Dynamical approach to random matrix theory

László Erdős, Horng-Tzer Yau

May 9, 2017

*Partially supported by ERC Advanced Grant, RANMAT 338804
†Partially supported by the NSF grant DMS-1307444 and a Simons Investigator Award
AMS Subject Classification (2010): 15B52, 82B44

Keywords: Random matrix, local semicircle law, Dyson sine kernel, Wigner-Dyson-Mehta conjecture, Tracy-Widom distribution, Dyson Brownian motion.
Preface

This book is a concise and self-contained introduction of the recent techniques to prove local spectral universality for large random matrices. Random matrix theory is a fast expanding research area and this book mainly focuses on the methods we participated in developing over the past few years. Many other interesting topics are not included, nor are several new developments within the framework of these methods. We have chosen instead to present key concepts that we believe are the core of these methods and should be relevant for future applications. We keep technicalities to a minimum to make the book accessible to graduate students. With this in mind, we include in this book the basic notions and tools for high dimensional analysis such as large deviation, entropy, Dirichlet form and logarithmic Sobolev inequality.

The material in this book originates from our joint works with a group of collaborators in the past several years. Not only were the main mathematical results in this book taken from these works, but the presentation of many sections followed the routes laid out in these papers. In alphabetical order, these coauthors were Paul Bourgade, Antti Knowles, Sandrine Péché, Jose Ramírez, Benjamin Schlein and Jun Yin. We would like to thank all of them.

This manuscript was developed and continuously improved over the last five years. We have taught this material in several regular graduate courses at Harvard, Munich and Vienna, in addition to various summer schools and short courses. We are thankful for the generous support of the Institute for Advanced Studies, Princeton, where part of this manuscript was written during the special year devoted to random matrices in 2013-2014. L.E. also thanks Harvard University for the continuous support during his numerous visits. L.E. was partially supported by the SFB TR 12 grant of the German Science Foundation and the ERC Advanced Grant, RANMAT 338804 of the European Research Council. H.-T. Y. would like to thank the National Center for the Theoretic Sciences at the National Taiwan University, where part of the manuscript was written, for the hospitality and support for his several visits. H.-T. Y. gratefully acknowledges the support from NSF DMS-1307444 and a Simons Investigator award.

Finally, we are grateful to the editorial support from the publisher, to Amol Aggarwal, Johannes Alt, Patrick Lopatto for careful reading of the manuscript and to Alex Gontar for his help in composing the bibliography.
### Contents

1 Introduction 6

2 Wigner matrices and their generalizations 10

3 Eigenvalue density 11

3.1 Wigner semicircle law and other canonical densities 11

3.2 The moment method 12

3.3 The resolvent method and the Stieltjes transform 14

4 Invariant ensembles 16

4.1 Joint density of eigenvalues for invariant ensembles 16

4.2 Universality of classical invariant ensembles via orthogonal polynomials 18

5 Universality for generalized Wigner matrices 24

5.1 Different notions of universality 24

5.2 The three-step strategy 25

6 Local semicircle law for universal Wigner matrices 27

6.1 Setup 27

6.2 Spectral information on $S$ 29

6.3 Stochastic domination 30

6.4 Statement of the local semicircle law 31

6.5 Appendix: Behaviour of $\Gamma$ and $\tilde{\Gamma}$ and the proof of Lemma 6.3 33

7 Weak local semicircle law 36

7.1 Proof of the weak local semicircle law, Theorem 7.1 36

7.2 Large deviation estimates 45

8 Proof of the local semicircle law 48

8.1 Tools 48

8.2 Self-consistent equations on two levels 50

8.3 Proof of the local semicircle law without using the spectral gap 52

9 Sketch of the proof of the local semicircle law using the spectral gap 62

10 Fluctuation averaging mechanism 65

10.1 Intuition behind the fluctuation averaging 65

10.2 Proof of Lemma 8.9 66

10.3 Alternative proof of (8.47) of Lemma 8.9 73

11 Eigenvalue location: the rigidity phenomenon 76

11.1 Extreme eigenvalues 76

11.2 Stieltjes transform and regularized counting function 76

11.3 Convergence speed of the empirical distribution function 79

11.4 Rigidity of eigenvalues 80

12 Universality for matrices with Gaussian convolutions 82

12.1 Dyson Brownian motion 82

12.2 Derivation of Dyson Brownian motion and perturbation theory 84

12.3 Strong local ergodicity of the Dyson Brownian motion 85

12.4 Existence and restriction of the dynamics 88
### 13 Entropy and the logarithmic Sobolev inequality (LSI)

- 13.1 Basic properties of the entropy .......................................................... 92
- 13.2 Entropy on product spaces and conditioning ......................................... 94
- 13.3 Logarithmic Sobolev inequality .............................................................. 96
- 13.4 Hypercontractivity ................................................................................. 102
- 13.5 Brascamp-Lieb inequality ....................................................................... 104
- 13.6 Remarks on the applications of the LSI to random matrices .................... 105
- 13.7 Extensions to the simplex, regularization of the DBM ............................. 108

### 14 Universality of the Dyson Brownian Motion

- 14.1 Main ideas behind the proof of Theorem 14.1 ........................................ 114
- 14.2 Proof of Theorem 14.1 ........................................................................ 115
- 14.3 Restriction to the simplex via regularization ......................................... 120
- 14.4 Dirichlet form inequality for any $\beta > 0$ ............................................. 121
- 14.5 From gap distribution to correlation functions: proof of Theorem 12.4 .... 122
- 14.6 Details of the proof of Lemma 14.8 ......................................................... 123

### 15 Continuity of local correlation functions under matrix OU process

- 15.1 Proof of Theorem 15.2 ........................................................................ 129
- 15.2 Proof of the correlation function comparison Theorem 15.3 .................. 131

### 16 Universality of Wigner matrices in small energy window: GFT

- 16.1 The Green function comparison theorems ............................................. 136
- 16.2 Conclusion of the three step strategy ..................................................... 139

### 17 Edge Universality

### 18 Further results and historical notes

- 18.1 Wigner matrices: bulk universality in different senses .......................... 150
- 18.2 Historical overview of the three-step strategy for bulk universality ......... 151
- 18.3 Invariant ensembles and log-gases ......................................................... 153
- 18.4 Universality at the spectral edge ............................................................ 155
- 18.5 Eigenvectors ......................................................................................... 155
- 18.6 General mean field models .................................................................. 157
- 18.7 Beyond mean field models: band matrices ........................................... 158
1 Introduction

“Perhaps I am now too courageous when I try to guess the distribution of the distances between successive levels (of energies of heavy nuclei). Theoretically, the situation is quite simple if one attacks the problem in a simpleminded fashion. The question is simply what are the distances of the characteristic values of a symmetric matrix with random coefficients.”

Eugene Wigner on the Wigner surmise, 1956

Random matrices appeared in the literature as early as 1928, when Wishart [138] used them in statistics. The natural question regarding their eigenvalue statistics, however, was not raised until the pioneering work [137] of Eugene Wigner in the 1950s. Wigner’s original motivation came from nuclear physics when he noticed from experimental data that gaps in energy levels of large nuclei tend to follow the same statistics irrespective of the material. Quantum mechanics predicts that energy levels are eigenvalues of a self-adjoint operator, but the correct Hamiltonian operator describing nuclear forces was not known at that time. In addition, the computation of the energy levels of large quantum systems would have been impossible even with the full Hamiltonian explicitly given. Instead of pursuing a direct solution of this problem, Wigner appealed to a phenomenological model to explain his observation. Wigner’s pioneering idea was to model the complex Hamiltonian by a random matrix with independent entries. All physical details of the system were ignored except one, the symmetry type: systems with time reversal symmetry were modeled by real symmetric random matrices, while complex Hermitian random matrices were used for systems without time reversal symmetry (e.g. with magnetic forces). This simple-minded model amazingly reproduced the correct gap statistics, indicating a profound universality principle working in the background.

Notwithstanding their physical significance, random matrices are also very natural mathematical objects and their studies could have been initiated by mathematicians driven by pure curiosity. A large number of random numbers and vectors have been known to exhibit universal patterns; the obvious examples are the law of large numbers and the central limit theorem. What are their analogues in the non-commutative setting, e.g. for matrices? Focusing on the spectrum, what do eigenvalues of typical large random matrices look like?

As the first result of this type, Wigner proved a type of law of large numbers for the density of eigenvalues, which we now explain. The (real or complex) Wigner ensembles consist of $N \times N$ self-adjoint matrices $H = (h_{ij})$ with matrix elements having mean zero and variance $1/N$ that are independent up to the symmetry constraint $h_{ij} = \bar{h}_{ji}$. The Wigner semicircle law states that the empirical density of the eigenvalues of $H$ is given by the semicircle law, $\varrho_{\text{sc}}(x) = \frac{1}{2\pi} \sqrt{4 - x^2}$, as $N \to \infty$, independent of the details of the distribution of $h_{ij}$.

On the scale of individual eigenvalues, Wigner predicted that the fluctuations of the gaps are universal and their distribution is given by a new law, the Wigner surmise. This might be viewed as the random matrix analogue of the central limit theorem.

After Wigner’s discovery, Dyson, Gaudin and Mehta achieved several fundamental mathematical results, in particular they were able to compute the gap distribution and the local correlation functions of the eigenvalues for Gaussian ensembles. They are called Gaussian Orthogonal Ensemble (GOE) and Gaussian Unitary Ensemble (GUE) corresponding to the two most important symmetry classes, the real symmetric and complex Hermitian matrices. The Wigner surmise turned out to be slightly wrong and the correct law is given by the Gaudin distribution. Dyson and Mehta gradually formulated what is nowadays known as the Wigner-Dyson-Mehta (WDM) universality conjecture. As presented in the classical treatise of Mehta [105], this conjecture asserts that the local eigenvalue statistics for large random matrices with independent entries are universal, i.e., they do not depend on the particular distribution of the matrix elements. In particular, they coincide with those in the Gaussian case that were computed explicitly.

On a more philosophic level, we can recast Wigner’s vision as the hypothesis that the eigenvalue gap distributions for large complicated quantum systems are universal in the sense that they depend only on the symmetry class of the physical system but not on other detailed structures. Therefore, the Wigner-Dyson-Mehta universality conjecture is merely a test of Wigner’s hypothesis for a special class of matrix models, the Wigner ensembles, which are characterized by the independence of their matrix elements. The other large
The conceptually novel point of our method is Step 2. The eigenvalue distributions of the Gaussian divisible ensembles, written in the form $e^{-t/2}H_0 + \sqrt{1-e^{-t}}H^G$, are the same as that of the solution of a \textit{matrix valued Ornstein-Uhlenbeck (OU) process} $H_t$ for any time $t \geq 0$. Dyson [45] observed half a century
ago that the dynamics of the eigenvalues of $H_t$ is given by an interacting stochastic particle system, called the Dyson Brownian motion (DBM). In addition, the invariant measure of this dynamics is exactly the eigenvalue distribution of GOE or GUE. This invariant measure is also a Gibbs measure of point particles in one dimension interacting via a long range logarithmic potential. Using a heuristic physical argument, Dyson remarked [45] that the DBM reaches its “local equilibrium” on a short time scale $t \gtrsim N^{-1}$. We will refer to this as Dyson’s conjecture, although it was rather an intuitive physical picture than an exact mathematical statement.

Since Dyson’s work in the sixties, there was virtually no progress in proving this conjecture. Besides the limit of available mathematical tools, one main factor is the vague formulation of the conjecture involving the notion of “local equilibrium”, which even nowadays is not well-defined for a Gibbs measure with a general long range interaction. Furthermore, a possible connection between Dyson’s conjecture and the solution of the WDM conjecture has never been elucidated in the literature.

In fact, “relaxation to local equilibrium” in a time scale $t$ refers to the phenomenon that after time $t$ the dynamics has changed the system, at least locally, from its initial state to a local equilibrium. It therefore appears counterintuitive that one may learn anything useful in this way about the WDM conjecture, since the latter concerns the initial state. The key point is that by applying local relaxation to all initial states (within a reasonable class) simultaneously, Step 2 generates a large set of random matrix ensembles for which universality holds. We prove that, for the purpose of universality, this set is sufficiently dense so that any Wigner matrix $H$ is sufficiently close to a Gaussian divisible ensemble of the form $e^{-t/2}H_0 + \sqrt{1-e^{-t}}H^G$ with a suitably chosen $H_0$. Originally, our motivation was driven not by Dyson’s conjecture, but by the desire to prove the universality for Gaussian divisible ensembles with the Gaussian component as small as possible. As it turned out, our method used in Step 2 actually proved a precisely formulated version of Dyson’s conjecture.

The applicability of our approach actually goes well beyond the original scope; in the last years the method has been extended to invariant ensembles, sample covariance matrices, adjacency matrices of sparse random graphs and certain band matrices. At the end of the book we will give a short overview of these applications. In the following, we outline the main contents of the book.

In Sections 2–4 we start with an elementary recapitulation of some well known concepts, such as the moment method, the orthogonal polynomial method, the emergence of the sine kernel, the Tracy-Widom distribution and the invariant ensembles. These sections are included only to provide backgrounds for the topics discussed in this book; details are often omitted, since these issues are discussed extensively in other books. The core material starts from Section 5 with a precise definition of different concepts of universality and with the formulation of a representative theorem (Theorem 5.1) that we eventually prove in this book. We also give here a detailed overview of the three step strategy.

The first step, the local version of Wigner’s semicircle law is given in Section 6. For pedagogical reasons, we give the proof on several levels. A weaker version is given in Section 7 whose proof is easier; an ambitious reader may skip this section. The proof of the stronger version is found in Section 8 which gives the optimal result in the bulk of the spectrum. To obtain the optimal result at the edge, an additional idea is needed; this is sketched in Section 9, but we do not give all the technical details. This section may be skipped at first reading. A key ingredient of the proofs, the fluctuation averaging mechanism, is presented separately in Section 10. Important consequences of the local law, such as rigidity, speed of convergence and eigenvector delocalization, are given in Section 11.

The second step, the analysis of the Dyson Brownian motion and the proof of Dyson’s conjecture, is presented in Section 12 and Section 14. Before the proof in Section 14, we added a separate Section 13 devoted to general tools of very large dimensional analysis, where we introduce the concepts of entropy, Dirichlet form, Bakry-Émery method, and logarithmic Sobolev inequality. Some concepts (e.g. the Brascamp-Lieb inequality and hypercontractivity) are not used in this book but we included them since they are useful and related topics.

The third step, the perturbative argument to compare arbitrary Wigner matrices with Gaussian divisible matrices, is given in two versions. A shorter but somewhat weaker version, the continuity of the correlation functions, is found in Section 15. A technically more involved version, the Green function comparison theorem is presented in Section 16. This completes the proof of our main result.
Separately, in Section 17 we prove the universality at the edge; the main input is the strong form of the local semicircle law at the edge. Finally, we give a historical overview, discuss newer results, and provide some outlook on related models in Section 18.

This book is not a comprehensive monograph on random matrices. Although we summarize a few general concepts at the beginning, we mostly focus on a concrete strategy and the necessary technical steps behind it. For readers interested in other aspects, in addition to the classical book of Mehta [105], several excellent works are available that present random matrices in a broader scope. The books by Anderson, Guionnet and Zeitouni [8] and Pastur and Shcherbina [112] contain extensive material starting from the basics. Tao’s book [127] provides a different aspect to this subject and is self-contained as a graduate textbook. Forrester’s monograph [72] is a handbook for any explicit formulas related to random matrices. Finally, [6] is an excellent comprehensive overview of diverse applications of random matrix theory in mathematics, physics, neural networks and engineering.

Finally, we list a few notational conventions. In order to focus on the essentials, we will not follow the dependencies of various constants on different parameters. In particular, we will use the generic letters $C$ and $c$ to denote positive constants, whose values may change from line to line and which may depend on some fixed basic parameters of the model. For two positive quantities $A$ and $B$, we will write $A \lesssim B$ to indicate that there exists a constant $C$ such that $A \leq CB$. If $A$ and $B$ are comparable in the sense that $A \lesssim B$ and $B \lesssim A$, then we write $A \asymp B$. We introduce the notation $[A, B] := \mathbb{Z} \cap [A, B]$ for the set of integers between any two real numbers $A < B$. 

