\sum_{ij}[\mathrm{E}|x_{jk}^{(n)}|^2 I(|x_{jk}^{(n)}|\geq \eta_n\sqrt{n}) + \mathrm{E}^2|x_{jk}^{(n)}|I(|x_{jk}^{(n)}|\geq \eta_n\sqrt{n})] \to 0.$$

Also, we have[2]

$$\mathrm{E}\left|\frac{1}{n^2}\sum_{i\neq j}(1-\delta_{ij}^{-1})^2\left|x_{ij}^{(n)}I(|x_{ij}^{(n)}|\leq \eta_n\sqrt{n}) - \mathrm{E}(x_{ij}^{(n)}I(|x_{ij}^{(n)}|\leq \eta_n\sqrt{n}))\right|^2\right|^4$$
$$\leq \frac{C}{n^8}\left[\sum_{i\neq j}\mathrm{E}|x_{ij}^{(n)}|^8 I(|x_{ij}^{(n)}|\leq \eta_n\sqrt{n}) + \left(\sum_{i\neq j}\mathrm{E}|x_{ij}^{(n)}|^4 I(|x_{ij}^{(n)}|\leq \eta_n\sqrt{n})\right)^2\right]$$
$$\leq Cn^{-2}[n^{-1}\eta_n^6 + \eta_n^4],$$

which is summable. From the two estimates above, we conclude that

$$L\left(F^{\widetilde{\mathbf{W}}_n}, F^{\widehat{\mathbf{W}}_n - \mathrm{E}\widehat{\mathbf{W}}_n}\right) \to 0, \text{ a.s.}$$

Step 5. Proof by MCT

Up to here, we have proved that we may truncate, centralize, and rescale the entries of the Wigner matrix at $\eta_n\sqrt{n}$ and remove the diagonal elements without changing the LSD. These four steps are almost the same as those we followed for the iid case.

Now, we assume that the variables are truncated at $\eta_n\sqrt{n}$ and then centralized and rescaled.

Again for simplicity, the truncated and centralized variables are still denoted by x_{ij}, We assume:

(i) The variables $\{x_{ij},\, 1\leq i<j\leq n\}$ are independent and $x_{ii}=0$.
(ii) $\mathrm{E}(x_{ij})=0$ and $\mathrm{Var}(x_{ij})=1$.
(iii) $|x_{ij}|\leq \eta_n\sqrt{n}$.

Similar to what we did in the last section, in order to prove Theorem 2.9, we need to show that:

(1) $E[\beta_k(\mathbf{W}_n)]$ converges to the k-th moment β_k of the semicircular distribution.
(2) For each fixed k, $\sum_n \mathrm{E}|\beta_k(\mathbf{W}_n) - \mathrm{E}(\beta_k(\mathbf{W}_n))|^4 < \infty$.

The proof of (1)
Let $\mathbf{i} = (i_1, \cdots, i_k) \in \{1, \cdots, n\}^k$. As in the iid case, we write

[2] Here we use the elementary inequality $\mathrm{E}|\sum X_i|^{2k} \leq C_k(\sum \mathrm{E}|X_i|^{2k} + (\sum \mathrm{E}|X_i|^2)^k)$ for some constant C_k if the X_i's are independent with zero means.

$$\mathrm{E}[\beta_k(\mathbf{W}_n)] = n^{-1-k/2} \sum_{\mathbf{i}} \mathrm{E}X(G(\mathbf{i})),$$

where $X(G(\mathbf{i})) = x_{i_1,i_2} x_{i_2,i_3} \cdots, x_{i_k,i_1}$, and $G(\mathbf{i})$ is the graph defined by $\mathbf{i}$.

By the same method for the iid case, we split $\mathrm{E}[\beta_k(\mathbf{W}_n)]$ into three sums according to the categories of graphs. We know that the terms in S_2 are all zero, that is, $S_2 = 0$.

We now show that $S_3 \to 0$. Split S_3 as $S_{31} + S_{32}$, where S_{31} consists of the terms corresponding to a $\Gamma_3(k,t)$-graph that contains at least one noncoincident edge with multiplicity greater than 2 and S_{32} is the sum of the remaining terms in S_3.

To estimate S_{31}, assume that the $\Gamma_3(k,t)$-graph contains ℓ noncoincident edges with multiplicities $\nu_1, \cdots, \nu_\ell$ among which at least one is greater than or equal to 3. Note that the multiplicities are subject to $\nu_1 + \cdots + \nu_\ell = k$. Also, each term in S_{31} is bounded by

$$n^{-1-k/2} \prod_{i=1}^{\ell} E|x_{a_i,b_i}|^{\nu_i} \le n^{-1-k/2} (\eta_n \sqrt{n})^{\sum(\nu_i - 2)} = n^{-1-\ell} \eta_n^{k-2\ell}.$$

Since the graph is connected and the number of its noncoincident edges is ℓ, the number of noncoincident vertices is not more than $\ell + 1$, which implies that the number of terms in S_{31} is not more than $n^{1+\ell}$. Therefore,

$$|S_{31}| \le C_k \eta_n^{k-2\ell} \to 0$$

since $k - 2\ell \ge 1$.

To estimate S_{32}, we note that the $\Gamma_3(k,t)$-graph contains exactly $k/2$ noncoincident edges, each with multiplicity 2 (thus k must be even). Then each term of S_{32} is bounded by $n^{-1-k/2}$. Since the graph is not in category 1, the graph of noncoincident edges must contain a cycle, and hence the number of noncoincident vertices is not more than $k/2$ and therefore

$$|S_{32}| \le C n^{-1} \to 0.$$

Then, the evaluation of S_1 is exactly the same as in the iid case and hence is omitted. Hence, we complete the proof of $\mathrm{E}\beta_k(\mathbf{W}_n) \to \beta_k$.

The proof of (2)

Unlike in the proof of (2.1.11), the almost sure convergence cannot follow by estimating the second moment of $\beta_k(\mathbf{W}_n)$. We need to estimate its fourth moment as

$$\begin{aligned} &\mathrm{E}(\beta_k(\mathbf{W}_n) - \mathrm{E}(\beta_k(\mathbf{W}_n)))^4 \\ &= n^{-4-2k} \sum_{\mathbf{i}_j, j=1,2,3,4} \mathrm{E} \prod_{j=1}^{4} \left(X[\mathbf{i}_j] - \mathrm{E}(X[\mathbf{i}_j])\right), \end{aligned} \tag{2.2.6}$$

where $\mathbf{i}_j$ is a vector of k integers not larger than n, $j = 1, 2, 3, 4$. As in the last section, for each $\mathbf{i}_j$, we construct a graph $G_j = G(\mathbf{i}_j)$.

Obviously, if, for some j, $G(\mathbf{i}_j)$ does not have any edges coincident with edges of the other three graphs, then the term in (2.2.6) equals 0 by independence. Also, if $G = \bigcup_{j=1}^{4} G_j$ has a single edge, the term in (2.2.6) equals 0 by centralization.

Now, let us estimate the nonzero terms in (2.2.6). Assume that G has ℓ noncoincident edges with multiplicities $\nu_1, \cdots, \nu_\ell$, subject to the constraint $\nu_1 + \cdots + \nu_\ell = 4k$. Then, the term corresponding to G is bounded by

$$n^{-4-2k} \prod_{j=1}^{\ell} (\eta_n \sqrt{n})^{\nu_j - 2} = \eta_n^{4k-2\ell} n^{-4-\ell}.$$

Since the graph of noncoincident edges of G can have at most two pieces of connected subgraphs, the number of noncoincident vertices of G is not greater than $\ell + 2$. If $\ell = 2k$, then $\nu_1 = \cdots = \nu_\ell = 2$. Therefore, there is at least one noncoincident edge consisting of edges from two different subgraphs and hence there must be a cycle in the graph of noncoincident edges of G. Therefore,

$$\begin{aligned} &\mathrm{E}(\beta_k(\mathbf{W}_n) - \mathrm{E}(\beta_k(\mathbf{W}_n)))^4 \\ &\le C_k n^{-2k-4} \left[\sum_{\ell < 2k} n^{\ell+2} (\eta_n^2 n)^{2k-\ell} + n^{2k+1} \right] \le C_k \eta_n n^{-2}, \end{aligned}$$

which is summable, and thus (2) is proved. Consequently, the proof of Theorem 2.9 is complete.

2.3 Semicircular Law by the Stieltjes Transform

As an illustration of the use of Stieltjes transforms, in this section we shall present a proof of Theorem 2.9 using them.

2.3.1 Stieltjes Transform of the Semicircular Law

Let $z = u + iv$ with $v > 0$ and $s(z)$ be the Stieltjes transform of the semicircular law. Then, we have

$$s(z) = \frac{1}{2\pi\sigma^2} \int_{-2\sigma}^{2\sigma} \frac{1}{x - z} \sqrt{4\sigma^2 - x^2}\, dx.$$

Letting $x = 2\sigma \cos y$, then

$$\begin{aligned} s(z) &= \frac{2}{\pi}\int_0^{\pi} \frac{1}{2\sigma\cos y - z}\sin^2 y dy \\ &= \frac{1}{\pi}\int_0^{2\pi} \frac{1}{2\sigma\frac{e^{iy}+e^{-iy}}{2} - z}\Big(\frac{e^{iy}-e^{-iy}}{2i}\Big)^2 dy \\ &= -\frac{1}{4i\pi}\oint_{|\zeta|=1} \frac{1}{\sigma(\zeta+\zeta^{-1})-z}(\zeta-\zeta^{-1})^2\zeta^{-1}d\zeta \text{ (setting } \zeta = e^{iy}) \\ &= -\frac{1}{4i\pi}\oint_{|\zeta|=1} \frac{(\zeta^2-1)^2}{\zeta^2(\sigma\zeta^2+\sigma-z\zeta)}d\zeta. \end{aligned} \tag{2.3.1}$$

We will use the residue theorem to evaluate the integral. Note that the integrand has three poles, at $\zeta_0 = 0$, $\zeta_1 = \frac{z+\sqrt{z^2-4\sigma^2}}{2\sigma}$, and $\zeta_2 = \frac{z-\sqrt{z^2-4\sigma^2}}{2\sigma}$, where here, and throughout the book, the square root of a complex number is specified as the one with the positive imaginary part. By this convention, we have

$$\sqrt{z} = \text{sign}(\Im z)\frac{|z|+z}{\sqrt{2(|z|+\Re z)}} \tag{2.3.2}$$

or

$$\Re(\sqrt{z}) = \frac{1}{\sqrt{2}}\text{sign}(\Im z)\sqrt{|z|+\Re z} = \frac{\Im z}{\sqrt{2(|z|-\Re z)}}$$

and

$$\Im(\sqrt{z}) = \frac{1}{\sqrt{2}}\sqrt{|z|-\Re z} = \frac{|\Im z|}{\sqrt{2(|z|+\Re z)}}.$$

This shows that the real part of $\sqrt{z}$ has the same sign as the imaginary part of z. Applying this to ζ_1 and ζ_2, we find that the real part of $\sqrt{z^2-4\sigma^2}$ has the same sign as z, which implies that $|\zeta_1| > |\zeta_2|$. Since $\zeta_1\zeta_2 = 1$, we conclude that $|\zeta_2| < 1$ and thus the two poles 0 and ζ_1 of the integrand are in the disk $|z| \le 1$. By simple calculation, we find that the residues at these two poles are

$$\frac{z}{\sigma^2} \text{ and } \frac{(\zeta_2^2-1)^2}{\sigma\zeta_2^2(\zeta_2-\zeta_1)} = \sigma^{-1}(\zeta_2-\zeta_1) = -\sigma^{-2}\sqrt{z^2-4\sigma^2}.$$

Substituting these into the integral of (2.3.1), we obtain the following lemma.