2 Wigner matrices and their generalizations

Consider $N \times N$ square matrices of the form

$$H = H^{(N)} = \begin{pmatrix} h_{11} & h_{12} & \ldots & h_{1N} \\ h_{21} & h_{22} & \ldots & h_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \ldots & h_{NN} \end{pmatrix}$$

(2.1)

with centered entries

$$\mathbb{E} h_{ij} = 0, \quad i, j = 1, 2, \ldots, N.$$  

(2.2)

The random variables $h_{ij}, i, j = 1, \ldots, N$ are real or complex random variables subject to the symmetry constraint $h_{ij} = \bar{h}_{ji}$ so that $H$ is either Hermitian (complex) or symmetric (real). Let $S = S^{(N)}$ denote the matrix of variances

$$S := \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1N} \\ s_{21} & s_{22} & \ldots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \ldots & s_{NN} \end{pmatrix}, \quad s_{ij} := \mathbb{E}|h_{ij}|^2.$$  

(2.3)

We assume that

$$\sum_{j=1}^{N} s_{ij} = 1, \quad \text{for every} \quad i = 1, 2, \ldots, N,$$

(2.4)

i.e., the deterministic $N \times N$ matrix of variances, $S = (s_{ij})$, is symmetric and doubly stochastic. The size of the random matrix, $N$, is the fundamental limiting parameter throughout this book. Most results become sharp in the large $N$ limit. Many quantities, such as the distribution of $H$, the matrix $S$, naturally depend on $N$, but for notational simplicity we will often omit this dependence from the notation.

The simplest case, when

$$s_{ij} = \frac{1}{N} \quad i, j = 1, 2, \ldots, N,$$

is called the Wigner matrix ensemble. We summarize this setup in the following definition:

**Definition 2.1.** An $N \times N$ symmetric or Hermitian random matrix (2.1) is called a universal Wigner matrix (ensemble) if the entries are centered (2.2), their variances $s_{ij} = \mathbb{E}|h_{ij}|^2$ satisfy

$$\sum_{j=1}^{N} s_{ij} = 1, \quad i = 1, 2, \ldots, N,$$

(2.5)

and $\{h_{ij} : i \leq j\}$ are independent. An important subclass of the universal Wigner ensembles is called generalized Wigner matrices (ensembles) if, additionally, the variances are comparable, i.e.,

$$0 < C_{\text{inf}} \leq N s_{ij} \leq C_{\text{sup}} < \infty, \quad i, j = 1, 2, \ldots, N,$$

(2.6)

holds with some fixed positive constants $C_{\text{inf}}, C_{\text{sup}}$. For Hermitian ensembles, we additionally require that for each $i, j$ the $2 \times 2$ covariance matrix is bounded by $C/N$ in matrix sense, i.e.,

$$\Sigma_{ij} := \begin{pmatrix} \mathbb{E}(\Re h_{ij})^2 & \mathbb{E}(\Re h_{ij})(\Im h_{ij}) \\ \mathbb{E}(\Re h_{ij})(\Im h_{ij}) & \mathbb{E}(\Im h_{ij})^2 \end{pmatrix} \preceq \frac{C}{N}.$$

In the special case $s_{ij} = 1/N$ and $\Sigma_{ij} = \frac{1}{2N} I_{2 \times 2}$ we recover the original definition of the Wigner matrices or Wigner ensemble [137].
In this book we will focus on the generalized Wigner matrices. These are mean-field models in the sense that the typical size of the matrix elements $h_{ij}$ are comparable, see (2.6). In Section 18.7 we will discuss models beyond the mean field regime.

The most prominent Wigner ensembles are the Gaussian Orthogonal Ensemble (GOE) and the Gaussian Unitary Ensemble (GUE); i.e., real symmetric and complex Hermitian Wigner matrices with rescaled matrix elements $\sqrt{Nh_{ij}}$ being standard Gaussian variables. More precisely, in the Hermitian case, $\sqrt{Nh_{ij}}$ is a standard complex Gaussian variable, i.e., $E[\sqrt{Nh_{ij}}^2] = 1$ with the real and imaginary part independent having the same variance when $i \neq j$. Furthermore, $\sqrt{Nh_{ij}}$ is a real standard Gaussian variable with variance 1. For the real symmetric case, we require that $E[\sqrt{Nh_{ii}}^2] = 1 = E[\sqrt{Nh_{kk}}^2]/2$ for any $i \neq j$.

For simplicity of the presentation, in the case of the Wigner ensembles, we will assume that not only the variances of $h_{ij}$, $i < j$ are identical, but they are identically distributed. In this case we fix a distribution $\nu$ and we assume that the rescaled matrix elements $\sqrt{Nh_{ij}}$ are distributed according to $\nu$. Depending on the symmetry type, the diagonal elements may have a slightly different distribution, but we will omit this subtlety from the discussion. The distribution $\nu$ will be called the single entry distribution of $H$. In the case of generalized Wigner matrices, we assume that the rescaled matrix elements $s_{ij}^{-1/2}h_{ij}$, with unit variance, all have law $\nu$.

We will need some decay property of $\nu$, and we will consider two decay types. Either we assume that $\nu$ has subexponential decay, i.e., that there are constants $C, \vartheta > 0$ such that for any $s > 0$

$$\int_{\mathbb{R}} 1(|x| \geq s) d\nu(x) \leq C \exp(-s^\vartheta); \tag{2.7}$$

or we assume that $\nu$ has a polynomial decay with arbitrary high power, i.e.,

$$\int_{\mathbb{R}} 1(|x| \geq s) d\nu(x) \leq C_\nu (1 + s)^{-p}, \quad \forall \ s > 0, \tag{2.8}$$

for any $p \in \mathbb{N}$.

Another class of random matrices, which even predate Wigner, are the random sample covariance matrices. These are matrices of the form

$$H = X^*X, \tag{2.9}$$

where $X$ is a rectangular $M \times N$ matrix with centered i.i.d. entries with variance $E[X_{ij}] = M^{-1}$. Note that the matrix elements of $H$ are not independent, but they are generated from the independent matrix elements of $X$ in a straightforward way. These matrices appear in statistical samples and were first considered by Wishart [138]. In the case when $X_{ij}$ are centered Gaussian, the random covariance matrices are called Wishart matrices or ensemble.

3 Eigenvalue density

3.1 Wigner semicircle law and other canonical densities

For real symmetric or complex Hermitian Wigner matrix $H$, let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ denote the eigenvalues. The empirical distribution of eigenvalues follows a universal pattern, the Wigner semicircle law. To formulate it more precisely, note that the typical spacing between neighboring eigenvalues is of order $1/N$, so in a fixed interval $[a, b] \subset \mathbb{R}$, one expects macroscopically many (of order $N$) eigenvalues. More precisely, it can be shown (first proof was given by Wigner [137]) that for any fixed $a \leq b$ real numbers,

$$\lim_{N \to \infty} \frac{1}{N} \# \left\{ i : \lambda_i \in [a, b] \right\} = \int_a^b \psi_{sc}(x) dx, \quad \psi_{sc}(x) := \frac{1}{2\pi} \sqrt{(4 - x^2)_+}, \tag{3.1}$$

where $(a)_+ := \max\{a, 0\}$ denotes the positive part of the number $a$. Note the emergence of the universal density, the semicircle law, that is independent of the details of the distribution of the matrix elements. The
semicircle law holds beyond Wigner matrices: it characterizes the eigenvalue density of the universal Wigner matrices (see Definition 2.1).

For the random covariance matrices (2.9), the empirical density of eigenvalues \( \lambda_i \) of \( H \) converges to the Marchenko-Pastur law \([103]\) in the limit when \( M, N \to \infty \) such that \( d = N/M \) is fixed \( 0 < d \leq 1 \):

\[
\lim_{N \to \infty} \frac{1}{N} \# \{ i : \lambda_i \in [a, b] \} = \int_a^b \varrho_{\text{MP}}(x) \, dx, \quad \varrho_{\text{MP}}(x) := \frac{1}{2\pi d} \sqrt{\frac{(\lambda_+ - x)(x - \lambda_-)}{x^2}} \tag{3.2}
\]

with \( \lambda_{\pm} := (1 \pm \sqrt{d})^2 \) being the spectral edges. Note that in case \( M \leq N \), the matrix \( H \) has macroscopically many zero eigenvalues, otherwise the spectra of \( XX^* \) and \( X^*X \) coincide so the Marchenko-Pastur law can be applied to all nonzero eigenvalues with the role of \( M \) and \( N \) exchanged.

### 3.2 The moment method

The eigenvalue density is commonly approached via the fairly robust moment method (see [8] for an exposé) that was also the original approach of Wigner to prove the semicircle law \([137]\). The following proof is formulated for Wigner matrices with identically distributed entries, in particular with variances \( s_{ij} = N^{-1} \), but a slight modification of the argument also applies to more general ensembles satisfying (2.5). The simplest moment is the second moment, \( \text{Tr} H^2 \):

\[
\sum_{i=1}^N \lambda_i^2 = \text{Tr} H^2 = \sum_{i,j=1}^N |h_{ij}|^2.
\]

Taking expectation and using (2.5) we have

\[
\frac{1}{N} \sum_i \mathbb{E} \lambda_i^2 = \frac{1}{N} \sum_{ij} s_{ij} = 1.
\]

In general, we can compute traces of all even powers of \( H \), i.e.,

\[
\mathbb{E} \text{Tr} H^{2k}
\]

by expanding the product as

\[
\mathbb{E} \sum_{i_1, i_2, \ldots, i_{2k}} h_{i_1i_2}h_{i_2i_3} \ldots h_{i_{2k}i_1}. \tag{3.3}
\]

Since the expectation of each matrix element vanishes, each factor \( h_{xy} \) must be paired with at least another copy \( h_{yx} = h_{xy} \), otherwise the expectation value is zero. In general, there is no need that the sequence form a perfect pairing. If we identify \( h_{yx} \) with \( h_{xy} \), the only restriction is that the matrix element \( h_{xy} \) should appear at least twice in the sequence. Under this condition, the main contribution comes from the index sequences which satisfy the perfect pairing condition such that each \( h_{xy} \) is paired with a unique \( h_{yx} \) in (3.3). These sequences can be classified according to their complexity, and it turns out that the main contribution comes from the so-called backtracking sequences.

An index sequence \( i_1i_2i_3 \ldots i_{2k}i_1 \), returning to the original index \( i_1 \), is called backtracking if it can be successively generated by a substitution rule

\[
a \to aba, \quad b \in \{1, 2, \ldots, N\}, \quad b \neq a, \tag{3.4}
\]

with an arbitrary index \( b \). For example, we represent the term

\[
h_{i_1i_2}h_{i_2i_3}h_{i_3i_4}h_{i_4i_5}h_{i_5i_6}h_{i_6i_7}h_{i_7i_1}, \quad i_1, i_2, \ldots, i_5 \in \{1, 2, \ldots, N\}, \quad k = 4
\]

in the expansion of \( \text{Tr} H^8 \) \((k = 4)\) by a walk of length \( 2 \times 4 \) on the set \( \{1, 2, \ldots, N\} \). This path is generated by the operation (3.4) in the following order

\[
i_1 \to i_1i_2i_1 \to i_1i_2i_3i_2i_1 \to i_1i_2i_3i_2i_4i_3i_2i_1 \to i_1i_2i_3i_2i_4i_5i_4i_2i_1, \quad \text{with} \quad i_1 \neq i_2, i_2 \neq i_3, i_3 \neq i_4, i_4 \neq i_5.
\]
It may happen that two non-successive labels coincide (e.g. $i_1 = i_3$), but the contribution of such terms is by a factor $1/N$ less than the leading term so we can neglect them. Thus we assume that all labels $i_1, i_2, i_3, i_4, i_5$ are distinct. We may bookkeep these paths by two objects:

a) a graph on vertices labelled by $1, 2, 3, 4, 5$ (i.e., by the indices of the labels $i_j$) and with edges defined by walking over the vertices in following order $1, 2, 3, 2, 4, 5, 4, 2, 1$, i.e., the order of the indices of the labels in (3.5);

b) an assignment of a distinct label $i_j, j = 1, 2, 3, 4, 5$ to each vertex.

It is easy to see that the graph generated by backtracking sequences is a (double-edged) tree on the vertices $1, 2, 3, 4, 5$. Now we count the combinatorics of the objects a) and b) separately. We start with a).

**Lemma 3.1.** The number of graphs with $k$ double-edges, derived from backtracking sequences, is explicitly given by the Catalan numbers, $C_k := \frac{1}{k+1} \binom{2k}{k}$.

**Proof.** There is a one to one correspondence between backtracking sequences and the number of nonnegative one dimensional random walks of length $2k$ returning to the origin at the time $2k$. This is defined by the simple rule, that a forward step $a \rightarrow b$ in the substitution rule (3.4) will be represented by a step to the right in the 1d random walk, while the backtracking step $b \rightarrow a$ in (3.4) is step to the left. Since at any moment the number of backtracking steps cannot be larger than the number of forward steps, the walk remains on the nonnegative semi-axis.

From this interpretation, it is easy to show that the recursive relation

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-k-1}, \quad C_0 = C_1 = 1,$$

holds for the number $C_n$ of the backtracking sequences of total length $2n$. Elementary combinatorics shows that the solution to this recursion is given by the Catalan numbers, which proves the lemma. We remark that $C_n$ has many other definitions. Alternatively, it is also the number of planar binary trees with $n+1$ vertices and one could also use this definition to verify the lemma.

The combinatorics of label-assignments in step b) for backtracking paths of length $2k$ is $N(N-1)\ldots(N-k) \approx N^{k+1}$ for any fixed $k$ in the $N \rightarrow \infty$ limit. It is easy to see that among all paths of length $2k$, the backtracking paths support the maximal number $(k+1)$ independent indices. The combinatorics for all non-backtracking paths is at most $CN^k$, i.e., by a factor $1/N$ less, thus they are negligible. Hence $E \text{Tr} \ H^{2k}$ can be computed fairly precisely for each finite $k$:

$$\frac{1}{N} E \text{Tr} \ H^{2k} = \frac{1}{k+1} \binom{2k}{k} + O(k N^{-1}).$$

(3.6)

Note that the number of independent labels, $N^{k+1}$, after dividing by $N$, exactly cancels the size of the $k$-fold product of variances, $(\mathbb{E}|h|^2)^k = N^{-k}$.

The expectation of the traces of odd powers of $H$ is negligible since they can never satisfy the pairing condition. One can easily check that

$$\int_\mathbb{R} x^{2k} \rho_{sc}(x) dx = C_k,$$

(3.7)

i.e., the semicircle law is identified as the probability measure on $\mathbb{R}$ whose even moments are the Catalan numbers and the odd moments vanish. This proves that

$$\frac{1}{N} E \text{Tr} P(H) \rightarrow \int_\mathbb{R} P(x) \rho_{sc}(x) dx$$

(3.8)

for any polynomial $P$. With standard approximation arguments, the polynomials can be replaced with any continuous functions with compact support and even with characteristic functions of intervals. This proves
the semicircle law (3.1) for the Wigner matrices in the “expectation sense”. To rigorously complete the proof of the semicircle law (3.1) without expectation, we also need to control the variance of $\text{Tr} P(H)$ in the left hand side of (3.8). Although this can also be done by the moment method, we will not get into further details here; see [8] for a detailed proof.

Finally, we also remark that there are index sequences in (3.3) that require to compute higher than the second moment of $h$. While the combinatorics of these terms is negligible, the proof above assumed polynomial decay (2.8), i.e., that all rescaled moments of $h$ are finite, $E|h_{ij}|^m \leq C_m N^{-m/2}$. This condition can be relaxed by a cutoff argument which we skip here. The interested reader should consult [8] for more details.

### 3.3 The resolvent method and the Stieltjes transform

An alternative approach to the eigenvalue density is via the Stieltjes transform. The empirical measure of the eigenvalues is defined by

$$\varrho_N(dx) := \frac{1}{N} \sum_{j=1}^{N} \delta(x - \lambda_j)dx. $$

The Stieltjes transform of the empirical measure is defined by

$$m(z) = m_N(z) := \frac{1}{N} \text{Tr} \frac{1}{H - z} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_j - z} = \int_{\mathbb{R}} \frac{d\varrho_N(x)}{x - z}, $$

for any $z = E + i\eta$, $E \in \mathbb{R}$, $\eta > 0$. Notice that $m$ is simply the normalized trace of the resolvent of the random matrix $H$ with spectral parameter $z$. The real part $E = \text{Re} z$ will often be referred to "energy", alluring to the quantum mechanical interpretation of the spectrum of $H$. An important property of the Stieltjes transform of any measure on $\mathbb{R}$ is that its imaginary part is positive whenever $\text{Im} z > 0$.