Lemma 2.11. *The Stieltjes transform for the semicircular law with scale parameter σ^2 is*

$$s(z) = -\frac{1}{2\sigma^2}(z - \sqrt{z^2-4\sigma^2}).$$

2.3.2 Proof of Theorem 2.9

At first, we truncate the underlying variables at $\eta_n\sqrt{n}$ and remove the diagonal elements and then centralize and rescale the off-diagonal elements as done in Steps 1–4 in the last section. That is, we assume that:

(i) For $i \neq j$, $|x_{ij}| \leq \eta_n\sqrt{n}$ and $x_{ii} = 0$.
(ii) For all $i \neq j$, $\mathrm{E}x_{ij} = 0$, $\mathrm{E}|x_{ij}|^2 = \sigma^2$.
(iii) The variables $\{x_{ij}, i < j\}$ are independent.

For brevity, we assume $\sigma^2 = 1$ in what follows.

By definition, the Stieltjes transform of $F^{\mathbf{W}_n}$ is given by

$$s_n(z) = \frac{1}{n}\mathrm{tr}(\mathbf{W}_n - z\mathbf{I}_n)^{-1}. \tag{2.3.3}$$

We shall then proceed in our proof by taking the following three steps:

(i) For any fixed $z \in \mathbb{C}^+ = \{z, \Im(z) > 0\}$, $s_n(z) - \mathrm{E}s_n(z) \to 0$, a.s.
(ii) For any fixed $z \in \mathbb{C}^+$, $\mathrm{E}s_n(z) \to s(z)$, the Stieltjes transform of the semicircular law.
(iii) Outside a null set, $s_n(z) \to s(z)$ for every $z \in \mathbb{C}^+$.

Then, applying Theorem B.9, it follows that, except for this null set, $F^{\mathbf{W}_n} \to F$ weakly.

Step 1. Almost sure convergence of the random part
For the first step, we show that, for each fixed $z \in \mathbb{C}^+$,

$$s_n(z) - \mathrm{E}(s_n(z)) \to 0 \quad \text{a.s.} \tag{2.3.4}$$

We need the extended Burkholder inequality.

Lemma 2.12. *Let $\{X_k\}$ be a complex martingale difference sequence with respect to the increasing σ-field $\{\mathcal{F}_k\}$. Then, for $p > 1$,*

$$\mathrm{E}\left|\sum X_k\right|^p \leq K_p\mathrm{E}\left(\sum |X_k|^2\right)^{p/2}.$$

Proof. Burkholder [67] proved the lemma for a real martingale difference sequence. Now, both $\{\Re X_k\}$ and $\{\Im X_k\}$ are martingale difference sequences. Thus, we have

$$\begin{aligned}
\mathrm{E}\left|\sum X_k\right|^p &\leq C_p\left[\mathrm{E}\left|\sum \Re X_k\right|^p + \mathrm{E}\left|\sum \Im X_k\right|^p\right] \\
&\leq C_p\left[K_p\mathrm{E}\left(\sum |\Re X_k|^2\right)^{p/2} + K_p\mathrm{E}\left(\sum |\Im X_k|^2\right)^{p/2}\right] \\
&\leq 2C_pK_p\mathrm{E}\left(\sum |X_k|^2\right)^{p/2},
\end{aligned}$$

where $C_p = 2^{p-1}$. This lemma is proved.

For later use, we introduce here another inequality proved in [67].

Lemma 2.13. *Let $\{X_k\}$ be a complex martingale difference sequence with respect to the increasing σ-field $\mathcal{F}_k$, and let E_k denote conditional expectation w.r.t. $\mathcal{F}_k$. Then, for $p \geq 2$,*

$$\mathrm{E}\left|\sum X_k\right|^p \leq K_p \left(\mathrm{E}\left(\sum \mathrm{E}_{k-1}|X_k|^2\right)^{p/2} + \mathrm{E}\sum |X_k|^p\right).$$

Similar to Lemma 2.12, Burkholder proved this lemma for the real case. Using the same technique as in the proof of Lemma 2.12, one may easily extend the Burkholder inequality to the complex case.

Now, we proceed to the proof of the almost sure convergence (2.3.4). Denote by $\mathrm{E}_k(\cdot)$ conditional expectation with respect to the σ-field generated by the random variables $\{x_{ij}, i, j > k\}$, with the convention that $\mathrm{E}_n s_n(z) = \mathrm{E}s_n(z)$ and $\mathrm{E}_0 s_n(z) = s_n(z)$. Then, we have

$$s_n(z) - \mathrm{E}(s_n(z)) = \sum_{k=1}^{n} [\mathrm{E}_{k-1}(s_n(z)) - \mathrm{E}_k(s_n(z))] := \sum_{k=1}^{n} \gamma_k,$$

where, by Theorem A.5,

$$\begin{aligned}
\gamma_k &= \frac{1}{n}\left(\mathrm{E}_{k-1}\mathrm{tr}(\mathbf{W}_n - z\mathbf{I})^{-1} - \mathrm{E}_k\mathrm{tr}(\mathbf{W}_n - z\mathbf{I})^{-1}\right) \\
&= \frac{1}{n}\left(\mathrm{E}_{k-1}[\mathrm{tr}(\mathbf{W}_n - z\mathbf{I})^{-1} - \mathrm{tr}(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}]\right. \\
&\quad \left. -\mathrm{E}_k[\mathrm{tr}(\mathbf{W}_n - z\mathbf{I})^{-1} - \mathrm{tr}(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}]\right) \\
&= \frac{1}{n}\left(\mathrm{E}_{k-1}\frac{1 + \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-2}\boldsymbol{\alpha}_k}{-z - \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k}\right. \\
&\quad \left. -\mathrm{E}_k\frac{1 + \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-2}\boldsymbol{\alpha}_k}{-z - \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k}\right),
\end{aligned}$$

where $\mathbf{W}_k$ is the matrix obtained from $\mathbf{W}_n$ with the k-th row and column removed and $\boldsymbol{\alpha}_k$ is the k-th column of $\mathbf{W}_n$ with the k-th element removed.

Note that

$$\begin{aligned}
&|1 + \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-2}\boldsymbol{\alpha}_k| \\
&\leq 1 + \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}(\mathbf{W}_k - \bar{z}\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k \\
&= v^{-1}\Im(z + \boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k)
\end{aligned}$$

which implies that

$$|\gamma_k| \leq 2/nv.$$

Noting that $\{\gamma_k\}$ forms a martingale difference sequence, applying Lemma 2.12 for $p = 4$, we have

$$\begin{aligned}\mathrm{E}|s_n(z)-\mathrm{E}(s_n(z))|^4 &\le K_4\mathrm{E}\left(\sum_{k=1}^{n}|\gamma_k|^2\right)^2\\ &\le K_4\left(\sum_{k=1}^{n}\frac{2}{n^2v^2}\right)^2\\ &\le \frac{4K_4}{n^2v^4}.\end{aligned}$$

By the Borel-Cantelli lemma, we know that, for each fixed $z\in\mathbb{C}^+$,

$$s_n(z)-\mathrm{E}(s_n(z))\to 0,\ \text{a.s.}$$

Step 2. Convergence of the expected Stieltjes transform
By Theorem A.4, we have

$$\begin{aligned}s_n(z) &= \frac{1}{n}\mathrm{tr}(\mathbf{W}_n-z\mathbf{I}_n)^{-1}\\ &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{-z-\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k}.\end{aligned}\tag{2.3.5}$$

Write $\varepsilon_k=\mathrm{E}s_n(z)-\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k$. Then we have

$$\begin{aligned}\mathrm{E}s_n(z) &= \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\frac{1}{-z-\mathrm{E}s_n(z)+\varepsilon_k}\\ &= -\frac{1}{z+\mathrm{E}s_n(z)}+\delta_n,\end{aligned}\tag{2.3.6}$$

where

$$\delta_n=\frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\left(\frac{\varepsilon_k}{(z+\mathrm{E}s_n(z))(-z-\mathrm{E}s_n(z)+\varepsilon_k)}\right).$$

Solving equation (2.3.6), we obtain two solutions:

$$\frac{1}{2}(-z+\delta_n\pm\sqrt{(z+\delta_n)^2-4}).$$

We show that

$$\mathrm{E}s_n(z)=\frac{1}{2}(-z+\delta_n+\sqrt{(z+\delta_n)^2-4}).\tag{2.3.7}$$

When fixing $\Re z$ and letting $\Im z=v\to\infty$, we have $\mathrm{E}s_n(z)\to 0$, which implies that $\delta_n\to 0$. Consequently,

$$\Im\Big(\frac{1}{2}(-z+\delta_n-\sqrt{(z+\delta_n)^2-4})\Big)\le-\frac{v-|\delta_n|}{2}\to-\infty,$$

which cannot be $\mathrm{E}s_n(z)$ since it violates the property that $\Im s_n(z)\ge 0$. Thus, assertion (2.3.7) is true when v is large. Now, we claim that assertion (2.3.7) is true for all $z\in\mathbb{C}^+$.

It is easy to see that $\mathrm{E}s_n(z)$ and $\frac{1}{2}(-z+\delta_n\pm\sqrt{(z+\delta_n)^2-4})$ are continuous functions on the upper half plane $\mathbb{C}^+$. If $\mathrm{E}s_n(z)$ takes a value on the branch $\frac{1}{2}(-z+\delta_n-\sqrt{(z+\delta_n)^2-4})$ for some z, then the two branches $\frac{1}{2}(-z+\delta_n\pm\sqrt{(z+\delta_n)^2-4})$ should cross each other at some point $z_0\in\mathbb{C}^+$. At this point, we would have $\sqrt{(z_0+\delta_n)^2-4}=0$ and hence $\mathrm{E}s_n(z_0)$ has to be one of the following:

$$\frac{1}{2}(-z_0+\delta_n)=\frac{1}{2}(-2z_0\pm 2).$$

However, both of the two values above have negative imaginary parts. This contradiction leads to the truth of (2.3.7).

From (2.3.7), to prove $\mathrm{E}s_n(z)\to s(z)$, it suffices to show that

$$\delta_n\to 0. \tag{2.3.8}$$

Now, rewrite

$$\begin{aligned}\delta_n&=-\frac{1}{n}\sum_{k=1}^n\frac{\mathrm{E}(\varepsilon_k)}{(z+\mathrm{E}s_n(z))^2}+\frac{1}{n}\sum_{k=1}^n\mathrm{E}\left(\frac{\varepsilon_k^2}{(z+\mathrm{E}s_n(z))^2(-z-\mathrm{E}s_n(z)+\varepsilon_k)}\right)\\&=J_1+J_2.\end{aligned}$$

By (A.1.10) and (A.1.12), we have

$$\begin{aligned}|\mathrm{E}\varepsilon_k|&=\left|\frac{1}{n}\mathrm{E}(\mathrm{tr}(\mathbf{W}_n-z\mathbf{I})^{-1}-\mathrm{tr}(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1})\right|\\&=\left|\frac{1}{n}\cdot\mathrm{E}\frac{1+\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-2}\boldsymbol{\alpha}_k}{-z-\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k}\right|\le\frac{1}{nv}.\end{aligned}$$

Note that

$$|z+\mathrm{E}s_n(z)|\ge\Im(z+\mathrm{E}s_n(z))=v+\mathrm{E}(\Im(s_n(z))\ge v.$$

Therefore, for any fixed $z\in\mathbb{C}^+$,

$$|J_1|\le\frac{1}{nv^3}\to 0.$$

On the other hand, we have

$$\begin{aligned}&|-z-\mathrm{E}s_n(z)+\varepsilon_k|=|-z-\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k|\\&\ge\Im(z+\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k)\end{aligned}$$

$$= v(1 + \boldsymbol{\alpha}_k^*((\mathbf{W}_k - z\mathbf{I}_{n-1})(\mathbf{W}_k - \bar{z}\mathbf{I}_{n-1}))^{-1}\boldsymbol{\alpha}_k) \geq v.$$

To prove $J_2 \to 0$, it is sufficient to show that

$$\max_k \mathrm{E}|\varepsilon_k|^2 \to 0.$$

Write $(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1} = (b_{ij})_{i,j\leq n-1}$. We then have

$$\begin{aligned}
\mathrm{E}|\varepsilon_k - \mathrm{E}\varepsilon_k|^2 &= \mathrm{E}|\boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k - \frac{1}{n}\mathrm{Etr}((\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1})|^2 \\
&= \mathrm{E}|\boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k - \frac{1}{n}\mathrm{tr}((\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1})|^2 \\
&\quad + \mathrm{E}\left|\frac{1}{n}\mathrm{tr}((\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}) - \frac{1}{n}\mathrm{Etr}((\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1})\right|^2.
\end{aligned}$$

By elementary calculations, we have

$$\begin{aligned}
&\mathrm{E}|\boldsymbol{\alpha}_k^*(\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k - \frac{1}{n}\mathrm{tr}((\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1})|^2 \\
&= \frac{1}{n^2}\left[\sum_{ij\neq k}[\mathrm{E}|b_{ij}|^2\mathrm{E}|x_{ik}|^2\mathrm{E}|x_{jk}|^2 + \mathrm{E}b_{ij}^2\mathrm{E}x_{ik}^2\mathrm{E}x_{jk}^2] + \sum_{i\neq k}\mathrm{E}|b_{ii}^2|(\mathrm{E}|x_{ik}^4| - 1)\right] \\
&\leq \frac{2}{n^2}\sum_{ij}\mathrm{E}|b_{ij}|^2 + \frac{\eta_n^2}{n}\sum_{i\neq k}\mathrm{E}|b_{ii}|^2 \\
&= \frac{2}{n^2}\mathrm{Etr}((\mathbf{W}_k - z\mathbf{I}_{n-1})(\mathbf{W}_k - \bar{z}\mathbf{I}_{n-1}))^{-1} + \frac{\eta_n^2}{n}\sum_{i\neq k}\mathrm{E}|b_{ii}|^2 \\
&\leq \frac{2}{nv^2} + \eta_n^2 \to 0.
\end{aligned} \tag{2.3.9}$$

By Theorem A.5, one can prove that

$$\mathrm{E}\left|\frac{1}{n}\mathrm{tr}((\mathbf{W}_n - z\mathbf{I}_{n-1})^{-1}) - \frac{1}{n}\mathrm{Etr}((\mathbf{W}_n - z\mathbf{I}_{n-1})^{-1})\right|^2 \leq 1/n^2v^2.$$

Then, the assertion $J_2 \to 0$ follows from the estimates above and the fact that

$$\mathrm{E}|\varepsilon_n|^2 = \mathrm{E}|\varepsilon_n - \mathrm{E}\varepsilon_n|^2 + |\mathrm{E}\varepsilon_n|^2.$$

The proof of the mean convergence is complete.

Step 3. Completion of the proof of Theorem 2.9
In this step, we need Vitali's convergence theorem.

Lemma 2.14. *Let $f_1, f_2, \cdots$ be analytic in D, a connected open set of $\mathbb{C}$, satisfying $|f_n(z)| \leq M$ for every n and z in D, and $f_n(z)$ converges as $n \to \infty$ for each z in a subset of D having a limit point in D. Then there exists a*

function f analytic in D for which $f_n(z) \to f(z)$ and $f_n'(z) \to f'(z)$ for all $z \in D$. Moreover, on any set bounded by a contour interior to D, the convergence is uniform and $\{f_n'(z)\}$ is uniformly bounded.

Proof. The conclusions on $\{f_n\}$ are from Vitali's convergence theorem (see Titchmarsh [275], p. 168). Those on $\{f_n'\}$ follow from the dominated convergence theorem (d.c.t.) and the identity

$$f_n'(z) = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{f_n(w)}{(w-z)^2} dw,$$

where $\mathcal{C}$ is a contour in D and enclosing z. The proof of the lemma is complete.

By Steps 1 and 2, for any fixed $z \in \mathbb{C}^+$, we have

$$s_n(z) \to s(z), \quad \text{a.s.},$$

where $s(z)$ is the Stieltjes transform of the standard semicircular law. That is, for each $z \in \mathbb{C}^+$, there exists a null set N_z (i.e., $P(N_z) = 0$) such that

$$s_n(z, \omega) \to s(z) \text{ for all } \omega \in N_z^c.$$

Now, let $\mathbb{C}_0^+ = \{z_m\}$ be a dense subset of $\mathbb{C}^+$ (e.g., all z of rational real and imaginary parts) and let $N = \cup N_{z_m}$. Then

$$s_n(z, \omega) \to s(z) \text{ for all } \omega \in N^c \text{ and } z \in \mathbb{C}_0^+.$$

Let $\mathbb{C}_m^+ = \{z \in \mathbb{C}^+, \Im z > 1/m, \ |z| \le m\}$. When $z \in \mathbb{C}_m^+$, we have $|s_n(z)| \le m$. Applying Lemma 2.14, we have

$$s_n(z, \omega) \to s(z) \text{ for all } \omega \in N^c \text{ and } z \in \mathbb{C}_m^+.$$

Since the convergence above holds for every m, we conclude that

$$s_n(z, \omega) \to s(z) \text{ for all } \omega \in N^c \text{ and } z \in \mathbb{C}^+.$$

Applying Theorem B.9, we conclude that

$$F^{\mathbf{W}_n} \xrightarrow{w} F, \text{ a.s.}$$

Chapter 3
Sample Covariance Matrices and the Marčenko-Pastur Law

The sample covariance matrix is a most important random matrix in multivariate statistical inference. It is fundamental in hypothesis testing, principal component analysis, factor analysis, and discrimination analysis. Many test statistics are defined by its eigenvalues.

The definition of a sample covariance matrix is as follows. Suppose that $\{x_{jk},\ j,k=1,2,\cdots\}$ is a double array of iid complex random variables with mean zero and variance σ^2. Write $\mathbf{x}_j=(x_{1j},\cdots,x_{pj})'$ and $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_n)$. The sample covariance matrix is defined by

$$\mathbf{S}=\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})^*,$$

where $\bar{\mathbf{x}}=\frac{1}{n}\sum\mathbf{x}_j$.

However, in most cases of spectral analysis of large dimensional random matrices, the sample covariance matrix is simply defined as

$$\mathbf{S}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^*=\frac{1}{n}\mathbf{X}\mathbf{X}^* \tag{3.0.1}$$

because the $\bar{\mathbf{x}}\bar{\mathbf{x}}^*$ is a rank 1 matrix and hence the removal of $\bar{\mathbf{x}}$ does not affect the LSD due to Theorem A.44.

In spectral analysis of large dimensional sample covariance matrices, it is usual to assume that the dimension p tends to infinity proportionally to the degrees of freedom n, namely $p/n\to y\in(0,\infty)$.

The first success in finding the limiting spectral distribution of the large sample covariance matrix $\mathbf{S}_n$ (named the Marčenko-Pastur (M-P) law by some authors) was due to Marčenko and Pastur [201]. Succeeding work was done in Bai and Yin [37], Grenander and Silverstein [137], Jonsson [169], Silverstein [256], Wachter [291], and Yin [300]. When the entries of $\mathbf{X}$ are not independent, Yin and Krishnaiah [303] investigated the limiting spectral distribution of $\mathbf{S}$ when the underlying distribution is isotropic. The theorem

Z. Bai and J.W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices, Second Edition*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0661-8_3,

in the next section is a consequence of a result in Yin [300], where the real case is considered.

3.1 M-P Law for the iid Case

3.1.1 Moments of the M-P Law

The M-P law $F_y(x)$ has a density function

$$p_y(x) = \begin{cases} \frac{1}{2\pi xy\sigma^2}\sqrt{(b-x)(x-a)}, & \text{if } \ a \le x \le b, \\ 0, & \text{otherwise,} \end{cases} \tag{3.1.1}$$

and has a point mass $1 - 1/y$ at the origin if $y > 1$, where $a = \sigma^2(1 - \sqrt{y})^2$ and $b = \sigma^2(1 + \sqrt{y})^2$. Here, the constant y is the dimension to sample size ratio index and σ^2 is the scale parameter. If $\sigma^2 = 1$, the M-P law is said to be the standard M-P law.

The moments $\beta_k = \beta_k(y, \sigma^2) = \int_a^b x^k p_y(x)dx$. In the following, we shall determine the explicit expression of β_k. Note that, for all $k \ge 1$,

$$\beta_k(y, \sigma^2) = \sigma^{2k}\beta_k(y, 1).$$

We need only compute β_k for the standard M-P law.