For large $z$ one can expand $m_N$ as follows

$$m_N(z) = \frac{1}{N} \text{Tr} \frac{1}{H - z} = -\frac{1}{Nz} \sum_{m=0}^{\infty} \left( \frac{H}{z} \right)^m, $$

so after taking the expectation, using (3.6) and neglecting the error terms, we get

$$\mathbb{E}m_N(z) \approx -\sum_{k=0}^{\infty} \frac{1}{k+1} \left( \frac{2k}{k} \right) (-z)^{-(2k+1)}, $$

which, after some calculus, can be identified as the Laurent series of $\frac{1}{2}(-z + \sqrt{z^2 - 4})$. The approximation becomes exact in the $N \to \infty$ limit. Although the expansion (3.10) is valid only for large $z$, given that the limit is an analytic function of $z$, one can extend the relation

$$\lim_{N \to \infty} \mathbb{E}m_N(z) = \frac{1}{2}(-z + \sqrt{z^2 - 4}) $$

by analytic continuation to the whole upper half plane $z = E + i\eta$, $\eta > 0$. It is an easy exercise to see that this is exactly the Stieltjes transform of the semicircle density, i.e.,

$$m_{sc}(z) := \frac{1}{2}(-z + \sqrt{z^2 - 4}) = \int_{\mathbb{R}} \frac{\varrho_{sc}(x)dx}{x - z}. $$

The square root function is chosen with a branch cut in the segment $[-2, 2]$ so that $\sqrt{z^2 - 4} \approx z$ at infinity. This guarantees that $\text{Im} m_{sc}(z) > 0$ for $\text{Im} z > 0$. Since the Stieltjes transform identifies the measure uniquely, and pointwise convergence of Stieltjes transforms implies weak convergence of measures, we obtain

$$\mathbb{E} \varrho_N(dx) \rightharpoonup \varrho_{sc}(x).$$
The relation (3.12) actually holds with high probability, i.e., for any $z$ with $\text{Im} \, z > 0$,

$$
\lim_{N \to \infty} m_N(z) = \frac{1}{2}(-z + \sqrt{z^2 - 4}),
$$

(3.15)
in probability. In the next sections we will prove this limit with an effective error term via the resolvent method.

The semicircle law can be identified in many different ways. The moment method in Section 3.2 utilized the fact that the moments of the semicircle density are given by the Catalan numbers (3.7), which also emerged as the normalized traces of powers of $H$, see (3.6). The resolvent method relies on the fact that $m_N$ approximately satisfies a self-consistent equation, $m_N \approx -(z + m_N)^{-1}$, that is very close to the quadratic equation that $m_{sc}$ from (3.13) satisfies:

$$
m_{sc}(z) = -\frac{1}{z + m_{sc}(z)}.
$$

In other words, in the resolvent method the semicircle density emerges via a specific relation for its Stieltjes transform. It turns out that this approach allows us to perform a much more precise analysis, especially in the short scale regime, where $\text{Im} \, z$ approaches to 0 as a function of $N$. Since the Stieltjes transform of a measure at spectral parameter $z = E + i\eta$ essentially identifies the measure around $E$ on scale $\eta > 0$, a precise understanding of $m_N(z)$ for small $\text{Im} \, z$ will yield a local version of the semicircle law. This will be explained in Section 6.
4 Invariant ensembles

4.1 Joint density of eigenvalues for invariant ensembles

There is another natural way to define probability distributions on real symmetric or complex Hermitian matrices apart from directly imposing a given probability law $\nu$ on their entries. They are obtained by defining a density function directly on the set of matrices:

$$P(H)dH := Z^{-1} \exp (-N \text{Tr} V(H))dH.$$  \hspace{1cm} (4.1)

Here $dH = \prod_{i<j} dH_{ij}$ is the flat Lebesgue measure (in case of complex Hermitian matrices and $i<j$, $dH_{ij}$ is the Lebesgue measure on the complex plane $\mathbb{C}$). The function $V : \mathbb{R} \to \mathbb{R}$ is assumed to grow mildly at infinity (some logarithmic growth would suffice) to ensure that the measure defined in (4.2) is finite, and $Z$ is the normalization factor. Probability distributions of the form (4.1) are called invariant ensembles since they are invariant under the orthogonal or unitary conjugation (in case of symmetric or Hermitian matrices, respectively). For example, in the Hermitian case, for any fixed unitary matrix $U$, the transformation $H \to U^*HU$ leaves the distribution (4.1) invariant thanks to $\text{Tr} V(U^*HU) = \text{Tr} V(H)$.

Wigner matrices and invariant ensembles form two different universes with quite different mathematical tools available for their studies. In fact, these two classes are almost disjoint, the Gaussian ensembles being the only invariant Wigner matrices. This is the content of the following lemma:

Lemma 4.1 ([37] or Theorem 2.6.3 [105]). Suppose that the real symmetric or complex Hermitian matrix ensembles given in (4.1) have independent entries $h_{ij}$, $i \leq j$. Then $V(x)$ is a quadratic polynomial, $V(x) = ax^2 + bx + c$ with $a > 0$. This means that apart from a trivial shift and normalization, the ensemble is GOE or GUE.

For ensembles that remain invariant under the transformations $H \to U^*HU$ for any unitary matrix $U$ (or, in case of symmetric matrices $H$, for any orthogonal matrix $U$), the joint probability density function of all the $N$ eigenvalues can be explicitly computed. These ensembles are typically given by

$$P(H)dH = Z^{-1} \exp \left(-\frac{\beta}{2} N \text{Tr} V(H)\right)dH.$$  \hspace{1cm} (4.2)

which is of the form (4.1) with a traditional extra factor $\beta/2$ that makes some later formulas nicer. The parameter $\beta$ is determined by the symmetry type; $\beta = 1$ for real symmetric ensembles and $\beta = 2$ for complex Hermitian ensembles.

The joint (symmetrized) probability density of the eigenvalues of $H$ can be computed explicitly:

$$p_N(\lambda_1, \lambda_2, \ldots, \lambda_N) = \text{const.} \prod_{i<j} (\lambda_i - \lambda_j)^{\beta} e^{-\frac{\beta}{4} N \sum_{j=1}^N V(\lambda_j)}.$$  \hspace{1cm} (4.3)

In particular, for the Gaussian case, $V(\lambda) = \frac{1}{2} \lambda^2$ is quadratic and thus the joint distribution of the GOE ($\beta = 1$) and GUE ($\beta = 2$) eigenvalues is given by

$$p_N(\lambda_1, \lambda_2, \ldots, \lambda_N) = \text{const.} \prod_{i<j} (\lambda_i - \lambda_j)^{\beta} e^{-\frac{\beta}{4} N \sum_{j=1}^N \lambda_j^2}.$$  \hspace{1cm} (4.4)

In particular, the eigenvalues are strongly correlated. (In this section we neglect the ordering of the eigenvalues and we will consider symmetrized statistics.)

The emergence of the Vandermonde determinant in (4.3) is a result of integrating out the "angle" variables in (4.2), i.e., the unitary matrix in the diagonalization of $H = U\Lambda U^*$. For illustration, we now show this...
formula for a $2 \times 2$ matrix. Consider first the real case. By diagonalization, any real symmetric $2 \times 2$ matrix can be written in the form

$$H = \begin{pmatrix} x & z \\ z & y \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}, \quad x, y, z \in \mathbb{R}. \quad (4.5)$$

Direct calculation shows that the Jacobian of the coordinate transformation from $(x, y, z)$ to $(\lambda_1, \lambda_2, \theta)$ is

$$J = (\lambda_1 - \lambda_2). \quad (4.6)$$

The complex case is slightly more complicated. We can write with $z = u + iv$ and $x, y \in \mathbb{R}$

$$H = \begin{pmatrix} x & z \\ \bar{z} & y \end{pmatrix} = e^{iA} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} e^{-iA}, \quad (4.7)$$

where $A$ is a Hermitian matrix with trace zero, thus it can be written in the form

$$A = \begin{pmatrix} a & b + ic \\ b - ic & -a \end{pmatrix}, \quad a, b, c \in \mathbb{R}. \quad (4.8)$$

This parametrization of $SU(2)$ with three real degrees of freedom is standard but for our purpose we only need two in (4.7) in addition to the two degrees of freedoms from $\lambda$’s. The reason is that the two phases of the eigenvectors are redundant and the trace zero condition only takes out one degree of freedom, leaving one more superfluous parameter. We will see that eventually it plays no role. First we evaluate the Jacobian at $A = 0$ from the formula (4.7), we only need to keep the leading order in $A$ which gives

$$\begin{pmatrix} x & z \\ \bar{z} & y \end{pmatrix} = \Lambda + i[A, \Lambda] + O(\|A\|^2), \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \quad (4.9)$$

Thus

$$\begin{pmatrix} x & z \\ \bar{z} & y \end{pmatrix} = \begin{pmatrix} \lambda_1 & i(b + ic)(\lambda_2 - \lambda_1) \\ -i(b - ic)(\lambda_2 - \lambda_1) & \lambda_2 \end{pmatrix} + O(\|A\|^2), \quad (4.10)$$

and the Jacobian of the transformation from $(x, y, z)$ to $(\lambda_1, \lambda_2, b, c)$ at $b = c = 0$ is of the form

$$C(\lambda_2 - \lambda_1)^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & i & -i \\ 0 & 0 & -1 & -1 \end{bmatrix} = C(\lambda_1 - \lambda_2)^2 \quad (4.11)$$

with some constant $C$. To compute the Jacobian not at the identity, we first notice that by rotation invariance, the measure factorizes. This means that its density with respect to the Lebesgue measure can be written of the form $f(\Lambda)g(U)$ with some functions $f$ and $g$, in fact $g(U)$ is constant (the marginal on the unitary part is the Haar measure). The function $f$ may be computed at any point, in particular at $U = I$, this was the calculation (4.11) yielding $f(\Lambda) = C(\lambda_1 - \lambda_2)^2$. This proves (4.4) for $N = 2$, modulo the case of multiple eigenvalue $\lambda_1 = \lambda_2$ where the parametrization of $U$ is even more redundant. But this set has zero measure, so it is negligible (see [8] for a precise argument). The formula (4.9) is the basis for the proof for general $N$ which we leave it to the readers. The detailed proof can be found in [8] or [37].

It is often useful to think of the measure (4.4) as a Gibbs measure on $N$ ”particles” or ”points” $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ of the form

$$\mu_N(d\lambda) = p_N(\lambda) \frac{e^{-\beta N \mathcal{H}(\lambda)}}{Z}, \quad \mathcal{H}(\lambda) := \frac{1}{2} \sum_{i=1}^{N} V(\lambda_i) - \frac{1}{N} \sum_{i<j} \log |\lambda_j - \lambda_i| \quad (4.12)$$

with the confining potential $V(\lambda)$ and logarithmic interaction. (This connection was exploited first in [46]).

We adopt the standard convention in random matrix theory that the Hamiltonian $\mathcal{H}$ expresses energy per
particle, in contrast to the standard statistical physics terminology where the "Hamiltonian" refers to the total energy. This explains the unusual $N$ factor in the exponent. Notice the emergence of the Vandermonde determinant in (4.3), which directly comes from integrating out the Haar measure and the symmetry type of the ensemble, appears through the exponent $\beta$. Only the "classical" $\beta = 1, 2$ or 4 cases correspond to matrix ensembles of the form (4.2), namely, to the real symmetric, complex Hermitian and quaternion self-dual matrices. We will not give the precise definition of the latter (see, e.g. Chapter 7 of [105] or [64]), just mention that this is the natural generalization of symmetric or Hermitian matrices to quaternion entries and they have real eigenvalues.

We remark that despite the convenience of the explicit formula (4.3) or (4.12) for the joint density, computing various statistics, such as correlation functions or even the density of a single eigenvalue, is a highly nontrivial task. For example, the density involves "only" integrating out all but one eigenvalue, but the measure is far from being a product, so these integrations cannot be performed directly when $N$ is large.

The measure (4.12) has a strong and long range interaction, while conventional methods of statistical physics are well suited for short range interactions. In fact, from this point of view $\beta$ can be any positive number and does not have to be restricted to the specific values $\beta = 1, 2, 4$. For other values of $\beta$ there is no invariant matrix ensemble behind the measure (4.12), but it is still a very interesting statistical mechanical system, called the log gas or $\beta$-ensemble. If the potential $V$ is quadratic, then (4.12) coincides with (4.4) and it is called the Gaussian $\beta$-ensemble. We will briefly discuss log gases in Section 18.3.

### 4.2 Universality of classical invariant ensembles via orthogonal polynomials

The classical values $\beta = 1, 2, 4$ in the ensemble (4.12) are specific not only because they originate from an invariant matrix ensemble (4.3). For these specific values an extra mathematical structure emerges, namely the orthogonal polynomials with respect to the weight function $e^{-NV(\lambda)}$ on the real line. Owing to this structure, much more is known about these ensembles than about log gases with general $\beta$. This approach was originally applied by Mehta and Gaudin [105, 106] to compute the gap distribution for the Gaussian case that involved the classical Hermite orthonormal polynomials. Dyson [47] computed the local correlation functions for a related ensemble (circular ensemble) that was extended to the standard Gaussian ensembles by Mehta [104]. Later a general method using orthogonal polynomials and the Riemann-Hilbert problem has been developed to tackle a very general class of invariant ensembles (see, e.g. [20, 37, 40–42, 71, 105, 111] and references therein).

For simplicity, to illustrate the connection, we will consider the Hermitian case $\beta = 2$ with a Gaussian potential $V(\lambda) = \lambda^2/2$ (which, by Lemma 4.1, is also a Wigner matrix ensemble, namely the GUE). To simplify the presentation further, for the purpose of this subsection only, we rescale the eigenvalues

$$x = \sqrt{N}\lambda,$$

which effectively removes the factor $N$ from the exponent in (4.3). (This simple scaling works only in the pure Gaussian case, and it is only a technical convenience to simplify formulas.)

After the rescaling and setting $\beta = 2$, the measure we will consider is given by a density which we denote by

$$\tilde{\rho}_N(x_1, x_2, \ldots, x_N) = \text{const.} \prod_{i<j} (x_i - x_j)^2 \prod_{j=1}^N e^{-1/2x_j^2}. \quad (4.14)$$

Let $P_k(x)$ be the $k$-th orthogonal polynomial on $\mathbb{R}$ with respect to the weight function $e^{-x^2/2}$ with leading coefficient 1. Let

$$\psi_k(x) := \frac{e^{-x^2/4}P_k(x)}{\|e^{-x^2/4}P_k\|_{L^2(\mathbb{R})}}$$

be the corresponding orthonormal function, i.e.,

$$\int_{\mathbb{R}} \psi_k(x)\psi_l(x)dx = \delta_{k,l}. \quad (4.15)$$
In the particular case of the Gaussian weight function, $P_k$ is given by the Hermite polynomials

$$P_k(x) = H_k(x) := (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2},$$

and

$$\psi_k(x) = \frac{P_k(x)}{(2\pi)^{1/4} (k!)^{1/2}} e^{-x^2/4},$$

(4.16)

but for the following discussion we will not need these explicit formulae, only the orthonormality relation (4.15).

The key observation is that, by simple properties of the Vandermonde determinant, we have

$$\Delta_N(x) = \prod_{1 \leq i < j \leq N} (x_j - x_i) = \det \left( x_i^{j-1} \right)_{i,j=1}^N = \det \left( P_{j-1}(x_i) \right)_{i,j=1}^N,$$

(4.17)

by exploiting that $P_j(x) = x^j + \ldots$ is a polynomial of degree $j$ with leading coefficient equal one. Define the kernel

$$K_N(x,y) := \sum_{k=0}^{N-1} \psi_k(x) \psi_k(y),$$

(4.18)

i.e., the projection kernel onto the subspace spanned by the first $K$ orthonormal functions. Then (4.17) immediately implies

$$\hat{p}_N(x_1, \ldots, x_N) = C_N \left[ \det \left( P_{j-1}(x_i) \right)_{i,j=1}^N \right]^2 \prod_{i=1}^N e^{-x_i^2/2}$$

(4.19)

$$= C_N' \left[ \det \left( \psi_{j-1}(x_i) \right)_{i,j=1}^N \right]^2 = C_N' \det \left( K_N(x_i, x_j) \right)_{i,j=1}^N,$$

where in the last step we used that the product of the matrices $\left( \psi_{j-1}(x_i) \right)_{i,j=1}^N$ and $\left( \psi_{j-1}(x_k) \right)_{j,k=1}^N$ is exactly $\left( K_N(x_i, x_k) \right)_{i,k=1}^N$ and we did not follow the precise constants for simplicity. The determinantal structure (4.19) for the joint density functions is remarkable: it allows one to describe a function of $N$ variables in terms of a kernel of two variables only. This structure greatly simplifies the analysis.

We now define the correlation functions that play a crucial role to describe universality.

**Definition 4.2.** Let $p_N(\lambda_1, \lambda_2, \ldots, \lambda_N)$ be the joint symmetrized probability distribution of the eigenvalues (without rescaling (4.13)). For any $n \geq 1$, the $n$-point correlation function is defined by

$$p_N^{(n)}(\lambda_1, \lambda_2, \ldots, \lambda_n) := \int_{\mathbb{R}^{N-n}} p_N(\lambda_1, \ldots, \lambda_n, \lambda_{n+1}, \ldots, \lambda_N) d\lambda_{n+1} \ldots d\lambda_N.$$  

(4.20)

We also define the rescaled correlation function $\hat{p}_N^{(n)}(x_1, x_2, \ldots, x_n)$ in a similar way.

**Remark.** In other sections of this book we usually label the eigenvalues in increasing order so that their probability density, denoted by $\hat{p}_N(\lambda)$, is defined on the set

$$\Xi^{(N)} := \{\lambda_1 < \lambda_2 < \ldots < \lambda_N\} \subset \mathbb{R}^N.$$

For the purpose of Definition 4.2, however, we dropped this restriction and we consider $p_N(\lambda_1, \lambda_2, \ldots, \lambda_N)$ to be a symmetric function of $N$ variables, $\lambda = (\lambda_1, \ldots, \lambda_N)$ on $\mathbb{R}^N$. The relation between the ordered and unordered densities is clearly $\hat{p}_N(\lambda) = N! p_N(\lambda) \cdot 1(\lambda \in \Xi^{(N)})$.

The significance of the correlation functions is that with their help one can compute the expectation value of any symmetrized observable. For example, for any test function $O$ of two variables we have, directly from the definition of the correlation functions, that

$$\frac{1}{N(N-1)} \sum_{i \neq j} O(\lambda_i, \lambda_j) = \int_{\mathbb{R}^2} O(\lambda_1, \lambda_2) p_N^{(2)}(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2,$$

(4.21)
where the expectation is w.r.t. the probability density \( p_N \) or in this case w.r.t. the original random matrix ensemble. Similar formula holds for observables of any number of variables.

To compute the correlation functions of a determinantal joint density (4.19), we start with the following prototype calculation for \( N = 3, n = 2 \)

\[
\int_\mathbb{R} \, dx_3 \begin{bmatrix} K_3(x_1, x_1) & K_3(x_1, x_2) & K_3(x_1, x_3) \\ K_3(x_2, x_1) & K_3(x_2, x_2) & K_3(x_2, x_3) \\ K_3(x_3, x_1) & K_3(x_3, x_2) & K_3(x_3, x_3) \end{bmatrix} \]

(4.22)

\[
= \int_\mathbb{R} \, dx_3 \begin{bmatrix} K_3(x_2, x_1) & K_3(x_2, x_2) \\ K_3(x_3, x_1) & K_3(x_3, x_2) \end{bmatrix} K_3(x_1, x_3) 
- \int_\mathbb{R} \, dx_3 \begin{bmatrix} K_3(x_1, x_1) & K_3(x_1, x_2) \\ K_3(x_3, x_1) & K_3(x_3, x_2) \end{bmatrix} K_3(x_2, x_3) 
+ \int_\mathbb{R} \, dx_3 \begin{bmatrix} K_3(x_1, x_1) & K_3(x_1, x_2) \\ K_3(x_2, x_1) & K_3(x_2, x_2) \end{bmatrix} K_3(x_3, x_3). \]

From the definition (4.18) and the orthonormality of the \( \psi \)'s we have the reproducing property

\[
\int_\mathbb{R} \, dy K_N(x, y) K_N(y, z) = K_N(x, z) \]

(4.23)

and the normalization

\[
\int_\mathbb{R} \, dx K_N(x, x) = N. \]

(4.24)

Thus (4.22) equals to

\[
\begin{bmatrix} K_3(x_2, x_1) & K_3(x_2, x_2) \\ K_3(x_1, x_1) & K_3(x_1, x_2) \end{bmatrix} = \begin{bmatrix} K_3(x_1, x_1) & K_3(x_1, x_2) \\ K_3(x_2, x_1) & K_3(x_2, x_2) \end{bmatrix} + 3 \begin{bmatrix} K_3(x_1, x_1) & K_3(x_1, x_2) \\ K_3(x_2, x_1) & K_3(x_2, x_2) \end{bmatrix} 
\]

(4.25)

\[
= \begin{bmatrix} K_3(x_1, x_1) & K_3(x_1, x_2) \\ K_3(x_2, x_1) & K_3(x_2, x_2) \end{bmatrix}. \]

(4.26)

It is easy to generalize this computation to get

\[
\tilde{p}_N^{(n)}(x_1, \ldots, x_n) = \frac{(N-n)!}{N!} \det[K_N(x_i, x_j)]_{i,j=1}^n, \]

(4.27)

i.e., the correlation functions continue to have a determinantal structure. Here the constant is obtained by the normalization condition that \( \tilde{p}_N^{(n)} \) is a probability density. Thus we have an explicit formula for the correlation functions in terms of the kernel \( K_N \). We note that this structure is very general and it is not restricted to Hermite polynomials, it only requires a system of orthogonal polynomials.

To understand the behavior of \( K_N \), first we recall a basic algebraic property of the orthogonal polynomials, the Christoffel–Darboux formula:

\[
K_N(x, y) = \sum_{j=0}^{N-1} \psi_j(x) \psi_j(y) = \sqrt{N} \left[ \psi_N(x) \psi_{N-1}(y) - \psi_N(y) \psi_{N-1}(x) \right]. \]

(4.28)

Since the Hermite polynomials and the orthonormal functions \( \psi_N \) differ only by an exponential factor (4.16), and these factors in \( \psi(x) \psi(y) \) on both side of (4.28) are cancelled, so (4.28) it is just a property of the Hermite polynomials. We now sketch a proof of this identity. Multiplying both side by \( (x-y) \), we need to prove that

\[
\sum_{j=0}^{N-1} \psi_j(x) \psi_j(y) (x-y) = \sqrt{N} \left[ \psi_N(x) \psi_{N-1}(y) - \psi_N(y) \psi_{N-1}(x) \right]. \]

(4.29)
The left side is a polynomial of degree $N$ in $x$ and degree $N$ in $y$, up to common exponential factors. The multiplication of $x$ or $y$ can be computed by the following identity

$$x\psi_j(x) = \sqrt{j+1}\psi_{j+1}(x) + \sqrt{j}\psi_{j-1}(x)$$  \hspace{1cm} (4.30)

that directly follows from the “three-term” relation for the Hermite polynomials. Collecting all the terms generated in this way, we obtain the right hand side of (4.29). Details can be found in Lemma 3.2.7 of [8].

It is well known that orthogonal polynomials of high degree have asymptotic behavior (Plancherel-Rotach asymptotics). For the Hermite orthonormal functions $\psi$ these formulas read as follows:

$$\psi_{2m}(x) = \frac{(-1)^m}{N^{1/4}√π} \cos(√Nx) + o(N^{-1/4}), \quad \psi_{2m+1}(x) = \frac{(-1)^m}{N^{1/4}√π} \sin(√Nx) + o(N^{-1/4}),$$  \hspace{1cm} (4.31)

as $N \to \infty$ for any $m$ such that $|2m - N| \leq C$. The approximation is uniform for $|x| \leq CN^{-1/2}$. We can thus compute that

$$K_N(x, y) \approx \frac{1}{π} \frac{\sin(√Nx) \cos(√Ny) - \sin(√Ny) \cos(√Nx)}{x - y} = \frac{\sin √N(x - y)}{π(x - y)},$$  \hspace{1cm} (4.32)

i.e., the celebrated sine-kernel emerged [47, 104].

To rewrite this formula into a canonical form, recall that we have done a rescaling (4.13) $λ = x/√N$ where $λ$ is the original variable. The two point function in the original variables was denoted by $p_N^{(2)}$ and the rescaled variables by $\tilde{p}_N^{(2)}$. The relation between $p_N^{(2)}$ and $\tilde{p}_N^{(2)}$ is determined by

$$p_N^{(2)}(\lambda_1, \lambda_2)d\lambda_1d\lambda_2 = \tilde{p}_N^{(2)}(x_1, x_2)dx_1dx_2,$$  \hspace{1cm} (4.33)

and thus we have

$$p_N^{(2)}(\lambda_1, \lambda_2) = N\tilde{p}_N^{(2)}(√N\lambda_1, √N\lambda_2).$$  \hspace{1cm} (4.34)

Now we introduce another rescaling of the eigenvalues that rescales the typical gap between them to order 1. We set

$$\lambda_j = \frac{a_j}{q_{sc}(0)N}, \quad q_{sc}(0) = \frac{1}{π},$$  \hspace{1cm} (4.35)

where $q_{sc}$ is the semicircle density, see (3.1). Using the expression (4.20) for correlation functions, we have, in terms of original variable, that

$$\frac{1}{[q_{sc}(0)]^2}p_N^{(2)}\left(\frac{a_1}{q_{sc}(0)N}, \frac{a_2}{q_{sc}(0)N}\right) = \det \left(\begin{array}{cc}
K_1 & K_2 \\
K_1 & K_2
\end{array}\right),$$  \hspace{1cm} (4.36)

where

$$K_{12} := \frac{1}{q_{sc}(0)} \frac{1}{√N - 1} K_N\left(\frac{a_1}{ρ_{sc}(0)√N}, \frac{a_2}{ρ_{sc}(0)√N}\right) \to S(a_1 - a_2), \quad S(x) := \frac{\sin x}{πx},$$  \hspace{1cm} (4.37)

where we used (4.32). Due to the rescaling, this calculation reveals the correlation functions around $E = 0$. The general formula for any fixed energy $E$ in the bulk, i.e., $|E| < 2$, can be obtained similarly and it is given by

$$\frac{1}{[q_{sc}(E)]^n}p_N^{(n)}\left(E + \frac{α_1}{Nq_{sc}(E)}, E + \frac{α_2}{Nq_{sc}(E)}, \ldots, E + \frac{α_n}{Nq_{sc}(E)}\right) = q_{GOE}^{(n)}(\alpha) := \det \left(S(α_i - α_j)\right)_{i,j=1}^n$$  \hspace{1cm} (4.38)

as weak convergence of functions in the variables $α = (α_1, \ldots, α_n)$. Note that the limit is universal in a sense that it is independent of the energy $E$. Formula (4.38) holds for the GUE case. The corresponding expression for GOE is more involved [8,105]

$$q_{GOE}^{(n)}(\alpha) := \det \left(K(α_i - α_j)\right)_{i,j=1}^n, \quad K(x) := \left(-\frac{1}{2} sgn(x) + \int_0^x S(t)dt \right) S(x).$$  \hspace{1cm} (4.39)
Here the determinant is understood as the trace of the quaternion determinant after the canonical correspondence between quaternions $a \cdot 1 + b \cdot i + c \cdot j + d \cdot k$, $a, b, c, d \in \mathbb{C}$, and $2 \times 2$ complex matrices given by

$$1 \leftrightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad i \leftrightarrow \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \quad j \leftrightarrow \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad k \leftrightarrow \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$ 

The main technical input is the refined asymptotic formulae (4.31) for orthogonal polynomials. In case of the classical orthogonal polynomials (appearing in the standard Gaussian Wigner and Wishart ensembles) they are usually obtained by a Laplace asymptotics from their integral representation. For a general potential $V$ the corresponding analysis is quite involved and depends on the regularity properties of $V$. One successful approach was initiated by Fokas, Its and Kitaev [71] and by P. Deift and collaborators via the Riemann-Hilbert method, see [37] and references therein. An alternative method was presented in [98,101] using more direct methods from orthogonal polynomials.

### 4.2.1 Edge universality: the Airy kernel

Near the spectral edges, i.e., for energy $E = \pm 2$, a different scaling has to be used. Recall the formula

$$K_N(x, y) = \sqrt{N} \left[ \frac{\psi_N(x)\psi_{N-1}(y) - \psi_N(y)\psi_{N-1}(x)}{x - y} \right]$$

in the rescaled variables $x, y$. We will need the following identity of the derivatives of the Hermite functions

$$\psi_N'(x) = -\frac{x}{2} \psi_N(x) + \sqrt{N} \psi_{N-1}(x). \quad (4.40)$$

Thus we can rewrite

$$K_N(x, y) = \left[ \frac{\psi_N(x)\psi_N'(y) - \psi_N(y)\psi_N'(x)}{x - y} - \frac{1}{2} \psi_N(x)\psi_N(y) \right]. \quad (4.41)$$

Define a new rescaled function

$$\Psi_N(u) := N^{1/12} \psi_N \left( 2\sqrt{N} + \frac{u}{N^{1/6}} \right). \quad (4.42)$$

The Plancherel-Rotach edge asymptotics for $\Psi$ asserts that

$$\lim_{N \to \infty} |\Psi_N(z) - \text{Ai}(z)| = 0 \quad (4.43)$$

in any compact domain in $\mathbb{C}$, where $\text{Ai}(x)$ is the Airy function, i.e.,

$$\text{Ai}(x) = \frac{1}{\pi} \int_0^\infty \cos \left( \frac{1}{3} t^3 + xt \right) dt.$$ 

It is well-known that the Airy function is the solution to the second order differential equation, $y'' - xy = 0$, with vanishing boundary condition at $x = \infty$.

We now define the Airy kernel by

$$A(u, v) := \frac{\text{Ai}(u)\text{Ai}'(v) - \text{Ai}'(u)\text{Ai}(v)}{u - v}.$$ 

Under the edge scaling (4.42), we have

$$N^{-1/6} K_N \left( 2\sqrt{N} + \frac{u}{N^{1/6}}, 2\sqrt{N} + \frac{v}{N^{1/6}} \right) \to A(u, v). \quad (4.44)$$

In these variables, we have, by (4.34), that

$$p_N^{(2)} \left( 2 + \frac{\alpha_1}{N^{2/3}}, 2 + \frac{\alpha_2}{N^{2/3}} \right) = N^{-1/6} p_N^{(2)} \left( 2\sqrt{N} + \frac{\alpha_1}{N^{1/6}}, 2\sqrt{N} + \frac{\alpha_1}{N^{1/6}} \right). \quad (4.45)$$

22
By using (4.20), we can continue with
\[
p_N^{(2)} \left( 2 + \frac{\alpha_1}{\sqrt[3]{N^2}}, 2 + \frac{\alpha_2}{\sqrt[3]{N^2}} \right) = N \frac{1}{N(N-1)} \det \left[ K_N \left( 2\sqrt{N} + \frac{\alpha_i}{N^{1/3}}, 2\sqrt{N} + \frac{\alpha_j}{N^{1/3}} \right) \right]_{i,j=1}^2 \\
\simeq N^{-2/3} \det \left[ N^{-1/6} K_N \left( 2\sqrt{N} + \frac{\alpha_i}{N^{1/3}}, 2\sqrt{N} + \frac{\alpha_j}{N^{1/3}} \right) \right]_{i,j=1}^2,
\]
and similar formulas hold for any \( k \)-point correlation functions. Using the limiting statement (4.44), in terms of the original variables, we obtain
\[
N^{k/3}p_N^{(k)} \left( 2 + \frac{\alpha_1}{N^{2/3}}, 2 + \frac{\alpha_2}{N^{2/3}}, \ldots, 2 + \frac{\alpha_k}{N^{2/3}} \right) \rightarrow \det \left( A(\alpha_i, \alpha_j) \right)_{i,j=1}^k \quad (4.45)
\]
in a weak sense. In particular, the last formula with \( k = 2 \) implies, for any smooth test function \( O \) with compact support, that
\[
\sum_{j \neq k} E\left[ O(N^{2/3}(\lambda_j - 2)), N^{2/3}(\lambda_k - 2) \right] = N(N-1)N^{-4/3} \int_{\mathbb{R}^2} d\alpha_1 d\alpha_2 O(\alpha_1, \alpha_2) p_N^{(2)} \left( 2 + \frac{\alpha_1}{N^{2/3}}, 2 + \frac{\alpha_2}{N^{2/3}} \right) \\
\rightarrow \int_{\mathbb{R}^2} d\alpha_1 d\alpha_2 O(\alpha_1, \alpha_2) \det \left( A(\alpha_i, \alpha_j) \right)_{i,j=1}^2. \quad (4.46)
\]
Similar statement holds at the lower spectral edge, \( E = -2 \).
5 Universality for generalized Wigner matrices

5.1 Different notions of universality

The universality of eigenvalue statistics can be considered via several different notions of convergence which yield somewhat different concepts of universality. The local statistics can either be expressed in terms of local correlation functions rescaled around some energy $E$ or one may ask for the gap statistics for a gap $\lambda_{j+1} - \lambda_j$ with a given (typically $N$-dependent) label $j$. These will be called fixed energy and fixed gap universality and these two concepts do not coincide. To see this, notice that since eigenvalues fluctuate on a scale much larger than the typical eigenvalue spacing, the label $j$ of the eigenvalue $\lambda_j$ closest to a fixed energy $E$ is not a deterministic function of $E$. Furthermore, one may look for cumulative statistics of gaps averaged over a mesoscopic scale, i.e., both above concepts have a natural averaged version. We now define these four concepts precisely.