Lemma 3.1. *We have*

$$\beta_k = \sum_{r=0}^{k-1} \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r} y^r.$$

Proof. By definition,

$$\begin{aligned}
\beta_k &= \frac{1}{2\pi y}\int_a^b x^{k-1}\sqrt{(b-x)(x-a)}dx \\
&= \frac{1}{2\pi y}\int_{-2\sqrt{y}}^{2\sqrt{y}} (1+y+z)^{k-1}\sqrt{4y-z^2}dz \quad (\text{with } \ x = 1+y+z) \\
&= \frac{1}{2\pi y}\sum_{\ell=0}^{k-1}\binom{k-1}{\ell}(1+y)^{k-1-\ell}\int_{-2\sqrt{y}}^{2\sqrt{y}} z^\ell\sqrt{4y-z^2}dz \\
&= \frac{1}{2\pi y}\sum_{\ell=0}^{[(k-1)/2]}\binom{k-1}{2\ell}(1+y)^{k-1-2\ell}(4y)^{\ell+1}\int_{-1}^{1} u^{2\ell}\sqrt{1-u^2}du, \\
&\qquad\qquad (\text{by setting } z = 2\sqrt{y}u)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\pi y} \sum_{\ell=0}^{[(k-1)/2]} \binom{k-1}{2\ell} (1+y)^{k-1-2\ell} (4y)^{\ell+1} \int_0^1 w^{\ell-1/2} \sqrt{1-w} dw \\
&\qquad \text{(setting } u = \sqrt{w}\text{)} \\
&= \frac{1}{2\pi y} \sum_{\ell=0}^{[(k-1)/2]} \binom{k-1}{2\ell} (1+y)^{k-1-2\ell} (4y)^{\ell+1} \int_0^1 w^{\ell-1/2} \sqrt{1-w} dw \\
&= \sum_{\ell=0}^{[(k-1)/2]} \frac{(k-1)!}{\ell!(\ell+1)!(k-1-2\ell)!} y^{\ell} (1+y)^{k-1-2\ell} \\
&= \sum_{\ell=0}^{[(k-1)/2]} \sum_{s=0}^{k-1-2\ell} \frac{(k-1)!}{\ell!(\ell+1)!s!(k-1-2\ell-s)!} y^{\ell+s} \\
&= \sum_{\ell=0}^{[(k-1)/2]} \sum_{r=\ell}^{k-1-\ell} \frac{(k-1)!}{\ell!(\ell+1)!(r-\ell)!(k-1-r-\ell)!} y^{r} \\
&= \frac{1}{k} \sum_{r=0}^{k-1} \binom{k}{r} y^r \sum_{\ell=0}^{\min(r,k-1-r)} \binom{s}{\ell} \binom{k-r}{k-r-\ell-1} \\
&= \frac{1}{k} \sum_{r=0}^{k-1} \binom{k}{r} \binom{k}{r+1} y^r = \sum_{r=0}^{k-1} \frac{1}{r+1} \binom{k}{r} \binom{k-1}{r} y^r.
\end{aligned}
$$

By definition, we have $\beta_{2k} \le b^{2k} = (1+\sqrt{y})^{4k}$. From this, it is easy to see that the Carleman condition is satisfied.

3.1.2 Some Lemmas on Graph Theory and Combinatorics

To use the moment method to show the convergence of the ESD of large dimensional sample covariance matrices to the M-P law, we need to define a class of Δ-graphs and establish some lemmas concerning some counting problems related to Δ-graphs.

Suppose that $i_1, \cdots, i_k$ are k positive integers (not necessarily distinct) not greater than p and $j_1, \cdots, j_k$ are k positive integers (not necessarily distinct) not larger than n. A Δ-graph is defined as follows. *Draw two parallel lines, referring to the I line and the J line. Plot* $i_1, \cdots, i_k$ *on the I line and* $j_1, \cdots, j_k$ *on the J line, and draw* k *(down) edges from* i_u *to* j_u, $u = 1, \cdots, k$ *and* k *(up) edges from* j_u *to* i_{u+1}, $u = 1, \cdots, k$ *(with the convention that* $i_{k+1} = i_1$*).* The graph is denoted by $G(\mathbf{i}, \mathbf{j})$, where $\mathbf{i} = (i_1, \cdots, i_k)$ and $\mathbf{j} = (j_1, \cdots, j_k)$. An example of a Δ-graph is shown in Fig. 3.1.

Two graphs are said to be isomorphic if one becomes the other by a suitable permutation on $(1, 2, \cdots, p)$ and a suitable permutation on $(1, 2, \cdots, n)$.

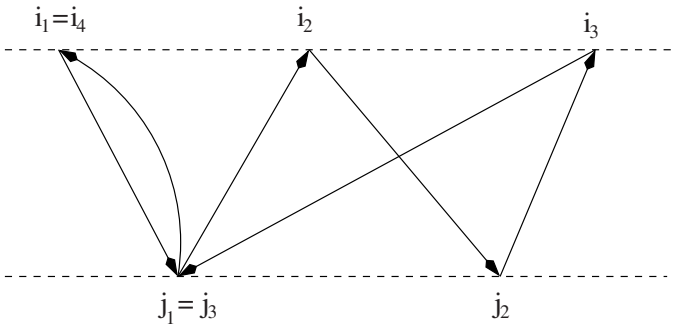

Fig. 3.1 A Δ-graph.

For each isomorphism class, there is only one graph, called *canonical*, satisfying $i_1 = j_1 = 1$, $i_u \leq \max\{i_1, \cdots, i_{u-1}\} + 1$, and $j_u \leq \max\{j_1, \cdots, j_{u-1}\} + 1$. A canonical Δ-graph $G(\mathbf{i}, \mathbf{j})$ is denoted by $\Delta(k, r, s)$ if G has $r+1$ noncoincident I-vertices and s noncoincident J-vertices. A canonical $\Delta(k, r, s)$ can be directly defined in the following way:

1. Its vertex set $V = V_I + V_J$, where $V_I = \{1, \cdots, r+1\}$, called the **I-vertices**, and $V_J = \{1, \cdots, s\}$, called the **J-vertices**.
2. There are two functions, $f : \{1, \cdots, k\} \mapsto \{1, \cdots, r+1\}$ and $g : \{1, \cdots, k\} \mapsto \{1, \cdots, s\}$, satisfying

$$\begin{aligned} &f(1) = 1 = g(1) = f(k+1), \\ &f(i) \leq \max\{f(1), \cdots, f(i-1)\} + 1, \\ &g(j) \leq \max\{g(1), \cdots, g(j-1)\} + 1. \end{aligned}$$

3. Its edge set $E = \{e_{1d}, e_{1u}, \cdots, e_{kd}, e_{ku}\}$, where $e_{1d}, \cdots, e_{kd}$ are called the down edges and $e_{1u}, \cdots, e_{ku}$ are called the up edges.
4. $F(e_{jd}) = (f(j), g(j))$ and $F(e_{ju}) = (g(j), f(j+1))$ for $j = 1, \cdots, k$.

In the case where $f(j+1) = \max\{f(1), \cdots, f(j)\} + 1$, the edge $e_{j,u}$ is called an up innovation, and in the case where $g(j) = \max\{g(1), \cdots, g(j-1)\} + 1$, the edge $e_{j,d}$ is called a down innovation. Intuitively, an up innovation leads to a new I-vertex and a down innovation leads to a new J-vertex. We make the convention that the first down edge is a down innovation and the last up edge is not an innovation.

Similar to the Γ-graphs, we classify $\Delta(k, r, s)$-graphs into three categories:

Category 1 (denoted by $\Delta_1(k, r)$): Δ-graphs in which each down edge must coincide with one and only one up edge. If we glue the coincident edges, the resulting graph is a tree of k edges. In this category, $r + s = k$ and thus s is suppressed for simplicity.
Category 2 ($\Delta_2(k, r, s)$): Δ-graphs that contain at least one single edge.

Category 3 ($\Delta_3(k,r,s)$): Δ-graphs that do not belong to $\Delta_1(k,r)$ or $\Delta_2(k,r,s)$.

Similar to the arguments given in Subsection 2.1.2, the number of graphs in each isomorphism class for a given canonical $\Delta(k,r,s)$ is given by the following lemma.

Lemma 3.2. *For a given k, r, and s, the number of graphs in the isomorphism class for each canonical $\Delta(k,r,s)$-graph is*

$$p(p-1)\cdots(p-r)n(n-1)\cdots(n-s+1) = p^{r+1}n^s[1+O(n^{-1})].$$

For a Δ_3-graph, we have the following lemma.

Lemma 3.3. *The total number of noncoincident vertices of a $\Delta_3(k,r,s)$-graph is less than or equal to k.*

Proof. Let G be a graph of $\Delta_3(k,r,s)$. Note that any Δ-graph is connected. Since G is not in category 2, it does not contain single edges and hence the number of noncoincident edges is not larger than k. If the number of noncoincident edges is less than k, then the lemma is proved. If the number of noncoincident edges is exactly k, the graph of noncoincident edges must contain a cycle since it is not in category 1. In this case, the number of noncoincident vertices is also not larger than k and the lemma is proved.

A more difficult task is to count the number of $\Delta_1(k,r)$-graphs, as given in the following lemma.

Lemma 3.4. *For k and r, the number of $\Delta_1(k,r)$-graphs is*

$$\frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}.$$

Proof. Define two characteristic sequences $\{u_1,\cdots,u_k\}$ and $\{d_1,\cdots,d_k\}$ of the graph G by

$$u_\ell = \begin{cases} 1, & \text{if } \ f(\ell+1) = \max\{f(1),\cdots,f(\ell)\}+1, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$d_\ell = \begin{cases} -1, & \text{if } \ f(\ell) \notin \{1, f(\ell+1),\cdots,f(k)\}, \\ 0, & \text{otherwise.} \end{cases}$$

We can interpret the intuitive meaning of the characteristic sequences as follows: $u_\ell = 1$ if and only if the ℓ-th up edge is an up innovation and $d_\ell = -1$ if and only if the ℓ-th down edge coincides with the up innovation that leads to this I-vertex. An example with $r=2$ and $s=3$ is given in Fig. 3.2.

By definition, we always have $u_k = 0$, and since $f(1) = 1$, we always have $d_1 = 0$. For a $\Delta_1(k,r)$-graph, there are exactly r up innovations and hence there are r u-variables equal to 1. Since there are r I-vertices other than 1, there are then r d-variables equal to -1.

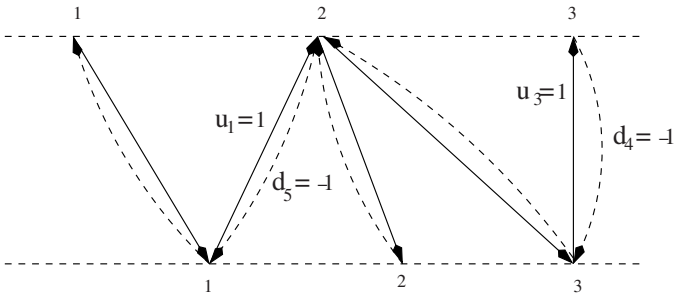

Fig. 3.2 Definition of (u, d) sequence

From its definition, one sees that $d_\ell = -1$ means that after plotting the ℓ-th down edge $(f(\ell), g(\ell))$, the future path will never revisit the I-vertex $f(\ell)$. This means that the edge $(f(\ell), g(\ell))$ must coincide with the up innovation leading to the vertex $f(\ell)$. Since there are $s = k - r$ down innovations to lead out the s J-vertices, $d_\ell = 0$ therefore implies that the edge $(f(\ell), g(\ell))$ must be a down innovation.