The correlation functions and the gaps need to be rescaled by the limiting local density, $\rho(E)$, to get a universal limit. In case of generalized Wigner matrices we have $\rho(E) = \rho_{sc}(E)$, but the definitions below hold in more general setup as well. We also remark that all results in this book hold for both real symmetric matrices or complex Hermitian matrices. For simplicity of notations, we formulate all concepts and later state all results in terms of real symmetric matrices.

We recall the notation $[A,B] := \mathbb{Z} \cap [A,B]$ for the set of integers between two real numbers $A < B$.

(i) **Fixed energy universality (in the bulk):** For any $n \geq 1$, $F : \mathbb{R}^n \to \mathbb{R}$ smooth and compactly supported function and for any $\mu > 0$, we have, uniformly in $E \in [-2 + \kappa, 2 - \kappa],
\lim_{N \to \infty} \frac{1}{\rho(E)^n} \int_{\mathbb{R}^n} d\alpha F(\alpha) p_N^{(n)}(E + \frac{\alpha}{N \rho(E)}) = \int_{\mathbb{R}^n} d\alpha F(\alpha) q_{GOE}^{(n)}(\alpha), \quad (5.1)

where $\alpha = (\alpha_1, \ldots, \alpha_n)$ . Here $p_N^{(n)}$ is the $n$-point function of the matrix ensemble and $q_{GOE}^{(n)}$ is the limiting $n$-point function of the GOE defined in (4.39). To shorten the argument of $p_N^{(n)}$, we used the convention that $E + \alpha = (E + \alpha_1, \ldots, E + \alpha_n)$ for any $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^N$.

(ii) **Averaged energy universality (in the bulk, on scale $N^{-1+\xi}$):** For any $n \geq 1$, $F : \mathbb{R}^n \to \mathbb{R}$ smooth and compactly supported function and for some $0 < \xi < 1$ and for any $\kappa > 0$, we have, uniformly in $E \in [-2 + \kappa, 2 - \kappa],
\lim_{N \to \infty} \frac{1}{\rho(E)^n} \int_{E-b}^{E+b} \frac{dx}{2b} \int_{\mathbb{R}^n} d\alpha F(\alpha) p_N^{(n)}(x + \frac{\alpha}{N \rho(E)}) \frac{d\alpha}{2} = \int_{\mathbb{R}^n} d\alpha F(\alpha) q_{GOE}^{(n)}(\alpha), \quad (5.2)

where $b = b_N := N^{-1+\xi}$.

(iii) **Fixed gap universality (in the bulk):** Fix any positive number $0 < \alpha < 1$ and an integer $n$. For any smooth compactly supported function $G : \mathbb{R}^n \to \mathbb{R}$ and for any $k, m \in [\alpha N, (1 - \alpha)N]$ we have
\[
\lim_{N \to \infty} \left| \mathbb{E}^{\mu_N} G((N \rho(\lambda_k))(\lambda_k - \lambda_{k+1}), \ldots, (N \rho(\lambda_k))(\lambda_k - \lambda_{k+n})) \right. \\
- \mathbb{E}^{\langle GOE \rangle} G((N \rho(\lambda_m))(\lambda_m - \lambda_{m+1}), \ldots, (N \rho(\lambda_m))(\lambda_m - \lambda_{m+n})) \left. \right| = 0, \quad (5.3)

where $\mu_N$ denotes the law of the random matrix ensemble under consideration.

(iv) **Averaged gap universality (in the bulk, on scale $N^{-1+\xi}$):** With the same notations as in (iii) and $\ell = N^{\xi}$ with $0 < \xi < 1$, we have
\[
\lim_{N \to \infty} \left| \frac{1}{2\ell + 1} \sum_{j=k-\ell}^{k+\ell} \mathbb{E}^{\mu_N} G((N \rho(\lambda_j))(\lambda_j - \lambda_{j+1}), \ldots, (N \rho(\lambda_j))(\lambda_j - \lambda_{j+n})) \right. \\
- \mathbb{E}^{\langle GOE \rangle} G((N \rho(\lambda_m))(\lambda_m - \lambda_{m+1}), \ldots, (N \rho(\lambda_m))(\lambda_m - \lambda_{m+n})) \left. \right| = 0. \quad (5.4)

Note that in the bulk $\ell = N^\xi$ consecutive eigenvalues range over a scale $N^{-1+\xi}$ in the spectrum, hence the name.

Although we have formulated the universality notions in terms of large $N$ limits, all limit statements in this book have effective error bounds of the form $N^{-c}$ for some $c > 0$. For the four notions of universality stated here, the fixed energy (fixed gap, resp.) universality obviously implies averaged energy (averaged gap, resp.) universality. However, the fixed gap universality and the fixed energy universality are not logical consequences of each other. On the other hand, under suitable conditions, the averaged energy universality and averaged gap universality are equivalent. In Section 14 we will prove that the latter implies the former, this is the direction that we actually use in the proof. The opposite direction goes along similar arguments and will be omitted.

In this book, we will focus on establishing the averaged energy universality. From this one, the average gap universality follows by the equivalence just mentioned. We now state precisely our representative universality theorem that will be proven in this book. It asserts average energy universality (in the bulk) for generalized Wigner matrices, where the scale of the energy average is reduced to $N^{-1+\xi}$ for arbitrary $\xi > 0$.

For convenience we assume that the normalized matrix entries

$$\zeta_{ij} := \sqrt{N} h_{ij}\quad (5.5)$$

have a polynomial decay of arbitrary high degree, i.e., for all $p \in \mathbb{N}$ there is a constant $\mu_p$ such that

$$\mathbb{E}|\zeta_{ij}|^p \leq \mu_p \quad (5.6)$$

for all $N$, $i$, and $j$. Recall from (2.6) that $s_{ij} \approx N^{-1}$, so this condition is equivalent to (2.8). We make this assumption in order to streamline notation, but in fact, our results hold, with the same proof, provided (5.6) is valid for some large but fixed $p$. In fact, even $p = 4 + \varepsilon$ is sufficient, see Section 18. On the other hand, if we strengthen the decay condition to uniform subexponential decay (2.7), then certain estimates concerning the local semicircle law (e.g., Theorem 6.7) become stronger, although we will not them in this book (see [70]).

**Theorem 5.1.** Let $H$ be an $N \times N$ generalized real symmetric Wigner matrix. Suppose that the rescaled matrix elements $\sqrt{N} h_{ij}$ satisfy the decay condition (5.6). Then averaged energy universality holds in the sense of (5.1) on scale $N^{-1+\xi}$ for any $0 < \xi < 1$.

The rest of this book is devoted to a proof of Theorem 5.1 and related questions on eigenvectors. The proof of this theorem will be based on the following three step strategy.

### 5.2 The three-step strategy

**Step 1. Local semicircle law and delocalization of eigenvectors:** It states that the density of eigenvalues is given by the semicircle law not only as a weak limit on macroscopic scales (3.1), but also in a high probability sense with an effective convergence speed and down to short scales containing only $N^\xi$ eigenvalues for all $\xi > 0$. This will imply the rigidity of eigenvalues, i.e., that the eigenvalues are near their classical locations in the sense to be made clear in Theorem 11.5. We also obtain precise estimates on the matrix elements of the Green function which in particular imply complete delocalization of eigenvectors.

**Step 2. Universality for Gaussian divisible ensembles:** The Gaussian divisible ensembles are matrices of the form $H_t = e^{-t/2}H_0 + \sqrt{1-e^{-t}}H^G$ where $t > 0$ is a parameter, $H_0$ is a (generalized) Wigner matrix and $H^G$ is an independent GOE matrix. The parametrization of $H_t$ is chosen so that $H_t$ can be obtained by an Ornstein-Uhlenbeck process starting from $H_0$. More precisely, consider the following matrix Ornstein-Uhlenbeck process

$$dH_t = \frac{1}{\sqrt{N}}dB_t - \frac{1}{2}H_t dt \quad (5.7)$$

25
with initial data $H_0$, where $B_t = \{b_{ij,t}\}_{i,j=1}^{N}$ is a symmetric $N \times N$ matrix such that its matrix elements $b_{ij,t}$ for $i < j$ and $b_{ii,t}/\sqrt{2}$ are independent standard Brownian motions starting from zero. Then $H_t$ and $e^{-t/2}H_0 + \sqrt{1 - e^{-t}}H^G$ have the same distribution.

The aim of Step 2 is to prove the bulk universality of $H_t$ for $t = N^{-\tau}$ for the entire range of $0 < \tau < 1$. This is connected to the local ergodicity of the Dyson Brownian motion which we now define.

**Definition 5.2.** Given a real parameter $\beta \geq 1$, consider the following system of stochastic differential equations (SDE)

$$
d\lambda_i = \frac{\sqrt{2}}{\sqrt{\beta N}} dB_i + \left( -\frac{\lambda_i}{2} + \frac{1}{N} \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \right) dt, \quad i \in [1, N],
$$

where $(B_i)$ is a collection of real-valued, independent standard Brownian motions. The solution of this SDE is called the Dyson Brownian Motion (DBM).

In a seminal paper [45], Dyson observed that the eigenvalue flow of the matrix OU process is exactly the DBM with $\beta = 1, 2$ corresponding to real symmetric or complex Hermitian ensembles. Furthermore, the invariant measure of the DBM is given by the Gaussian $\beta$-ensemble defined in (4.4). Dyson further conjectured that the time to “local equilibrium” for DBM is of order $1/N$ while the time to global equilibrium is of order one. It should be noted that there is no standard notion for the “local equilibrium”; we will instead take a practical point of view to interpret Dyson’s conjecture as that the local statistics of the DBM at any time $t \gg N^{-1}$ satisfy the universality defined in earlier in this section. In other words, Dyson’s conjecture is exactly that the local statistics of $H_t$ are universal for $t = N^{-\tau}$ for any $0 < \tau < 1$.

**Step 3. Approximation by a Gaussian divisible ensemble:** It is a simple density argument in the space of matrix ensembles which shows that for any probability distribution of the matrix elements there exists a Gaussian divisible distribution with a small Gaussian component, as in Step 2, such that the two associated Wigner ensembles have asymptotically identical local eigenvalue statistics. The general result to compare any two matrix ensembles with matching moments will be given in Theorem 16.1. Alternatively, to follow the evolution of the Green function under the OU flow, we can use the following continuity of matrix OU process:

**Step 3a. Continuity of eigenvalues under matrix OU process.** In Theorem 15.2 will show that the changes of the local statistics in the bulk under the flow (5.7) up to time scales $t \ll N^{-1/2}$ are negligible, see Lemma 15.4. This clearly can be used in combination with Step 2 to complete the proof of Theorem 5.1.

The three step strategy outlined here is very general and it has been applied to many different models as we will explain in Section 18. It can also be extended to the edges of the spectrum, yielding the universality at the spectral edges. This will be reviewed in Section 17.
6 Local semicircle law for universal Wigner matrices

6.1 Setup

We recall the definition of the universal Wigner matrices (Definition 2.1), in particular, the matrix elements may have different distributions but independence (up to symmetry) is always assumed. The fundamental data of this model is the $N \times N$ matrix of variances $S = (s_{ij})$, where

$$s_{ij} := \mathbb{E} |h_{ij}|^2,$$

and we assume that $S$ is (doubly) stochastic:

$$\sum_j s_{ij} = 1 \quad (6.1)$$

for all $i$. We will assume the polynomial decay analogous to (5.6)

$$\mathbb{E} |h_{ij}|^p \leq \mu_p [s_{ij}]^{p/2}, \quad (6.2)$$

where $\mu_p$ depends only on $p$ and is uniform in $i, j$ and $N$.

We introduce the parameter $M := [\max_{i,j} s_{ij}]^{-1}$ that expresses the maximal size of $s_{ij}$:

$$s_{ij} \leq M^{-1} \quad (6.3)$$

for all $i$ and $j$. We regard $N$ as the fundamental parameter and $M = M_N$ as a function of $N$. In this section we do not assume lower bound on $s_{ij}$. Notice that for generalized Wigner matrices $M$ is comparable with $N$ and one may use $N$ everywhere instead of $M$ in all error bounds in this section. However, we wish to keep $M$ as a separate parameter and assume only that

$$N^\delta \leq M \leq N \quad (6.4)$$

for some fixed $\delta > 0$. For standard Wigner matrices, $h_{ij}$ are identically distributed, hence $s_{ij} = \frac{1}{N}$ and $M = N$. Another motivating example where $M$ may substantially differ from $N$ is the random band matrices that play a key role interpolating between Wigner matrices and random Schrödinger operators (see Section 18.7).

Example 6.1. Random band matrices are characterized by translation invariant variances of the form

$$s_{ij} = \frac{1}{W f \left( \frac{|i-j|}{N} \right)}, \quad (6.5)$$

where $f$ is a smooth, symmetric probability density on $\mathbb{R}$, $W$ is a large parameter, called the band width, and $|i-j|_N$ denotes the periodic distance on the discrete torus $\mathbb{T}$ of length $N$. In this case $M$ is comparable with $W$ which is typically much smaller than $N$. The generalization in higher spatial dimensions is straightforward, in this case the rows and columns of $H$ are labelled by a discrete $d$ dimensional torus $\mathbb{T}^d$ of length $L$ with $N = L^d$.

Recall from (3.1) that the global eigenvalue density of $H$ in the $N \to \infty$ limit follows the celebrated Wigner semicircle law,

$$\rho_{sc}(x) := \frac{1}{2\pi} \sqrt{(4-x^2)_+}, \quad (6.6)$$

and its Stieltjes transform with the spectral parameter $z = E + i\eta$ is defined by

$$m_{sc}(z) := \int_{\mathbb{R}} \frac{\rho_{sc}(x)}{x-z} \, dx. \quad (6.7)$$

The spectral parameter $z$ will sometimes be omitted from the notation. The two endpoints $\pm 2$ of the support of $\rho_{sc}$ are called the spectral edges.
It is well-known (see Section 3.3) that the Stieltjes transform $m_{sc}$ is the unique solution of the quadratic equation

$$m(z) + \frac{1}{m(z)} + z = 0$$

with $\text{Im} \ m(z) > 0$ for $\text{Im} \ z > 0$. Thus we have

$$m_{sc}(z) = \frac{-z + \sqrt{z^2 - 4}}{2},$$

where the square root is chosen so that $\text{Im} m_{sc}(z) > 0$ for $\text{Im} z > 0$. In particular, $\sqrt{z^2 - 4} \approx z$ for $z$ large so that it cancels the $-z$ term in the above formula. Under this convention, we collect basic bounds on $m_{sc}$ in the following lemma.

**Lemma 6.2.** We have for all $z = E + i\eta$ with $\eta > 0$ that

$$|m_{sc}(z)| = |m_{sc}(z) + z|^{-1} \leq 1.$$  \hspace{1cm} (6.10)

Furthermore, there is a constant $c > 0$ such that for $E \in [-10, 10]$ and $\eta \in (0, 10]$ we have

$$c \leq |m_{sc}(z)| \leq 1 - c\eta,$$  \hspace{1cm} (6.11)

$$|1 - m_{sc}^2(z)| \approx \sqrt{\kappa + \eta},$$  \hspace{1cm} (6.12)

where $\kappa := |E| - 2$ denotes the distance of $E$ to the spectral edges.

**Proof.** The proof is an elementary exercise using (6.9). \hfill \Box

We define the **Green function** or the **resolvent** of $H$ through

$$G(z) := (H - z)^{-1},$$

and denote its entries by $G_{ij}(z)$. We recall from (3.9) that the Stieltjes transform of the empirical spectral measure

$$\varrho(dx) = \varrho_N(dx) := \frac{1}{N} \sum_\alpha \delta(\lambda_\alpha - x)dx$$

for the eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ of $H$ is

$$m(z) = m_N(z) := \int_\mathbb{R} \frac{\varrho_N(dx)}{x - z} = \frac{1}{N} \text{Tr}G(z) = \frac{1}{N} \sum_\alpha \frac{1}{z - \lambda_\alpha}. \hspace{1cm} (6.14)$$

We remark that every quantity related to the random matrix $H$, such as the eigenvalues, Green function, empirical density of states and its Stieltjes transform, all depend on $N$, but this dependence will often be omitted in the notation for brevity. In some formulas, especially in statements of the main results we will put back the $N$ dependence to stress its presence.

Since

$$\frac{1}{\pi} \text{Im} \frac{1}{z - \lambda_\alpha} = \theta_\eta(E - \lambda_\alpha), \quad \text{with} \quad \theta_\eta(x) := \frac{1}{\pi} \frac{\eta}{x^2 + \eta^2}$$

is an approximation to the identity (i.e., delta function) at the scale $\eta = \text{Im} z$, we have $\pi^{-1} \text{Im} m_N(z) = \varrho_N \ast \theta_\eta(z)$, i.e., the imaginary part of $m_N(z)$ is the density of the eigenvalues “at the scale $\eta$”. Thus the convergence of the Stieltjes transform $m_N(z)$ to $m_{sc}(z)$ as $N \to \infty$ will show that the empirical local density of the eigenvalues around the energy $E$ in a window of size $\eta$ converges to the semicircle law $\varrho_{sc}(E)$. Therefore the key task is to control $m_N(z)$ for small $\eta$. 

28
6.2 Spectral information on $S$

We will show that the diagonal matrix elements of the resolvent $G_{ii}(z)$ satisfy a system of self-consistent vector equation of the form

$$\frac{1}{G_{ii}(z)} \approx z + \sum_{j=1}^{N} s_{ij} G_{jj}(z) \quad (6.15)$$

with very high probability. This equation will be viewed as a small perturbation of the deterministic equation

$$\frac{1}{m_i(z)} = z + \sum_{j=1}^{N} s_{ij} m_j(z) \quad (6.16)$$

which, under the side condition that $\text{Im } m_i > 0$, has a unique solution, namely $m_i(z) = m_{sc}(z)$ for every $i$ (see (6.8)). Here the stochasticity condition (6.1) is essential. For the stability analysis of (6.16), the invertibility of the operator $1 - m_{sc}^2(z)S$ plays a key role.

Therefore an important parameter of the model is

$$\Gamma(z) := \left\| \frac{1}{1 - m_{sc}^2(z)S} \right\|_{\infty \to \infty}, \quad \text{Im } z > 0. \quad (6.17)$$

Note that $S$, being a stochastic matrix, satisfies $-1 \leq S \leq 1$, and 1 is an eigenvalue with eigenvector $\mathbf{e} = \frac{1}{N}(1, 1, \ldots, 1)$, $S\mathbf{e} = \mathbf{e}$. For convenience we assume that 1 is a simple eigenvalue of $S$ (which holds if $S$ is irreducible and aperiodic). Another important parameter is

$$\tilde{\Gamma}(z) := \left\| \frac{1}{1 - m_{sc}^2(z)S} \right\|_{\infty \to \infty} \left\| \mathbf{e} \right\|_{\infty \to \infty}, \quad (6.18)$$

i.e., the norm of $(1 - m_{sc}^2S)^{-1}$ restricted to the subspace orthogonal to the constants. Recalling that $-1 \leq S \leq 1$ and using the upper bound $|m_{sc}| \leq 1$ from (6.11), we find that there is a constant $c > 0$ such that

$$c \leq \tilde{\Gamma} \leq \Gamma. \quad (6.19)$$

Since $|m_{sc}(z)| \leq 1 - c\eta$, we can expand $(1 - m_{sc}^2S)^{-1}$ into a geometric series. Using that $\|S\|_{\ell_\infty \to \ell_\infty} \leq 1$ from (6.1), we obtain the trivial upper bound

$$\Gamma \leq C\eta^{-1}. \quad (6.20)$$

We remark that we also have the following easy lower bound on $\Gamma$:

$$\Gamma \geq |1 - m_{sc}^2|^{-1} \geq \frac{1}{2}. \quad (6.21)$$

This can be seen by applying the matrix $(1 - m_{sc}^2S)^{-1}$ to the constant vector and using the definition of $\Gamma$.

For standard Wigner matrices, $s_{ij} = \frac{1}{N}$, and for bounded energies $|E| \leq C$, we easily obtain that

$$\Gamma(z) = \frac{1}{|1 - m_{sc}^2(z)|} \leq \frac{1}{\sqrt{\kappa_E + \eta}}, \quad \tilde{\Gamma}(z) = 1, \quad (6.22)$$

where

$$\kappa_E := \min\{|E - 2|, |E + 2|\} \quad (6.23)$$

is the distance of $E = \text{Re } z$ from the spectral edges. To see this, we note that in this case $S = P_e$, the projection operator onto the direction $\mathbf{e}$. Hence the operator $[1 - m_{sc}^2(z)S]^{-1}$ can be computed explicitly. The comparison relation in (6.22) follows from (6.12). In the general case, we have the same relations up to a constant factor:
Lemma 6.3. For generalized Wigner matrices (Definition 2.1) and for bounded range of energies, say for any $|E| \leq 10$, we have
\[ \frac{c}{|1 - m^2(z)|} \leq \Gamma(z) \leq \frac{C}{|1 - m^2(z)|} \asymp \frac{C}{\sqrt{\kappa E + \eta}}, \quad c \leq \tilde{\Gamma}(z) \leq C, \quad (6.24) \]
where the constant $C$ depends on $C_{\inf}$ and $C_{\sup}$ in (2.6) and the constant $c$ in the lower bound is universal.

The proof will be given in Section 6.5.

6.3 Stochastic domination

The following definition introduces a notion of a high-probability bound that is suited for our purposes. It first appeared in [55] and it relieves us from the burden to keep track of exceptional sets of small probability where some bound does not hold.

Definition 6.4 (Stochastic domination). Let 
\[ X = (X^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}), \quad Y = (Y^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}) \]
be two families of nonnegative random variables, where $U^{(N)}$ is a possibly $N$-dependent parameter set. We say that $X$ is stochastically dominated by $Y$, uniformly in $u$, if for all (small) $\varepsilon > 0$ and (large) $D > 0$ we have
\[ \sup_{u \in U^{(N)}} \mathbb{P}\left[X^{(N)}(u) > N^{-\varepsilon}Y^{(N)}(u)\right] \leq N^{-D} \]
for large enough $N \geq N_0(\varepsilon, D)$. Unless stated otherwise, throughout this paper the stochastic domination will always be uniform in all parameters apart from the parameter $\delta$ in (6.4) and the sequence of constants $\mu_p$ in (5.6); thus, $N_0(\varepsilon, D)$ also depends on $\delta$ and $\mu_p$. If $X$ is stochastically dominated by $Y$, uniformly in $u$, we use the notation $X \prec Y$. Moreover, if for some complex family $X$ we have $|X| \prec Y$ we also write $X = O_{\prec}(Y)$.

The following proposition collects some basic properties of the stochastic domination, the proofs are left as an exercise.

Proposition 6.5. The relation $\prec$ satisfies the following properties:

i) $\prec$ is transitive: $X \prec Y$ and $Y \prec Z$ imply $X \prec Z$;

ii) $\prec$ satisfies the familiar arithmetic rules of order relations, i.e., if $X_1 \prec Y_1$ and $X_2 \prec Y_2$ then $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1 X_2 \prec Y_1 Y_2$;

iii) Moreover, the following cancellation property holds
\[ \text{if } X \prec Y + N^{-\varepsilon}X \text{ for some } \varepsilon > 0, \text{ then } X \prec Y; \quad (6.25) \]

iv) Furthermore, if $X \prec Y$, $EY \geq N^{-C}$ and $|X| \leq N^{C}$ almost surely with some fixed exponent $C$, then for any $\varepsilon > 0$ and sufficiently large $N \geq N_0(\varepsilon)$ we have
\[ EY \leq N^{\varepsilon}EY. \quad (6.26) \]

Later in Lemma 10.1 the relation (6.26) will be extended to partial expectations.

We now define appropriate subsets of the spectral parameter $z$.

Definition 6.6 (Spectral domain). We call an $N$-dependent family 
\[ D \equiv D^{(N)} \subset \{ z : |E| \leq 10, M^{-1} \leq \eta \leq 10 \} \]
a spectral domain. (Recall that $M \equiv M_N$ depends on $N$.)
We always consider families \( X^{(N)}(u) = X_i^{(N)}(z) \) indexed by \( u = (z, i) \), where \( z \) takes on values in some spectral domain \( D \), and \( i \) takes on values in some finite (possibly \( N \)-dependent or empty) index set. The stochastic domination \( X \prec Y \) of such families will always be uniform in \( z \) and \( i \), and we usually do not state this explicitly. Usually, which spectral domain \( D \) is meant will be clear from the context, in which case we shall not mention it explicitly.

For example, using Chebyshev’s inequality and (6.2) one easily finds that

\[
h_{ij} \prec (s_{ij})^{1/2} \prec M^{1/2},
\]

uniformly in \( i \) and \( j \), so that we may also write \( h_{ij} = O_{\prec}((s_{ij})^{1/2}) \). The definition of \( \prec \) with the polynomial factors \( N^{-\varepsilon} \) and \( N^{-D} \) are tailored for the assumption (6.2). We remark that if the analogous subexponential decay (2.7) is assumed then a stronger form of stochastic domination can be introduced but we will not pursue this direction here.

### 6.4 Statement of the local semicircle law

The local semicircle law is very sensitive to the closeness of \( \eta = \text{Im} \, z \) to the real axis; when \( \eta \) is too small, the resolvent and its trace becomes strongly fluctuating. So our results will hold only above a certain threshold for \( \eta \). We now define the lower threshold for \( \eta \) that depends on the energy \( E \in [-10, 10] \)

\[
\tilde{\eta}_E := \min \left\{ \eta : \frac{1}{M\xi} \leq \min \left\{ \frac{M^{-\gamma}}{\Gamma(E + i\xi)^3}, \frac{M^{-2\gamma}}{\Gamma(E + i\xi)^4 \text{Im} m_{sc}(E + i\xi)} \right\} \right. \text{ holds for all } \xi \geq \eta \right\}. \tag{6.28}
\]

Although this expression looks complicated, we shall see that it comes out naturally in the analysis of the self-consistent equations for the Green functions. Here \( \gamma > 0 \) is a parameter that can be chosen arbitrarily small; for all practical purposes the reader can neglect it. For generalized Wigner matrices, \( M \approx N \), from (6.24) we have

\[
\tilde{\eta}_E \leq CN^{-1+2\gamma}, \tag{6.29}
\]

i.e., we will get the local semicircle law on the smallest possible scale \( \eta \gg N^{-1} \), modulo an \( M\gamma \) correction with an arbitrary small exponent. We remark that if we assume subexponential decay (2.7) instead of the polynomial decay (6.2), then the small \( M\gamma \) correction can be replaced with a \((\log M)^C \) factor.

Finally we define our fundamental control parameter

\[
\Pi(z) := \sqrt{\frac{\text{Im} m_{sc}(z)}{M\eta}} + \frac{1}{M\eta}. \tag{6.30}
\]

We can now state the main result of this section, which in this full generality first appeared in [55]. Previous results that have cumulatively led to this general formulation will be summarized at the end of the section.

**Theorem 6.7** (Local semicircle law [55]). Consider a universal Wigner matrix satisfying the polynomial decay condition (6.2) and (6.3). Then, uniformly in the energy \(|E| \leq 10\) and \( \eta \in [\tilde{\eta}_E, 10] \), we have the bounds

\[
\max_{i,j} |G_{ij}(z) - \delta_{ij} m_{sc}(z)| \prec \Pi(z) = \sqrt{\frac{\text{Im} m_{sc}(z)}{M\eta}} + \frac{1}{M\eta}, \quad z = E + i\eta, \tag{6.31}
\]

as well as

\[
|m_N(z) - m_{sc}(z)| \prec \frac{1}{M\eta}. \tag{6.32}
\]

Moreover, outside of the spectrum we have the stronger estimate

\[
|m_N(z) - m(z)| \prec \frac{1}{M(\kappa E + \eta)} + \frac{1}{(M\eta)^2 \sqrt{\kappa E + \eta}}. \tag{6.33}
\]
uniformly in \( z \in \{ z : 2 \leq |E| \leq 10, \bar{\eta}_E \leq \eta \leq 10, M\eta\sqrt{\kappa_E + \eta} \geq M' \} \) for any fixed \( \gamma > 0 \), where \( \kappa_E := |E| - 2 | \).

For generalized Wigner matrix, the threshold \( \bar{\eta}_E \) can be chosen \( \bar{\eta}_E = N^{-1+\gamma} \).

We point out two remarkable features of these bounds. The error term for the resolvent entries behaves essentially as \( (M\eta)^{-1/2} \), with an improvement near the edges where \( \text{Im} m_{sc} \) vanishes. The error bound for the Stieltjes transform, i.e., for the average of the diagonal resolvent entries, is one order better, \( (M\eta)^{-1} \), but without improvement near the edge.

The resolvent matrix element \( G_{ij} \) may be viewed as the scalar product \( \langle e_i, G e_j \rangle \) where \( e_i \) is the \( i \)-th coordinate vector. In fact, a more general version of (6.31), the isotropic local law also holds for generalized Wigner matrices:

**Theorem 6.8 (Isotropic law [21]).** For a generalized Wigner matrix with polynomial decay (5.6) and for any fixed unit vector \( v, w \) we have

\[
|\langle v, G(z)w \rangle - m_{sc}(z)\langle v, w \rangle| \lesssim \sqrt{\frac{\text{Im} m_{sc}(z)}{N\eta}} + \frac{1}{N\eta},
\]

uniformly in the set \( \{ z = E + i\eta : |E| \leq \omega^{-1}, N^{-1+\omega} \leq \eta \leq \omega^{-1} \} \) for any fixed \( \omega > 0 \).

Isotropic law was first proven in [88] for Wigner matrices under a vanishing third moment condition. The general case in the form above was given in [21]. We will not prove this result here since it is not needed for the proof of Theorem 5.1.

We will first prove a weaker version of Theorem 6.7 in Section 7, the so called weak local semicircle law, where the error term is not optimal. After that, we will prove Theorem 6.7 in Section 8 using \( \Gamma \) instead of \( \Gamma' \). This yields the same estimate as given in Theorem 6.7 but only on a smaller set of the spectral parameter for which the argument is somewhat simpler. The proof of Theorem 6.7 for the entire domain will only be sketched in Section 9 and we refer the reader to the original paper for the complete version. Section 7 is included mainly for pedagogical reasons to introduce the ideas of continuity argument and self-consistent equations. Section 8 demonstrates how to use the vector self-consistent equations in a simpler setup. In Section 9 we sketch the analysis of the same self-consistent equation but splitting it on space orthogonal to the constant vector and exploiting a spectral gap. Finally Section 10 is devoted to a key technical lemma, the fluctuation averaging lemma.

We close this section with a short summary of the main developments that have gradually led to the local semicircle law Theorem 6.7 in its general form. In Section 8.2 we will demonstrate that all proofs of the local semicircle law rely on some version of a self-consistent equation. At the beginning this was a scalar equation for \( m_N \) used in various forms by many authors, see e.g. Pastur [109], Girko [76] and Bai [10]. The self-consistent vector equation for \( v_i = G_{ii} - m_{sc} \) (see Section 8.2.2) first appeared in [69]. This allowed one to deviate from the identical distributions for \( h_{ij} \) and opened up the route to estimates on individual resolvent matrix elements. Finally, the self-consistent matrix equation for \( \mathbb{E}|G_{xy}|^2 \) first appeared in [54] and it yielded the diffusion profile for the resolvent. In this book we only need the vector equation.