From the argument above, one sees that $d_\ell = -1$ must follow a $u_j = 1$ for some $j < \ell$. Therefore, the two sequences should satisfy the restriction

$$u_1 + \cdots + u_{\ell-1} + d_2 + \cdots + d_\ell \geq 0, \quad \ell = 2, \cdots, k. \tag{3.1.2}$$

From the definition of the characteristic sequences, each $\Delta_1(k, r)$-graph defines a pair of characteristic sequences. Conversely, we shall show that each pair of characteristic sequences satisfying (3.1.2) uniquely defines a $\Delta_1(k, r)$-graph. In other words, the functions f and g in the definition of the Δ-graph G are uniquely determined by the two sequences of $\{u_\ell\}$ and $\{d_\ell\}$.

At first, we notice that $u_\ell = 1$ implies that $e_{\ell,u}$ is an up innovation and thus

$$f(\ell + 1) = 1 + \ ^{\#}\{j \leq \ell, u_j = 1\}.$$

Similarly, $d_\ell = 0$ implies that $e_{\ell,d}$ is a down innovation and thus

$$g(\ell) = \ ^{\#}\{j \leq \ell, d_j = 0\}.$$

However, it is not easy to define the values of f and g at other points. So, we will directly plot the $\Delta_1(k, r)$-graph from the two characteristic sequences.

Since $d_1 = 0$ and hence $e_{1,d}$ is a down innovation, we draw $e_{1,d}$ from the I-vertex 1 to the J-vertex 1. If $u_1 = 0$, then $e_{1,u}$ is not an up innovation and thus the path must return the I-vertex 1 from the J-vertex 1; i.e., $f(2) = 1$. If $u_1 = 1$, $e_{1,u}$ is an up innovation leading to the new I-vertex 2; that is, $f(2) = 2$. Thus, the edge $e_{1,u}$ is from the J-vertex 1 to the I-vertex 2. This shows that the first pair of down and up edges are uniquely determined by u_1 and d_1. Suppose that the first ℓ pairs of the down and up edges are uniquely

determined by the sequences $\{u_1, \cdots, u_\ell\}$ and $\{d_1, \cdots, d_\ell\}$. Also, suppose that the subgraph G_ℓ of the first ℓ pairs of down and up edges satisfies the following properties

1. G_ℓ is connected, and the undirectional noncoincident edges of G_ℓ form a tree.
2. If the end vertex $f(\ell+1)$ of $e_{\ell,u}$ is the I-vertex 1, then each down edge of G_ℓ coincides with an up edge of G_ℓ. Thus, G_ℓ does not have single innovations.
   If the end vertex $f(\ell+1)$ of $e_{\ell,u}$ is not the I-vertex 1, then from the I-vertex 1 to the I-vertex $f(\ell+1)$ there is only one path (chain without cycles) of down-up-down-up single innovations and all other down edges coincide with an up edge.

To draw the $\ell+1$-st pair of down and up edges, we consider the following four cases.

Case 1. $d_{\ell+1} = 0$ and $u_{\ell+1} = 1$. Then both edges of the $\ell+1$-st pair are innovations. Thus, adding the two innovations to G_ℓ, the resulting subgraph $G_{\ell+1}$ satisfies the two properties above with the path of down-up single innovations that consists of the original path of single innovations and the two new innovations. See Case 1 in Fig. 3.3.

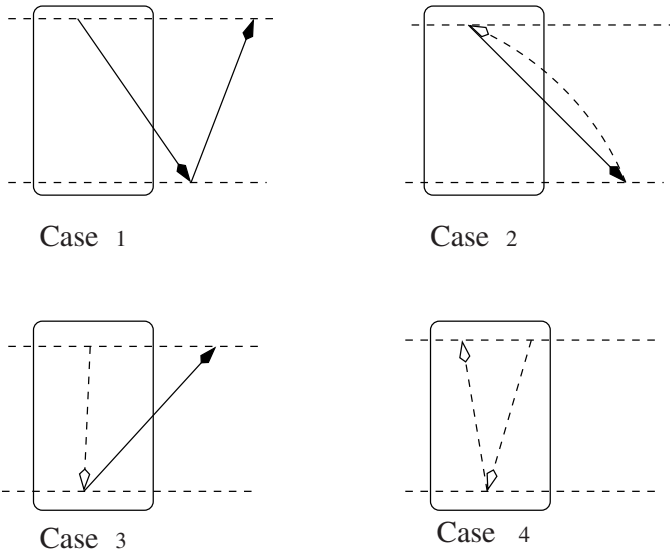

Fig. 3.3 Examples of the four cases. In the four graphs, the rectangle denotes the subgraph G_ℓ, solid arrows are new innovations, and broken arrows are new T_3 edges.

Case 2. $d_{\ell+1} = 0$ and $u_{\ell+1} = 0$. Then, $e_{\ell+1,d}$ is a down innovation and $e_{\ell+1,u}$ coincides with $e_{\ell+1,d}$. See Case 2 in Fig. 3.3. Thus, for the subgraph $G_{\ell+1}$,

the two properties above can be trivially seen from the hypothesis for the subgraph G_ℓ. The single innovation chain of $G_{\ell+1}$ is exactly the same as that of G_ℓ.

Case 3. $d_{\ell+1} = -1$ and $u_{\ell+1} = 1$. In this case, by (3.1.2) we have

$$u_1 + \cdots + u_\ell + d_2 + \cdots + d_\ell \geq 1$$

which implies that the total number of I-vertices of G_ℓ other than 1 (i.e., $u_1 + \cdots + u_\ell$) is greater than the number of I-vertices of G_ℓ from which the graph ultimately leaves (i.e., $d_2 + \cdots + d_\ell$). Therefore, $f(\ell+1) \neq 1$ because G_ℓ must contain single innovations by property 2. Then there must be a single up innovation leading to the vertex $f(\ell+1)$ and thus we can draw the down edge $e_{\ell+1,d}$ coincident with this up innovation. Then, the next up innovation $e_{\ell,u}$ starts from the end vertex to $g(\ell+1)$. See case 3 in Fig. 3.3. It is easy to see that the two properties above hold with the path of single innovations that is the original one with the last up innovation replaced by $e_{\ell+1,u}$.

Case 4. $d_{\ell+1} = -1$ and $u_{\ell+1} = 0$. Then, as discussed in case 3, $e_{\ell+1,d}$ can be drawn to coincide with the only up innovation ended at $f(\ell+1)$. Prior to this up innovation, there must be a single down innovation with which the up edge $e_{\ell,u}$ can be drawn to coincide. If the path of single innovations of G_ℓ has only one pair of down-up innovations, then $f(\ell+2) = 1$ and hence $G_{\ell+1}$ has no single innovations. If the path of single innovations of G_ℓ has more than two edges, then the remaining part of the path of single innovations of G_ℓ, with the last two innovations removed, forms a path of single innovations of $G_{\ell+1}$. See case 1 in Fig. 3.3. In either case, two properties for $G_{\ell+1}$ hold.

By induction, it is shown that two sequences subject to restriction (3.1.2) uniquely determine a $\Delta_1(k,r)$-graph. Therefore, counting the number of $\Delta_1(k,r)$-graphs is equivalent to counting the number of pairs of characteristic sequences.

Now, we count the number of characteristic sequences for given k and r. We have the following lemma.

Lemma 3.5. *For a given k and r $(0 \leq r \leq k-1)$, the number of $\Delta_1(k,r)$-graphs is*

$$\frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}.$$

Proof. Ignoring the restriction (3.1.2), we have $\binom{k-1}{r}\binom{k-1}{r}$ ways to arrange r ones in the $k-1$ positions $u_1, \cdots, u_{k-1}$ and to arrange r minus ones in the $k-1$ positions $d_2, \cdots, d_k$. If there is an integer $2 \leq \ell \leq k$ such that

$$u_1 + \cdots + u_{\ell-1} + d_1 + \cdots + d_\ell = -1,$$

then define

$$\tilde{u}_j = \begin{cases} u_j, & \text{if } j < \ell, \\ -d_{j+1}, & \text{if } \ell \leq j < k, \end{cases}$$

and

$$\tilde{d}_j = \begin{cases} d_j, & \text{if } 1 < j \le \ell, \\ -u_{j-1}, & \text{if } \ell < j \le k. \end{cases}$$

Then we have $r-1$ u's equal to one and $r+1$ d's equal to minus one. There are $\binom{k-1}{r-1}\binom{k-1}{r+1}$ ways to arrange $r-1$ ones in the $k-1$ positions $\tilde{u}_1, \cdots, \tilde{u}_{k-1}$, and to arrange $r+1$ minus ones in the $k-1$ positions $\tilde{d}_2, \cdots, \tilde{d}_k$.

Therefore, the number of pairs of characteristic sequences with indices k and r satisfying the restriction (3.1.2) is

$$\binom{k-1}{r}^2 - \binom{k-1}{r-1}\binom{k-1}{r+1} = \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}.$$

The proof of the lemma is complete.

3.1.3 M-P Law for the iid Case

In this section, we consider the LSD of the sample covariance matrix for the case where the underlying variables are iid.

Theorem 3.6. *Suppose that $\{x_{ij}\}$ are iid real random variables with mean zero and variance σ^2. Also assume that $p/n \to y \in (0, \infty)$. Then, with probability one, $F^{\mathbf{S}}$ tends to the M-P law, which is defined in (3.1.1).*

Yin [300] considered existence of the LSD of the sequence of random matrices $\mathbf{S}_n\mathbf{T}_n$, where $\mathbf{T}_n$ is a positive definite random matrix and is independent of $\mathbf{S}_n$. When $\mathbf{T}_n = \mathbf{I}_p$, Yin's result reduces to Theorem 3.6.

In this section, we shall give a proof of the following extension to the complex random sample covariance matrix.

Theorem 3.7. *Suppose that $\{x_{ij}\}$ are iid complex random variables with variance σ^2. Also assume that $p/n \to y \in (0, \infty)$. Then, with probability one, $F^{\mathbf{S}}$ tends to a limiting distribution the same as described in Theorem 3.6.*

Remark 3.8. The proofs will be separated into several steps. Note that the M-P law varies with the scale parameter σ^2. Therefore, in the proof we shall assume that $\sigma^2 = 1$, without loss of generality.

In most work in multivariate statistics, it is assumed that the means of the entries of $\mathbf{X}_n$ are zero. The centralization technique, which is Theorem A.44, relies on the interlacing property of eigenvalues of two matrices that differ by a rank-one matrix. One then sees that removing the common mean of the entries of $\mathbf{X}_n$ does not alter the LSD of sample covariance matrices.

Step 1. Truncation, Centralization, and Rescaling
Let C be a positive number, and define

$$\begin{aligned}
\hat{x}_{ij} &= x_{ij} I(|x_{ij}| \le C), \\
\tilde{x}_{ij} &= \hat{x}_{ij} - \mathrm{E}(\hat{x}_{11}), \\
\widehat{\mathbf{x}}_i &= (\hat{x}_{i1}, \cdots, \hat{x}_{ip})', \\
\widetilde{\mathbf{x}}_i &= (\tilde{x}_{i1}, \cdots, \tilde{x}_{ip})', \\
\widehat{\mathbf{S}}_n &= \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i^* = \frac{1}{n} \widehat{\mathbf{X}} \widehat{\mathbf{X}}^*, \\
\widetilde{\mathbf{S}}_n &= \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^* = \frac{1}{n} \widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^*.
\end{aligned}$$

Write the ESDs of $\widehat{\mathbf{S}}_n$ and $\widetilde{\mathbf{S}}_n$ as $F^{\widehat{\mathbf{S}}_n}$ and $F^{\widetilde{\mathbf{S}}_n}$, respectively. By Corollary A.42 and the strong law of large numbers, we have

$$\begin{aligned}
L^4(F^{\mathbf{S}}, F^{\widehat{\mathbf{S}}_n}) &\le \left(\frac{2}{np} \sum_{i,j} (|x_{ij}^2| + |\hat{x}_{ij}^2|) \right) \left(\frac{1}{np} \sum_{i,j} (|x_{ij} - \hat{x}_{ij}|^2) \right) \\
&\le \left(\frac{4}{np} \sum_{i,j} |x_{ij}^2| \right) \left(\frac{1}{np} \sum_{i,j} (|x_{ij}^2| I(|x_{ij}| > C)) \right) \\
&\to 4\mathrm{E}(|x_{ij}^2| I(|x_{ij}| > C)), \text{ a.s.}
\end{aligned} \tag{3.1.3}$$

Note that the right-hand side of (3.1.3) can be made arbitrarily small by choosing C large enough.

Also, by Theorem A.44, we obtain

$$||F^{\widehat{\mathbf{S}}_n} - F^{\widetilde{\mathbf{S}}_n}|| \le \frac{1}{p} \mathrm{rank}(\mathrm{E}\widehat{\mathbf{X}}) = \frac{1}{p}. \tag{3.1.4}$$

Write $\tilde{\sigma}^2 = \mathrm{E}(|\tilde{x}_{jk}|^2) \to 1$, as $C \to \infty$. Applying Corollary A.42, we obtain

$$\begin{aligned}
L^4(F^{\widetilde{\mathbf{S}}_n}, F^{\tilde{\sigma}^{-2}\widetilde{\mathbf{S}}_n}) &\le 2 \left(\frac{1 + \tilde{\sigma}^2}{np\tilde{\sigma}^2} \sum_{i,j} |\tilde{x}_{ij}|^2 \right) \left(\frac{1 - \tilde{\sigma}^2}{np\tilde{\sigma}^2} \sum_{i,j} |\tilde{x}_{ij(c)}|^2 \right) \\
&\to 2(1 - \tilde{\sigma}^4), \text{ a.s.}
\end{aligned} \tag{3.1.5}$$

Note that the right-hand side of the inequality above can be made arbitrarily small by choosing C large. Combining (3.1.3), (3.1.4), and (3.1.5), in the proof of Theorem 3.7 we may assume that the variables x_{jk} are uniformly bounded with mean zero and variance 1. For abbreviation, in proofs given in the next step, we still use $\mathbf{S}_n$, $\mathbf{X}_n$ for the matrices associated with the truncated variables.

Step 2. Proof for the M-P Law by MCT

Now, we are able to employ the moment approach to prove Theorem 3.7. By elementary calculus, we have

$$\begin{aligned}\beta_k(\mathbf{S}_n) &= \int x^k F^{\mathbf{S}_n}(dx) \\ &= p^{-1}n^{-k} \sum_{\{i_1,\cdots,i_k\}} \sum_{\{j_1,\cdots,j_k\}} x_{i_1j_1}\bar{x}_{i_2j_1}x_{i_2j_2}\cdots x_{i_kj_k}\bar{x}_{i_1j_k} \\ &:= p^{-1}n^{-k} \sum_{\mathbf{i},\mathbf{j}} X_{G(\mathbf{i},\mathbf{j})},\end{aligned}$$

where the summation runs over all $G(\mathbf{i},\mathbf{j})$-graphs as defined in Subsection 3.1.2, the indices in $\mathbf{i} = (i_1,\cdots,i_k)$ run over $1,2,\cdots,p$, and the indices in $\mathbf{j} = (j_1,\cdots,j_k)$ run over $1,2,\cdots,n$.

To complete the proof of the almost sure convergence of the ESD of $\mathbf{S}_n$, we need only show the following two assertions:

$$\begin{aligned}\mathrm{E}(\beta_k(\mathbf{S}_n)) &= p^{-1}n^{-k} \sum_{\mathbf{i},\mathbf{j}} \mathrm{E}(x_{G(\mathbf{i},\mathbf{j})}) \\ &= \sum_{r=0}^{k-1} \frac{y_n^r}{r+1} \binom{k}{r} \binom{k-1}{r} + O(n^{-1}) \qquad (3.1.6)\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Var}(\beta_k(\mathbf{S}_n)) \\ &= p^{-2}n^{-2k} \sum_{\mathbf{i}_1,\mathbf{j}_1,\mathbf{i}_2,\mathbf{j}_2} [\mathrm{E}(x_{G_1(\mathbf{i}_1,\mathbf{j}_1)}x_{G_2(\mathbf{i}_2,\mathbf{j}_2)} - \mathrm{E}(x_{G_1(\mathbf{i}_1,\mathbf{j}_1)})\mathrm{E}(x_{G_2(\mathbf{i}_2,\mathbf{j}_2)}))] \\ &= O(n^{-2}), \qquad (3.1.7)\end{aligned}$$

where $y_n = p/n$, and the graphs G_1 and G_2 are defined by $(\mathbf{i}_1,\mathbf{j}_1)$ and $(\mathbf{i}_2,\mathbf{j}_2)$, respectively.

The proof of (3.1.6). On the left-hand side of (3.1.6), two terms are equal if their corresponding graphs are isomorphic. Therefore, by Lemma 3.2, we may rewrite

$$\mathrm{E}(\beta_k(\mathbf{S}_n)) = p^{-1}n^{-k} \sum_{\Delta(k,r,s)} p(p-1)\cdots(p-r)n(n-1)\cdots(n-s+1)\mathrm{E}(X_{\Delta(k,r,s)}), \qquad (3.1.8)$$

where the summation is taken over canonical $\Delta(k,r,s)$-graphs. Now, split the sum in (3.1.8) into three parts according to $\Delta_1(k,r)$ and $\Delta_j(k,r,s)$, $j = 2,3$. Since the graph in $\Delta_2(k,r,s)$ contains at least one single edge, the corresponding expectation is zero. That is,

$$S_2 = p^{-1}n^{-k} \sum_{\Delta_2(k,r,s)} p(p-1)\cdots(p-r)n(n-1)\cdots(n-s+1)\mathrm{E}(X_{\Delta_2(k,r,s)}) = 0.$$

By Lemma 3.3, for a graph of $\Delta_3(k,r,s)$, we have $r+s<k$. Since the variable $x_{\Delta(k,r,s)}$ is bounded by $(2C/\tilde{\sigma})^{2k}$, we conclude that

$$\begin{aligned} S_3 &= p^{-1}n^{-k} \sum_{\Delta_3(k,r,s)} p(p-1)\cdots(p-r)n(n-1)\cdots(n-s+1)\mathrm{E}(X_{\Delta(k,r,s)}) \\ &= O(n^{-1}). \end{aligned}$$

Now let us evaluate S_1. For a graph in $\Delta_1(k,r)$ (with $s=k-r$), each pair of coincident edges consists of a down edge and an up edge; say, the edge (i_a,j_a) must coincide with the edge (j_a,i_a). This pair of coincident edges corresponds to the expectation $\mathrm{E}(|X_{i_a,j_a}|^2)=1$. Therefore, $\mathrm{E}(X_{\Delta_1(k,r)})=1$. By Lemma 3.4,

$$\begin{aligned} S_1 &= p^{-1}n^{-k} \sum_{\Delta_1(k,r)} p(p-1)\cdots(p-r)n(n-1)\cdots(n-s+1)\mathrm{E}(X_{\Delta_1(k,r)}) \\ &= \sum_{r=0}^{k-1} \frac{y_n^r}{r+1}\binom{k}{r}\binom{k-1}{r} + O(n^{-1}) \\ &= \beta_k + o(1), \end{aligned}$$

where $y_n = p/n \to y \in (0,\infty)$. The proof of (3.1.6) is complete.

The proof of (3.1.7). Recall

$$\begin{aligned} &\mathrm{Var}(\beta_k(\mathbf{S}_n)) \\ &= p^{-2}n^{-2k} \sum_{\mathbf{i},\mathbf{j}} [\mathrm{E}(X_{G_1(\mathbf{i}_1,\mathbf{j}_1)}X_{G_2(\mathbf{i}_2,\mathbf{j}_2)}) - \mathrm{E}(X_{G_1(\mathbf{i}_1,\mathbf{j}_1)})\mathrm{E}(X_{G_2(\mathbf{i}_2,\mathbf{j}_2)})]. \end{aligned}$$

Similar to the proof of Theorem 2.5, if G_1 has no edges coincident with edges of G_2 or $G=G_1\cup G_2$ has an overall single edge, then

$$\mathrm{E}(X_{G_1(\mathbf{i}_1,\mathbf{j}_1)}X_{G_2(\mathbf{i}_2,\mathbf{j}_2)}) - \mathrm{E}(X_{G_1(\mathbf{i}_1,\mathbf{j}_1)})\mathrm{E}(X_{G_2(\mathbf{i}_2,\mathbf{j}_2)}) = 0$$

by independence between X_{G_1} and X_{G_2}.

Similar to the arguments in Subsection 2.1.3, one may show that the number of noncoincident vertices of G is not more than $2k$. By the fact that the terms are bounded, we conclude that assertion (3.1.7) holds and consequently conclude the proof of Theorem 3.7.

Remark 3.9. The existence of the second moment of the entries is obviously necessary and sufficient for the Marčenko-Pastur law since the limiting distribution involves the parameter σ^2.

3.2 Generalization to the Non-iid Case

Sometimes it is of practical interest to consider the case where the entries of $\mathbf{X}_n$ depend on n and for each n they are independent but not necessarily identically distributed. As in Section 2.2, we shall briefly present a proof of the following theorem.

Theorem 3.10. *Suppose that, for each n, the entries of $\mathbf{X}$ are independent complex variables with a common mean μ and variance σ^2. Assume that $p/n \to y \in (0, \infty)$ and that, for any $\eta > 0$,*

$$\frac{1}{\eta^2 np} \sum_{jk} \mathrm{E}(|x_{jk}^{(n)}|^2 I(|x_{jk}^{(n)}| \geq \eta\sqrt{n})) \to 0. \tag{3.2.1}$$

Then, with probability one, $F^{\mathbf{S}}$ tends to the Marčenko-Pastur law with ratio index y and scale index σ^2.

Proof. We shall only give an outline of the proof of this theorem. The details are left to the reader. Without loss of generality, we assume that $\mu = 0$ and $\sigma^2 = 1$. Similar to what we did in the proof of Theorem 2.9, we may select a sequence $\eta_n \downarrow 0$ such that condition (3.2.1) holds true when η is replaced by η_n. In the following, once condition (3.2.1) is used, we always mean this condition with η replaced by η_n.