The local semicircle law has three important features that have been gradually achieved. We explain them for the case when \( M \) is comparable with \( N \), the general case is only a technical extension. First, the local law in the bulk holds down to the scale expressed by the lower bound \( \eta \gg N^{-1} \). This is the smallest possible scale to control \( m(z) \), since at scale \( \eta \lesssim N^{-1} \) a few individual eigenvalues strongly influence its behavior and \( m(z) \) is not close to a deterministic quantity. This optimal scale in the bulk of the spectrum was first established for Wigner matrices in a series of papers [60–62] with the extension to generalized Wigner matrices in [69]. Second, the optimal speed of convergence in the entrywise bound (6.31) is \( N^{-1/2} \), while in the bound \( m - m_{sc} \) (6.32) it is \( N^{-1} \). These optimal \( N \)-dependences were first achieved in [68] by introducing the fluctuation averaging mechanism. Third, near the spectral edge the stability of the self-consistent equation deteriorates, manifested by the behavior of \( \Gamma(z) \) in (6.24) when \( \kappa_E \) is small. This can be compensated by the fact that the density is small near the edge. After many attempts and weaker results
in [64, 68, 69], the optimal form of this compensation was eventually found in [70], basically separating the analysis of the self-consistent equation onto the space orthogonal to constants. Since \( \widetilde{\Gamma} \) does not deteriorate near the edge, see (6.24), the estimate becomes optimal.

### 6.5 Appendix: Behaviour of \( \Gamma \) and \( \widetilde{\Gamma} \) and the proof of Lemma 6.3

In this appendix we give basic bounds on the parameters \( \Gamma \) and \( \widetilde{\Gamma} \). Readers interested only in the Wigner case may skip this section entirely; if \( s_{ij} = N^{-1} \), then the explicit formulas in (6.22) already suffice. As it turns out, the behaviour of \( \Gamma \) and \( \widetilde{\Gamma} \) is intimately linked with the spectrum of \( S \), more precisely with its spectral gaps. Recall that the spectrum of \( S \) lies in \([-1, 1]\), with 1 being a simple eigenvalue.

**Definition 6.9.** Let \( \delta_- \) be the distance from \(-1\) to the spectrum of \( S \), and \( \delta_+ \) the distance from 1 to the spectrum of \( S \) restricted to \( e^\perp \). In other words, \( \delta_{\pm} \) are the largest numbers satisfying

\[
S \preceq -1 + \delta_-, \quad S|_{e^\perp} \preceq 1 - \delta_+.
\]

For generalized Wigner matrices the lower and upper spectral gaps satisfy \( \delta_{\pm} \geq a \) with a constant \( a := \min\{C_{inf}, C_{sup}\} \), see (2.6). This simple fact follows easily by splitting

\[
S = (S - aee^*) + aee^*
\]

and noticing that the first term is \((1 - a)\) times a doubly stochastic matrix, hence its spectrum lies in \([-1 + a, 1 - a]\).

**Proof of Lemma 6.3.** The lower bound on \( \Gamma \) in (6.24) follows from \( (1 - m_{sc}^2S)^{-1}e = (1 - m_{sc}^2)^{-1}e \) combined with (6.12). For the upper bound, we first notice that \( 1 - m_{sc}^2S \) is invertible since \(-1 \leq S \leq 1\) and \( |m_{sc}| < 1 \), see (6.11). Since \( m_{sc} \) and its reciprocal are bounded (6.11), it is sufficient to bound the inverse of \( m_{sc}^2 - S \). Since the spectrum of \( S \) lies in the set \([-1 + \delta_-, 1 - \delta_+] \cup \{1\} \subset [-1 + a, 1 - a] \cup \{1\} \), and \( m_{sc}^2 \geq 1 \), we easily get

\[
\left\| \frac{1}{1 - m_{sc}^2} e^{\ell^2 \to \ell^2} \right\|_{\ell^2 \to \ell^2} \leq \left\| \frac{1}{m_{sc}^2 - S} \right\|_{\ell^2 \to \ell^2} \leq \frac{C}{\min\{a, |m_{sc}^2| - 1\}} \leq \frac{C_a}{|1 - m_{sc}^2|},
\]

(6.35)

In order to find the \( \ell^1 \to \ell^\infty \) norm, we solve \((1 - m_{sc}^2)\mathbf{v} = \mathbf{u}\) directly using (6.10):

\[
\|\mathbf{v}\|_\infty = \|\mathbf{u} + m_{sc}^2 \mathbf{Sv}\|_\infty \leq \|\mathbf{u}\|_\infty + \|S\|_{\ell^1 \to \ell^\infty} \|\mathbf{v}\|_1 \leq \|\mathbf{u}\|_\infty + N^{1/2}\|S\|_{\ell^1 \to \ell^\infty} \|\mathbf{v}\|_2 \\
\leq \|\mathbf{u}\|_\infty + N^{1/2}\|S\|_{\ell^1 \to \ell^\infty} \left\| \frac{1}{1 - m_{sc}^2} \right\|_{\ell^2 \to \ell^2} \|\mathbf{u}\|_2 \\
\leq (1 + \frac{C_a N\|S\|_{\ell^1 \to \ell^\infty}}{|1 - m_{sc}^2|}) \|\mathbf{u}\|_\infty.
\]

Here we used (6.35) and that \( \|\mathbf{v}\|_1 = \sum_i |v_i| \leq N^{1/2}\left( \sum_i |v_i|^2 \right)^{1/2} \) and \( \|\mathbf{u}\|_2 \leq N^{1/2} \|\mathbf{u}\|_\infty \). Since for generalized Wigner matrices \( N\|S\|_{\ell^1 \to \ell^\infty} = N \cdot \max_{ij} s_{ij} \leq C_{sup} \) from (2.6), this proves

\[
\left\| \frac{1}{1 - m_{sc}^2} \right\|_{\ell^\infty \to \ell^\infty} \leq \frac{C}{|1 - m_{sc}^2|},
\]

where the constant depends on \( C_{inf} \) and \( C_{sup} \). This proves the upper bound on \( \Gamma \) in (6.24).

Finally, we bound \( \widetilde{\Gamma} \); the lower bound was already given in (6.19). For the upper bound we follow the argument above but we restrict \( S \) to \( e^\perp \). Since the spectrum of this restriction lies in \([-1 + a, 1 - a]\), we immediately get the bound \( C/a \) for the \( \ell^2 \)-norm of \( (1 - m_{sc}^2S)^{-1}|_{e^\perp} \) in the right hand side of (6.35). This can be lifted to the same estimate for the \( \ell^\infty \) norm. This completes the proof of the lemma. \( \square \)
This simple proof of Lemma 6.3 used both spectral gaps and that \( s_{ij} = O(N^{-1}) \). Lacking these information in the general case, the following proposition gives explicit bounds on \( \Gamma \) and \( \bar{\Gamma} \) depending on the spectral gaps \( \delta_{\pm} \) in the general case. We recall the notations \( z = E + i\eta, \kappa = \kappa_E := |E| - 2 \) and define

\[
\theta(z) := \begin{cases} 
\frac{\kappa + \eta}{\sqrt{\kappa + \eta}} & \text{if } |E| \leq 2 \\
\frac{1}{\sqrt{|E|}} & \text{if } |E| > 2,
\end{cases} \tag{6.36}
\]

**Proposition 6.10.** For the matrix elements of \( S \) we assume \( 0 \leq s_{ij} = s_{ji} \leq M \) and \( \sum_j s_{ij} = 1 \). Then there is a universal constant \( C \) such that the following holds uniformly in the domain \( \{ z = E + i\eta : |E| \leq 10, M^{-1} \leq \eta \leq 10 \} \), and in particular in any spectral domain \( D \).

(i) We have the estimate

\[
\frac{1}{C\sqrt{\kappa + \eta}} \leq \Gamma(z) \leq \frac{C\log N}{1 - \max_{\pm} \left| \frac{1 \pm m_{sc} \xi}{2} \right|} \leq \frac{C\log N}{\min\{\eta + E^2, \theta\}}. \tag{6.37}
\]

(ii) In the presence of a gap \( \delta_- \) we may improve the upper bound to

\[
\Gamma(z) \leq \frac{C\log N}{\min\{\delta_- + \eta + E^2, \theta\}}. \tag{6.38}
\]

(iii) For \( \bar{\Gamma} \) we have the bounds

\[
C^{-1} \leq \bar{\Gamma}(z) \leq \frac{C\log N}{\min\{\delta_- + \eta + E^2, \delta_+ + \theta\}}. \tag{6.39}
\]

**Proof.** The first bound of (6.37) follows from \((1 - m_{sc}^2 S)^{-1} e = (1 - m_{sc}^2)^{-1} e \) combined with (6.12). In order to prove the second bound of (6.37), we write

\[
\frac{1}{1 - m_{sc}^2 S} = \frac{1}{2} \left( \frac{1 - 1 + m_{sc}^2}{2} \right),
\]

and observe that

\[
\left\| \frac{1 + m_{sc}^2}{2} \right\|_{\ell^2 \to \ell^2} \leq \left( \max_{\pm} \left| \frac{1 \pm m_{sc}^2}{2} \right| \right)^2 =: q. \tag{6.40}
\]

Therefore

\[
\left\| \frac{1}{1 - m_{sc}^2 S} \right\|_{\ell^\infty \to \ell^\infty} \leq \sum_{n=0}^{n_0} \left( \frac{1 + m_{sc}^2}{2} \right)^n + \sqrt{N} \sum_{n=n_0}^{\infty} \left( \frac{1 + m_{sc}^2}{2} \right)^n \leq n_0 + \sqrt{N} q^{n_0} \leq \frac{C\log N}{1 - q},
\]

where in the last step we chose \( n_0 = C_0 \) for large enough \( C_0 \). Here we used that \( \|S\|_{\ell^\infty \to \ell^\infty} \leq 1 \) and (6.11) to estimate the summands in the first sum. This concludes the proof of the second bound of (6.37).

The third bound of (6.37) follows from the elementary estimates

\[
\left| \frac{1 - m_{sc}^2}{2} \right| \leq 1 - c(\eta + E^2), \quad \left| \frac{1 + m_{sc}^2}{2} \right| \leq 1 - c \left( \frac{\Im m_{sc}}{\Im m_{sc} + \eta} \right) \leq 1 - c\theta \tag{6.41}
\]

for some universal constant \( c > 0 \), where in the last step we used Lemma 6.2.
The estimate (6.38) follows similarly. Due to the gap $\delta_-$ in the spectrum of $S$, we may replace the estimate (6.40) with
\[
\left\| \frac{1 + m_{sc}^2 S}{2} \right\|_{\ell^2 \to \ell^2} \leq \max \left\{ 1 - \delta_- - \eta - E^2, \left| \frac{1 + m_{sc}^2}{2} \right| \right\}.
\] (6.42)
Hence (6.38) follows using (6.41).

The lower bound of (6.39) was proved in (6.19). The upper bound is proved similarly to (6.38), except that (6.42) is replaced with
\[
\left\| \frac{1 + m_{sc}^2 S}{2} \right\|_{\ell^2 \to \ell^2} \leq \max \left\{ 1 - \delta_- - \eta - E^2, \min \left\{ 1 - \delta_+, \left| \frac{1 + m_{sc}^2}{2} \right| \right\} \right\}.
\]
This concludes the proof of (6.39).
7 Weak local semicircle law

Before we prove the local semicircle law in the strong form Theorem 6.7, for pedagogical reasons we first prove the following weaker version whose proof is easier. For simplicity, in this section we consider the Wigner case, i.e., \( s_{ij} = 1/N \) and \( M = N \). In the bulk, this weaker estimate is still effective for all \( \eta \) down to the smallest scales \( N^{-1+\varepsilon} \), but the power of \( 1/N \) in the error estimate is not optimal (1/2 instead of 1 in (6.32)). Near the edge the bound is even weaker; the power of \( 1/N \) is reduced to 1/4, indicating that this proof is not sufficiently strong near the edge.

**Theorem 7.1** (Weak local semicircle law). Let \( z = E + i\eta \) and \( \kappa := |E| - 2 \). Let \( H \) be a Wigner matrix, let \( G(z) = (H - z)^{-1} \) be its resolvent and set \( m_N(z) = 1/N \text{Tr} G(z) \). We assume that the single entry distribution satisfies the decay condition (5.6). Choose any \( \gamma > 0 \). Then for \( z \in D_\gamma := \{ E + i\eta : |E| \leq 10, N^{-1+\gamma} \leq \eta \leq 10 \} \) we have

\[
|m_N(z) - m_{sc}(z)| < \min \left\{ \frac{1}{\sqrt{N\eta\kappa}}, \frac{1}{(N\eta)^{1/4}} \right\}.
\]  

(7.1)

For definiteness, we present the proof for the Hermitian case, but all formulas below carry over to the other symmetry classes with obvious modifications.

7.1 Proof of the weak local semicircle law, Theorem 7.1

The proof is divided into four steps.

7.1.1 Step 1. Schur complement formula.

The first step to prove the weak local semicircle law is the to use the Schur complement formula, which we state in the following lemma. The proof is an elementary exercise.

**Lemma 7.2** (Schur formula). Let \( A, B, C \) be \( n \times n \), \( m \times n \) and \( m \times m \) matrices. We define \( (m+n) \times (m+n) \) matrix \( D \) as

\[
D := \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}
\]

(7.2)

and \( n \times n \) matrix \( \hat{D} \) as

\[
\hat{D} := A - B^*C^{-1}B.
\]

(7.3)

Then \( \hat{D} \) is invertible if \( D \) is invertible and for any \( 1 \leq i, j \leq n \), we have

\[
(D^{-1})_{ij} = (\hat{D}^{-1})_{ij}
\]

(7.4)

for the corresponding matrix elements. \( \square \)

Recall that \( G_{ij} = G_{ij}(z) \) denotes the matrix element of the resolvent

\[
G_{ij} = \left( \frac{1}{H - z} \right)_{ij}.
\]

Let \( H^{(i)} \) be the matrix where all matrix elements of \( H = (h_{ab}) \) in the \( i \)-th column and row are set to be zero:

\[
(H^{(i)})_{ab} := h_{ab} \cdot 1(a \neq i) \cdot 1(b \neq i), \quad a, b = 1, 2, \ldots, N.
\]

In other words, \( H^{(i)} \) is the \( i \)-th minor \( H^{[i]} \) of \( H \) augmented to an \( N \times N \) matrix by adding a zero row and column. Recall that the minor \( H^{[i]} \) is an \( (N - 1) \times (N - 1) \) matrix with the \( i \)-th row and column removed:

\[
H^{[i]}_{ab} := h_{ab}, \quad a, b \neq i.
\]
Denote the Green function of $H^{(i)}$ by $G^{(i)}(z) = (H^{(i)} - z)^{-1}$ which is again an $N \times N$ matrix. Notice that

$$G^{(i)}_{ab} = \begin{cases} (z)^{-1} & \text{if } a = b = i \\ 0 & \text{if exactly one of } a \text{ or } b \text{ equals } i \\ (H^{[i]} - z)^{-1}_{ab} & \text{if } a \neq i, b \neq i. \end{cases} \tag{7.5}$$

We warn the reader that in some earlier papers on the subject a different convention was used, where $H^{(i)}$ and $G^{(i)}$ denoted the $(N - 1) \times (N - 1)$ minors $(H^{[i]}$ with the current notation) and their Green functions. The current convention simplifies several formulas although the mathematical contents of both versions are identical.

With similar conventions, we can define $G^{(ij)}$, etc. The superscript in parenthesis for resolvents always means “after setting the corresponding row and column of $H$ to be zero” (in some terminology, this procedure is also described as ”zero out the corresponding row and column”), in particular, by independence of matrix elements, this means that the matrix $G^{(ij)}$, say, is independent of the $i$-th and $j$-th row and column of $H$. This helps to decouple dependencies in formulae. Let

$$a^i = (h_{1i}, h_{2i}, \ldots, h_{N_i})^t \tag{7.6}$$

be the $i$-th column of $H$, after setting the $i$-th entry zero.

Using Lemma 7.2 for $n = 1$, $m = N - 1$, and (7.5), (7.6) we have

$$G_{ii} = \frac{1}{h_{ii} - z - a^i \cdot G^{(i)} a^i}, \tag{7.7}$$

where

$$a^i \cdot G^{(i)} a^i = \sum_{k,l \neq i} h_{ik} G^{(i)}_{kl} h_{li} = \sum_{k,l \neq i} h_{ik} G^{[i]}_{kl} h_{li}. \tag{7.8}$$

(We sometimes use $u \cdot v$ instead of $u^* v$ or $(u, v)$ for the usual Hermitian scalar product.) We now introduce some notations:

**Definition 7.3** (Partial expectation and independence). Let $X \equiv X(H)$ be a random variable. For $i \in \{1, \ldots, N\}$ define the operations $P_i$ and $Q_i$ through

$$P_i X := \mathbb{E}(X|H^{(i)}), \quad Q_i X := X - P_i X.$$

We call $P_i$ partial expectation in the index $i$. Moreover, we say that $X$ is independent of a set $T \subset \{1, \ldots, N\}$ if $X = P_i X$ for all $i \in T$.