Applying Theorem A.44 and the Bernstein inequality, by condition (3.2.1), we may truncate the variables $x_{ij}^{(n)}$ at $\eta_n\sqrt{n}$. Then, applying Corollary A.42, by condition (3.2.1), we may recentralize and rescale the truncated variables. Thus, in the rest of the proof, we shall drop the superscript (n) from the variables for brevity. We further assume that

$$\begin{aligned} &1)\ |x_{ij}| < \eta_n\sqrt{n}, \\ &2)\ \mathrm{E}(x_{ij}) = 0 \ \text{ and } \ \mathrm{Var}(x_{ij}) = 1. \end{aligned} \tag{3.2.2}$$

By arguments to those in the proof of Theorem 2.9, one can show the following two assertions:

$$\mathrm{E}(\beta_k(\mathbf{S}_n)) = \sum_{r=0}^{k-1} \frac{y_n^r}{r+1} \binom{k}{r} \binom{k-1}{r} + o(1) \tag{3.2.3}$$

and

$$\mathrm{E}\,|\beta_k(\mathbf{S}_n) - \mathrm{E}\,(\beta_k(\mathbf{S}_n))|^4 = o(n^{-2}). \tag{3.2.4}$$

The proof of Theorem 3.10 is then complete.

3.3 Proof of Theorem 3.10 by the Stieltjes Transform

As an illustration applying Stieltjes transforms to sample covariance matrices, we give a proof of Theorem 3.10 in this section. Using the same approach of truncating, centralizing, and rescaling as we did in the last section, we may assume the additional conditions given in (3.2.2).

3.3.1 Stieltjes Transform of the M-P Law

Let $z = u + iv$ with $v > 0$ and $s(z)$ be the Stieltjes transform of the M-P law.

Lemma 3.11.

$$s(z) = \frac{\sigma^2(1-y) - z + \sqrt{(z - \sigma^2 - y\sigma^2)^2 - 4y\sigma^4}}{2yz\sigma^2}. \tag{3.3.1}$$

Proof. When $y < 1$, we have

$$s(z) = \int_a^b \frac{1}{x - z} \frac{1}{2\pi x y \sigma^2} \sqrt{(b-x)(x-a)}dx,$$

where $a = \sigma^2(1 - \sqrt{y})^2$ and $b = \sigma^2(1 + \sqrt{y})^2$.

Letting $x = \sigma^2(1 + y + 2\sqrt{y}\cos w)$ and then setting $\zeta = e^{iw}$, we have

$$\begin{aligned}
s(z) &= \int_0^\pi \frac{2}{\pi} \frac{1}{(1 + y + 2\sqrt{y}\cos w)(\sigma^2(1 + y + 2\sqrt{y}\cos w) - z)} \sin^2 w dw \\
&= \frac{1}{\pi} \int_0^{2\pi} \frac{((e^{iw} - e^{-iw})/2i)^2}{(1 + y + \sqrt{y}(e^{iw} + e^{-iw}))(\sigma^2(1 + y + \sqrt{y}(e^{iw} + e^{-iw})) - z)} dw \\
&= -\frac{1}{4i\pi} \oint_{|\zeta|=1} \frac{(\zeta - \zeta^{-1})^2}{\zeta(1 + y + \sqrt{y}(\zeta + \zeta^{-1}))(\sigma^2(1 + y + \sqrt{y}(\zeta + \zeta^{-1})) - z)} d\zeta \\
&= -\frac{1}{4i\pi} \oint_{|\zeta|=1} \frac{(\zeta^2 - 1)^2}{\zeta((1+y)\zeta + \sqrt{y}(\zeta^2 + 1))(\sigma^2(1+y)\zeta + \sqrt{y}\sigma^2(\zeta^2 + 1) - z\zeta)} d\zeta.
\end{aligned} \tag{3.3.2}$$

The integrand function has five simple poles at

$$\begin{aligned}
\zeta_0 &= 0, \\
\zeta_1 &= \frac{-(1+y) + (1-y)}{2\sqrt{y}}, \\
\zeta_2 &= \frac{-(1+y) - (1-y)}{2\sqrt{y}},
\end{aligned}$$

$$\zeta_3 = \frac{-\sigma^2(1+y)+z+\sqrt{\sigma^4(1-y)^2-2\sigma^2(1+y)z+z^2}}{2\sigma^2\sqrt{y}},$$
$$\zeta_4 = \frac{-\sigma^2(1+y)+z-\sqrt{\sigma^4(1-y)^2-2\sigma^2(1+y)z+z^2}}{2\sigma^2\sqrt{y}}.$$

By elementary calculation, we find that the residues at these five poles are

$$\frac{1}{y\sigma^2}, \quad \mp\frac{1-y}{yz} \quad \text{and} \quad \pm\frac{1}{\sigma^2 yz}\sqrt{\sigma^4(1-y)^2-2\sigma^2(1+y)z+z^2}.$$

Noting that $\zeta_3\zeta_4 = 1$ and recalling the definition for the square root of complex numbers, we know that both the real part and imaginary part of $\sqrt{\sigma^2(1-y)^2-2\sigma^2(1+y)z+z^2}$ and $-\sigma^2(1+y)+z$ have the same signs and hence $|\zeta_3|>1$, $|\zeta_4|<1$. Also, $|\zeta_1| = |-\sqrt{y}| < 1$ and $|\zeta_2| = |-1/\sqrt{y}| > 1$. By Cauchy integration, we obtain

$$\begin{aligned} s(z) &= -\frac{1}{2}\left(\frac{1}{y\sigma^2}-\frac{1}{\sigma^2 yz}\sqrt{\sigma^4(1-y)^2-2\sigma^2(1+y)z+z^2}-\frac{1-y}{yz}\right) \\ &= \frac{\sigma^2(1-y)-z+\sqrt{(z-\sigma^2-y\sigma^2)^2-4y\sigma^4}}{2yz\sigma^2}. \end{aligned}$$

This proves equation (3.3.1) when $y<1$.

When $y>1$, since the M-P law has also a point mass $1-1/y$ at zero, $s(z)$ equals the integral above plus $-(y-1)/yz$. In this case, $|\zeta_3| = |-\sqrt{y}| > 1$ and $|\zeta_4| = |-1/\sqrt{y}| < 1$, and thus the residue at ζ_4 should be counted into the integral. Finally, one finds that equation (3.3.1) still holds. When $y=1$, the equation is still true by continuity in y.

3.3.2 Proof of Theorem 3.10

Let the Stieltjes transform of the ESD of $\mathbf{S}_n$ be denoted by $s_n(z)$. Define

$$s_n(z) = \frac{1}{p}\mathrm{tr}(\mathbf{S}_n - z\mathbf{I}_p)^{-1}.$$

As in Section 2.3, we shall complete the proof by the following three steps:

(i) For any fixed $z\in\mathbb{C}^+$, $s_n(z) - \mathrm{E}s_n(z) \to 0$, a.s.
(ii) For any fixed $z\in\mathbb{C}^+$, $\mathrm{E}s_n(z) \to s(z)$, the Stieltjes transform of the M-P law.
(iii) Except for a null set, $s_n(z)\to s(z)$ for every $z\in\mathbb{C}^+$.

Similar to Section 2.3, the last step is implied by the first two steps and thus its proof is omitted. We now proceed with the first two steps.

Step 1. Almost sure convergence of the random part

$$s_n(z) - \mathrm{E}s_n(z) \to 0, \quad \text{a.s.} \tag{3.3.3}$$

Let $\mathrm{E}_k(\cdot)$ denote the conditional expectation given $\{\mathbf{x}_{k+1}, \cdots, \mathbf{x}_n\}$. Then, by the formula

$$(\mathbf{A} + \boldsymbol{\alpha}\boldsymbol{\beta}^*)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\boldsymbol{\alpha}\boldsymbol{\beta}^*\mathbf{A}^{-1}}{1 + \boldsymbol{\beta}^*\mathbf{A}^{-1}\boldsymbol{\alpha}} \tag{3.3.4}$$

we obtain

$$\begin{aligned} s_n(z) - \mathrm{E}s_n(z) &= \frac{1}{p}\sum_{k=1}^{n}[\mathrm{E}_k\mathrm{tr}(\mathbf{S}_n - z\mathbf{I}_p)^{-1} - \mathrm{E}_{k-1}\mathrm{tr}(\mathbf{S}_n - z\mathbf{I}_p)^{-1}] \\ &= \frac{1}{p}\sum_{k=1}^{n}\gamma_k, \end{aligned}$$

where, by Theorem A.5,

$$\begin{aligned} \gamma_k &= (\mathrm{E}_k - \mathrm{E}_{k-1})[\mathrm{tr}(\mathbf{S}_n - z\mathbf{I}_p)^{-1} - \mathrm{tr}(\mathbf{S}_{nk} - z\mathbf{I}_p)^{-1}] \\ &= -[\mathrm{E}_k - \mathrm{E}_{k-1}]\frac{\mathbf{x}_k^*(\mathbf{S}_{nk} - z\mathbf{I}_p)^{-2}\mathbf{x}_k}{1 + \mathbf{x}_k^*(\mathbf{S}_{nk} - z\mathbf{I}_p)^{-1}\mathbf{x}_k} \end{aligned}$$

and $\mathbf{S}_{nk} = \mathbf{S}_n - \mathbf{x}_k\mathbf{x}_k^*$. Note that

$$\begin{aligned} &\left|\frac{\mathbf{x}_k^*(\mathbf{S}_{nk} - z\mathbf{I}_p)^{-2}\mathbf{x}_k}{1 + \mathbf{x}_k^*(\mathbf{S}_{nk} - z\mathbf{I}_p)^{-1}\mathbf{x}_k}\right| \\ &\le \frac{\mathbf{x}_k^*((\mathbf{S}_{nk} - u\mathbf{I}_p)^2 + v^2\mathbf{I}_p)^{-1}\mathbf{x}_k}{\Im(1 + \mathbf{x}_k^*(\mathbf{S}_{nk} - z\mathbf{I}_p)^{-1}\mathbf{x}_k)} = \frac{1}{v}. \end{aligned}$$

Noticing that $\{\gamma_k\}$ forms a sequence of bounded martingale differences, by Lemma 2.12 with $p = 4$, we obtain

$$\begin{aligned} \mathrm{E}|s_n(z) - \mathrm{E}s_n(z)|^4 &\le \frac{K_4}{p^4}\mathrm{E}\left(\sum_{k=1}^{n}|\gamma_k|^2\right)^2 \\ &\le \frac{4K_4n^2}{v^4p^4} = O(n^{-2}), \end{aligned}$$

which, together with the Borel-Cantelli lemma, implies (3.3.3). The proof is complete.

Step 2. Mean convergence

We will show that

$$\mathrm{E}s_n(z) \to s(z), \tag{3.3.5}$$

where $s(z)$ is defined in (3.3.1) with $\sigma^2 = 1$.

By Theorem A.4, we have

$$s_n(z) = \frac{1}{p}\sum_{k=1}^{p} \frac{1}{\frac{1}{n}\boldsymbol{\alpha}_k'\overline{\boldsymbol{\alpha}}_k - z - \frac{1}{n^2}\boldsymbol{\alpha}_k'\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k}, \quad (3.3.6)$$

where $\mathbf{X}_k$ is the matrix obtained from $\mathbf{X}$ with the k-th row removed and $\boldsymbol{\alpha}_k'$ $(n \times 1)$ is the k-th row of $\mathbf{X}$.

Set

$$\varepsilon_k = \frac{1}{n}\boldsymbol{\alpha}_k'\overline{\boldsymbol{\alpha}}_k - 1 - \frac{1}{n^2}\boldsymbol{\alpha}_k'\mathbf{X}_k^*\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1}\right)^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k + y_n + y_n z\mathrm{E}s_n(z), \quad (3.3.7)$$

where $y_n = p/n$. Then, by (3.3.6), we have

$$\mathrm{E}s_n(z) = \frac{1}{1 - z - y_n - y_n z\mathrm{E}s_n(z)} + \delta_n, \quad (3.3.8)$$

where

$$\delta_n = -\frac{1}{p}\sum_{k=1}^{p}\mathrm{E}\left(\frac{\varepsilon_k}{(1 - z - y_n - y_n z\mathrm{E}s_n(z))(1 - z - y_n - y_n z\mathrm{E}s_n(z) + \varepsilon_k)}\right). \quad (3.3.9)$$

Solving $\mathrm{E}s_n(z)$ from equation (3.3.8), we get two solutions:

$$s_1(z) = \frac{1}{2y_n z}(1 - z - y_n + y_n z\delta_n + \sqrt{(1 - z - y_n - y_n z\delta_n)^2 - 4y_n z}),$$
$$s_2(z) = \frac{1}{2y_n z}(1 - z - y_n + y_n z\delta_n - \sqrt{(1 - z - y_n - y_n z\delta_n)^2 - 4y_n z}).$$

Comparing this with (3.3.1), it suffices to show that

$$\mathrm{E}s_n(z) = s_1(z) \quad (3.3.10)$$

and

$$\delta_n \to 0. \quad (3.3.11)$$

We show (3.3.10) first. Making $v \to \infty$, we know that $\mathrm{E}s_n(z) \to 0$ and hence $\delta_n \to 0$ by (3.3.8). This shows that $\mathrm{E}s_n(z) = s_1(z)$ for all z with large imaginary part. If (3.3.10) is not true for all $z \in \mathbb{C}^+$, then by the continuity of s_1 and s_2, there exists a $z_0 \in \mathbb{C}^+$ such that $s_1(z_0) = s_2(z_0)$, which implies that

$$(1 - z_0 - y_n + y_n z\delta_n)^2 - 4y_n z_0(1 + \delta_n(1 - z_0 - y_n)) = 0.$$

Thus,

$$\mathrm{E}s_n(z_0) = s_1(z_0) = \frac{1 - z_0 - y_n + y_n z_0 \delta_n}{2y_n z_0}.$$

Substituting the solution δ_n of equation (3.3.8) into the identity above, we obtain

$$\mathrm{E}s_n(z_0) = \frac{1 - z_0 - y_n}{y_n z_0} + \frac{1}{y_n + z_0 - 1 + y_n z_0 \mathrm{E}s_n(z_0)}. \tag{3.3.12}$$

Noting that for any Stieltjes transform $s(z)$ of probability F defined on $\mathbb{R}^+$ and positive y, we have

$$\begin{aligned} \Im(y + z - 1 + yzs(z)) &= \Im\left(z - 1 + \int_0^\infty \frac{yxdF(x)}{x - z}\right) \\ &= v\left(1 + \int_0^\infty \frac{yxdF(x)}{(x-u)^2 + v^2}\right) > 0. \end{aligned} \tag{3.3.13}$$

In view of this, it follows that the imaginary part of the second term in (3.3.12) is negative. If $y_n \leq 1$, it can be easily seen that $\Im(1 - z_0 - y_n)/(y_n z_0) < 0$. Then we conclude that $\Im \mathrm{E}s_n(z_0) < 0$, which is impossible since the imaginary part of the Stieltjes transform should be positive. This contradiction leads to the truth of (3.3.10) for the case $y_n \leq 1$.

For the general case, we can prove it in the following way. In view of (3.3.12) and (3.3.13), we should have

$$y_n + z_0 - 1 + y_n z_0 \mathrm{E}s_n(z_0) = \sqrt{y_n z_0}. \tag{3.3.14}$$

Now, let $\underline{s}_n(z)$ be the Stieltjes transform of the matrix $\frac{1}{n}\mathbf{X}^*\mathbf{X}$. Noting that $\frac{1}{n}\mathbf{X}^*\mathbf{X}$ and $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^*$ have the same set of nonzero eigenvalues, we have the relation between s_n and $\underline{s}_n$ given by

$$s_n(z) = y_n^{-1}\underline{s}_n(z) - \frac{1 - 1/y_n}{z}.$$

Note that the equation above is true regardless of whether $y_n > 1$ or $y_n \leq 1$. From this we have

$$y_n - 1 + y_n z_0 \mathrm{E}s_n(z_0) = z_0 \mathrm{E}\underline{s}_n(z_0).$$

Substituting this into (3.3.14), we obtain

$$1 + \mathrm{E}\underline{s}_n(z_0) = \sqrt{y}/\sqrt{z_0},$$

which leads to a contradiction that the imaginary part of LHS is positive and that of the RHS is negative. Then, (3.3.10) is proved.

Now, let us consider the proof of (3.3.11). Rewrite

$$\begin{aligned}\delta_n &= -\frac{1}{p}\sum_{k=1}^{p}\left(\frac{\mathrm{E}\varepsilon_k}{(1-z-y_n-y_nz\mathrm{E}s_n(z))^2}\right)\\ &\quad+\frac{1}{p}\sum_{k=1}^{p}\mathrm{E}\left(\frac{\varepsilon_k^2}{(1-z-y_n-y_nz\mathrm{E}s_n(z))^2(1-z-y_n-y_nz\mathrm{E}s_n(z)+\varepsilon_k)}\right)\\ &= J_1+J_2.\end{aligned}$$

At first, by assumptions given in (3.2.2), we note that

$$\begin{aligned}|\mathrm{E}\varepsilon_k| &= \left|-\frac{1}{n^2}\mathrm{Etr}\mathbf{X}_k^*\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*-z\mathbf{I}_{p-1}\right)^{-1}\mathbf{X}_k+y_n+y_nz\mathrm{E}s_n(z)\right|\\ &= \left|-\frac{1}{n}\mathrm{Etr}\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*-z\mathbf{I}_{p-1}\right)^{-1}\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*+y_n+y_nz\mathrm{E}s_n(z)\right|\\ &\le \frac{1}{n}+\frac{|z|y_n}{n}\mathrm{E}\left|\mathrm{tr}\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*-z\mathbf{I}_{p-1}\right)^{-1}-s_n(z)\right|\\ &\le \frac{1}{n}+\frac{|z|y_n}{nv}\to 0,\end{aligned}\tag{3.3.15}$$

which implies that $J_1\to 0$.

Now we prove $J_2\to 0$. Since

$$\begin{aligned}&\Im\left(1-z-y_n-y_nz\mathrm{E}s_n(z)+\varepsilon_k\right)\\ &=\Im\left(\frac{1}{n}\boldsymbol{\alpha}_k'\overline{\boldsymbol{\alpha}}_k-z-\frac{1}{n^2}\boldsymbol{\alpha}_k'\mathbf{X}_k^*\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*-z\mathbf{I}_{p-1}\right)^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k\right)\\ &=-v\left(1+\frac{1}{n^2}\boldsymbol{\alpha}_k'\mathbf{X}_k^*\left[\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*-u\mathbf{I}_{p-1}\right)^2+v^2\mathbf{I}_{p-1}\right]^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k\right)<-v,\end{aligned}$$

combining this with (3.3.13), we obtain

$$\begin{aligned}|J_2| &\le \frac{1}{pv^3}\sum_{k=1}^{p}\mathrm{E}|\varepsilon_k|^2\\ &= \frac{1}{pv^3}\sum_{k=1}^{p}[\mathrm{E}|\varepsilon_k-\widetilde{\mathrm{E}}(\varepsilon_k)|^2+\mathrm{E}|\widetilde{\mathrm{E}}\varepsilon_k-\mathrm{E}(\varepsilon_k)|^2+(\mathrm{E}(\varepsilon_k))^2],\end{aligned}$$

where $\widetilde{\mathrm{E}}(\cdot)$ denotes the conditional expectation given $\{\boldsymbol{\alpha}_j, j=1,...,k-1,k+1,...,p\}$. In the estimation of J_1, we have proved that

$$|\mathrm{E}(\varepsilon_k)|\le\frac{1}{n}+\frac{|z|y}{nv}\to 0.$$

Write $\mathbf{A}=(a_{ij})=\mathbf{I}_n-\frac{1}{n}\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*-z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k$. Then, we have

$$\varepsilon_k - \tilde{\mathrm{E}}\varepsilon_k = \frac{1}{n}\left(\sum_{i=1}^{n} a_{ii}(|x_{ki}|^2 - 1) + \sum_{i\neq j} a_{ij}x_{ki}\bar{x}_{kj}\right).$$

By elementary calculation, we have

$$\begin{aligned}
&\frac{1}{n^2}\tilde{\mathrm{E}}|\varepsilon_k' - \tilde{\mathrm{E}}\varepsilon_k|^2 \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}|a_{ii}|^2(\mathrm{E}|x_{ki}|^4 - 1) + \sum_{i\neq j}[|a_{ij}|^2\mathrm{E}|x_{ki}|^2\mathrm{E}|x_{kj}|^2 + a_{ij}^2\mathrm{E}x_{ki}^2\mathrm{E}x_{kj}^2]\right) \\
&\leq \frac{1}{n^2}\left(\sum_{i=1}^{n}|a_{ii}|^2(\eta_n^2 n) + 2\sum_{i\neq j}|a_{ij}|^2\right) \\
&\leq \frac{\eta_n^2}{v^2} + \frac{2}{nv^2}.
\end{aligned}$$

Here, we have used the fact that $|a_{ii}| \leq v^{-1}$.

Using the martingale decomposition method in the proof of (3.3.3), we can show that

$$\begin{aligned}
&\mathrm{E}|\widetilde{\mathrm{E}}\varepsilon_k - \mathrm{E}\varepsilon_k|^2 \\
&= \frac{|z|^2y^2}{n^2}\mathrm{E}\left|\mathrm{tr}\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1}\right)^{-1} - \mathrm{Etr}\left(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1}\right)^{-1}\right|^2 \\
&\leq \frac{|z|^2y^2}{nv^2} \to 0.
\end{aligned}$$

Combining the three estimations above , we have completed the proof of the mean convergence of the Stieltjes transform of the ESD of $\mathbf{S}_n$.

Consequently, Theorem 3.10 is proved by the method of Stieltjes transforms.