We can decompose $a^i \cdot G^{(i)} a^i$ into its expectation and fluctuation

$$a^i \cdot G^{(i)} a^i = P_i [a^i \cdot G^{(i)} a^i] + Z_i,$$

where

$$Z_i := Q_i [a^i \cdot G^{(i)} a^i]. \tag{7.9}$$

Since $G^{(i)}$ is independent of $a^i$, so we need to compute expectations and fluctuations of quadratic functions. The expectation is easy

$$P_i [a^i \cdot G^{(i)} a^i] = P_i \sum_{k,l} a_k^i G^{(i)}_{kl} a_l^i = \sum_{k,l \neq i} P_i [h_{ik} G^{(i)}_{kl} h_{li}] = \frac{1}{N} \sum_{k \neq i} G^{(i)}_{kk},$$

where in the last step we used that different matrix elements are independent, i.e., $P_i [h_{ik} h_{il}] = \frac{1}{N} \delta_{kl}$. The summations always run over all indices from 1 to $N$, apart from those that are explicitly excluded. We define

$$m_N^{(i)}(z) := \frac{1}{N - 1} \text{Tr} G^{(i)}(z) = \frac{1}{N - 1} \sum_{k \neq i} G^{(i)}_{kk}(z),$$

where we used $G^{[i]}_{kk} = G^{(i)}_{kk}$ for $k \neq i$ from (7.5). Hence we have the identity

$$G_{ii} = \frac{1}{h_{ii} - z - P_i [a^i \cdot G^{(i)} a^i]} - Z_i = \frac{1}{h_{ii} - z - (1 - \frac{1}{N}) m_N^{(i)}(z)} - Z_i. \tag{7.10}$$
7.1.2 Step 2. Interlacing of eigenvalues

We now estimate the difference between \( m_{sc} \) and \( m_N^{(i)} \). The first step is the following well known lemma. We include a short proof for completeness; for simplicity we consider the randomized setup with a continuous distribution to avoid multiple eigenvalues. The general case easily follows from standard approximation arguments.

**Lemma 7.4** (Interlacing of eigenvalues). Let \( H \) be a symmetric or Hermitian \( N \times N \) matrix with continuous distribution. Decompose \( H \) as follows

\[
H = \begin{pmatrix} h & a^* \\ a & B \end{pmatrix},
\]

(7.11)

where \( a = (h_{12}, \ldots, h_{1N})^* \) and \( B = H^{[1]} \) is the \( (N-1) \times (N-1) \) minor of \( H \) obtained by removing the first row and first column from \( H \). Denote by \( \mu_1 \leq \mu_2 \leq \ldots \leq \mu_N \) the eigenvalues of \( H \) and \( \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{N-1} \) the eigenvalues of \( B \). Then with probability one the eigenvalues of \( B \) are distinct and the eigenvalues of \( H \) and \( B \) are interlaced:

\[
\mu_1 < \lambda_1 < \mu_2 < \lambda_2 < \mu_3 < \ldots < \mu_{N-1} < \lambda_{N-1} < \mu_N.
\]

(7.12)

**Proof.** Since matrices with multiple eigenvalues form a lower dimensional submanifold within the set of all Hermitian matrices, the eigenvalues are distinct almost surely. Let \( \mu \) be one of the eigenvalues of \( H \) and let \( v = (v_1, \ldots, v_N)^t \) be a normalized eigenvector associated with \( \mu \). From the eigenvalue equation \( H v = \mu v \) and from (7.11) we find that

\[
h v_1 + a \cdot w = \mu v_1, \quad \text{and} \quad a v_1 + B w = \mu w
\]

(7.13)

with \( w = (v_2, \ldots, v_N)^t \). From these equations we obtain

\[
w = (\mu - B)^{-1} a v_1 \quad \text{and thus} \quad (\mu - h) v_1 = a \cdot (\mu - B)^{-1} a v_1 = \frac{v_1}{N} \sum_{\alpha} \frac{\xi_\alpha}{\mu - \lambda_\alpha},
\]

(7.14)

using the spectral representation of \( B \), where we set

\[
\xi_\alpha = |\sqrt{N} a \cdot u_\alpha|^2,
\]

with \( u_\alpha \) being the normalized eigenvector of \( B \) associated with the eigenvalue \( \lambda_\alpha \). From the continuity of the distribution it also that \( v_1 \neq 0 \) almost surely and thus we have

\[
\mu - h = \frac{1}{N} \sum_{\alpha} \frac{\xi_\alpha}{\mu - \lambda_\alpha},
\]

(7.15)

where \( \xi_\alpha \)'s are strictly positive almost surely (notice that \( a \) and \( u_\alpha \) are independent). In particular, this shows that \( \mu \neq \lambda_\alpha \) for any \( \alpha \). In the open interval \( \mu \in (\lambda_{\alpha-1}, \lambda_\alpha) \) the function

\[
\Phi(\mu) = \frac{1}{N} \sum_{\alpha} \frac{\xi_\alpha}{\mu - \lambda_\alpha}
\]

is strictly decreasing from \( \infty \) to \( -\infty \), therefore there is exactly one solution to the equation \( \mu - h = \Phi(\mu) \). Similar argument shows that there is also exactly one solution below \( \lambda_1 \) and above \( \lambda_{N-1} \). This completes the proof. \( \square \)

We can now compare the normalized traces of \( G \) and \( G^{[i]} \):

**Lemma 7.5.** Under the conditions of Lemma 7.4, for any \( 1 \leq i \leq N \), we have

\[
\left| m_N(z) - \left(1 - \frac{1}{N}\right) m_N^{(i)}(z) \right| \leq \frac{C}{N \eta}, \quad \eta = \text{Im} \ z > 0.
\]

(7.16)
Proof. With the notations of Lemma 7.4, let
\[ F(x) := \frac{1}{N} \# \{ \lambda_j \leq x \}, \quad F^{(i)}(x) := \frac{1}{N-1} \# \{ \mu_j \leq x \} \]
declare the normalized counting functions of the eigenvalues. The interlacing property of the eigenvalues of \( H \) and \( H^{[i]} \) (see (7.12)) in terms of these functions means that
\[ \sup_x |NF(x) - (N - 1)F^{(i)}(x)| \leq 1. \]
Then, after integrating by parts,
\[
\begin{align*}
|m_N(z) - (1 - \frac{1}{N})m_N^{(i)}(z)| &= \left| \int \frac{dF(x)}{x-z} - (1 - \frac{1}{N}) \int \frac{dF^{(i)}(x)}{x-z} \right| \\
&= \frac{1}{N} \left| \int NF(x) - (N - 1)F^{(i)}(x) \frac{dx}{(x-z)^2} \right| \\
&\leq \frac{1}{N} \int \frac{dx}{|x-z|^2} \leq \frac{C}{N\eta},
\end{align*}
\]
which proves (7.16). Note that the (7.16) would also hold without the prefactor \( (1 - \frac{1}{N}) \) since we have the trivial bound \( |m^{(i)}| \leq \eta^{-1} \).

We can now rewrite (7.10) into the following form, after summing over \( i \):
\[ m_N(z) = \frac{1}{N} \sum_i \frac{1}{-z - m_N(z) + \Omega_i}, \quad \text{with} \quad \Omega_i := h_{ii} - Z_i + O\left( \frac{1}{N\eta} \right). \tag{7.18} \]
At this stage we can explain the main idea for the proof of the local semicircle law (7.1). Equation (7.18) is the key self-consistent equation for \( m_N \), the Stieltjes transform of the empirical eigenvalue density of \( H \). Notice that if \( \Omega_i \) were zero, then we would have the equation
\[ m = -\frac{1}{m + z} \tag{7.19} \]
complemented with the side condition that \( \text{Im} \ m > 0 \). This is exactly the defining equation (6.8) of the Stieltjes transform of the semicircle density. In the next step we will give bounds on \( h_{ii} \) and \( Z_i \), and then we will effectively control the stability of (7.19) against small perturbations. This will give the desired estimate (7.1) on \( |m_N - m_{sc}| \).

7.1.3 Step 3. Large deviation estimate of the fluctuations.

The quantity \( Z_i \) defined in (7.9) will be viewed as a random variable in the probability space of the \( i \)-th column. We will need an upper bound on it in large deviation sense, i.e., with a very high probability. Since it is a quadratic function of the independent matrix elements of the \( i \)-th column, standard large deviation theorems do not directly apply. To focus on the main line of the proof and to keep technicalities at minimum, we postpone the discussion of full version of the quadratic large deviation bounds to Section 7.2. However, in order to get an idea of the size of \( Z_i \), we compute its second moment as follows:
\[ P_i |Z_i|^2 = \sum_{k \neq i} \sum_{k' \neq i} P_i \left[ \left( h_{ik} G_{kk'}^{(i)} h_{i'i'} - P_i \left[ h_{ik} G_{kk'}^{(i)} h_{i'i'} \right] \right) \left( \bar{G}_{kk'}^{(i)} \bar{G}_{kk'}^{(i)} - P_i \left[ \bar{G}_{kk'}^{(i)} \bar{G}_{kk'}^{(i)} \right] \right) \right]. \tag{7.20} \]
Since \( \mathbb{E} h = 0 \), the non-zero contributions to this sum come from index combinations when all \( h \) and \( \bar{h} \) are paired. For pedagogical simplicity, assume that \( \mathbb{E} h^2 = 0 \), this holds, for example, if the distribution of the real and imaginary parts are the same. Then \( h \) factors in the above expression have to be paired in such a
way that $h_{ik} = h_{ik'}$ and $h_{il} = h_{il'}$, i.e., $k = k'$, $l = l'$. Note that pairing $h_{ik} = h_{il}$ would give zero because the expectation is subtracted. The result is
\[
\mathbb{P}_i |Z_i|^2 = \frac{1}{N^2} \sum_{k,l \neq i} |G_{kl}^{(i)}|^2 + \frac{m_4 - 1}{N^2} \sum_{k \neq i} |G_{kk}^{(i)}|^2, \tag{7.21}
\]
where $m_4 = \mathbb{E}[\sqrt{N}h]^4$ is the fourth moment of the single entry distribution. The first term can be computed
\[
\frac{1}{N^2} \sum_{k,l \neq i} |G_{kl}^{(i)}|^2 = \frac{1}{N^2} \sum_{k,l \neq i} |G_{kl}^{[i]}|^2 = \frac{1}{N^2} \sum_{k \neq i} |(G_{kl}^{[i]}|^2)_{kk} = \frac{1}{N} \frac{1}{N} \sum_{k \neq i} \text{Im} G_{kk}^{[i]} = \frac{1}{N} \left( 1 - \frac{1}{N} \right) \text{Im} m_N^{(i)}, \tag{7.22}
\]
where $|G|^2 = GG^*$. In the middle step we have used the Ward identity valid for the resolvent $R(z) = (A - z)^{-1}$ of any Hermitian matrix $A$:
\[
|R(z)|^2 = \frac{1}{|A - E|^2 + \eta^2} = \frac{1}{\eta} \text{Im} R(z), \quad z = E + i\eta. \tag{7.23}
\]
We can estimate the second term in (7.21) by the general fact that the resolvent $R = (A - z)^{-1}$ of any Hermitian matrix $A$, we have
\[
\sum_k |R_{kk}|^2 \leq \sum_{\alpha} \frac{1}{|\zeta_{\alpha} - z|^2}, \tag{7.24}
\]
where $\zeta_{\alpha}$ are the eigenvalues of $A$. To see this, let $u_{\alpha}$ be the normalized eigenvectors. Then by spectral theorem
\[
R_{kk} = \sum_{\alpha} \frac{|u_{\alpha}(k)|^2}{\zeta_{\alpha} - z},
\]
and thus we have
\[
\sum_k |R_{kk}|^2 \leq \sum_k \sum_{\alpha, \beta} \frac{|u_{\alpha}(k)|^2 |u_{\beta}(k)|^2}{|\zeta_{\alpha} - z| |\zeta_{\beta} - z|} \leq \sum_{\alpha} \frac{1}{|\zeta_{\alpha} - z|^2} \sum_k |u_{\alpha}(k)|^2 \sum_{\beta} |u_{\beta}(k)|^2 = \sum_{\alpha} \frac{1}{|\zeta_{\alpha} - z|^2},
\]
where we have used the Schwarz inequality and that $\{u_{\beta}\}$ is an orthonormal basis. Applying this bound to the Green function of $H^{[i]}$ with eigenvalues $\mu_{\alpha}$, we have
\[
\frac{1}{N^2} \sum_{k \neq i} |G_{kk}^{[i]}|^2 = \frac{1}{N^2} \sum_{k \neq i} |G_{kk}^{[i]}|^2 \leq \frac{1}{N} \frac{1}{N} \sum_{\alpha = 1}^{N-1} \frac{\eta}{|\mu_{\alpha} - z|^2} = \frac{1}{N} \frac{1}{N} \left( 1 - \frac{1}{N} \right) \text{Im} m_N^{(i)}. \tag{7.25}
\]
By (7.16), we can estimate $m_N^{(i)}$ by $m_N$. Thus the estimates (7.22) and (7.25) confirm that the size of $Z_i$ is roughly
\[
|Z_i| \lesssim \frac{1}{\sqrt{N\eta}} \sqrt{\text{Im} m_N} \tag{7.26}
\]
in the second moment sense. In Section 7.2 we will prove that this inequality actually holds in large deviation sense, i.e., we have
\[
|Z_i| \sim \frac{1}{\sqrt{N\eta}} \sqrt{\text{Im} m_N}. \tag{7.27}
\]

The diagonal entry $h_{ii}$ can be easily estimated. Since the single entry distribution has finite moments (5.6), we have
\[
\mathbb{P}(|h_{ii}| \geq N^z N^{-1/2}) \leq C_p N^{-\varepsilon p}
\]
for each $i$ and for any $\varepsilon > 0$. Hence we can guarantee that all diagonal elements $h_{ii}$ simultaneously satisfy $|h_{ii}| \lesssim N^{-1/2}$. 

40
7.1.4 Step 4. Initial estimate at the large scales

To control the error terms in the self-consistent equation (7.18) we need two inputs. First, from now on we assume that (7.27) holds. Since $N\eta$ is large, this implies that $Z_i$ is small provided that $\text{Im} m_N$ is bounded. Second, we need to ensure that the denominator in the right hand side of (7.18) does not become too small. Since the main term in this denominator is $z + m_N(z)$, our task is to show that

$$\frac{1}{|z + m_N(z)|} < 1.$$  \hfill (7.28)

Notice that both inputs are in terms of the yet uncontrolled quantity $m_N$; they would be trivially available if $m_N$ were replaced with $m_{sc}$ (see Lemma 6.2). Since the smallness of $m_N - m_{sc}$ is our goal, to break this apparently circular argument we will use a bootstrap strategy. The convenient bootstrap parameter is $\eta$. We first establish the result for large $\eta$ in this section which is called the initial estimate. Then, in the next section, step by step we reduce the value of $\eta$ by using the control from the previous scale to estimate $\text{Im} m_N$ and $|z + m_N(z)|^{-1}$. This control will use the large deviation bounds on $Z_i$ which hold with very high probability. Hence at each step an exceptional event of very small probability will have to be excluded. This is the main reason why the bootstrap argument is done in small discrete steps, although in essence this is a continuity argument. We will explain this important argument in more details in the next section.

Both the initial estimate and the continuity argument use the following simple idea. By expanding $\Omega_i$ in the denominator in (7.18) and using (7.27) and $h_{ii} \preceq N^{-1/2}$, we have

$$\left| m_N(z) + \frac{1}{z + m_N(z)} \right| < \frac{1}{|z + m_N(z)|^2} \max_i |\Omega_i|$$

$$\preceq \frac{1}{|z + m_N(z)|^2} \left[ \sqrt{\text{Im} m_N} \frac{N}{N\eta} + N^{-1/2} + (N\eta)^{-1} \right]$$  \hfill (7.29)

provided that $\Omega_i$ can be considered as a small perturbation of $z + m_N(z)$, i.e.

$$\frac{1}{|z + m_N(z)|} \max_i |\Omega_i| \ll 1,$$  \hfill (7.30)

so that the expansion is justified. We can now use the following elementary lemma.

**Lemma 7.6.** Fix $z = E + i\eta$, with $|E| \leq 20$, $0 < \eta \leq 10$, and set $\kappa := |E| - 2$. Suppose that $m$ satisfies the inequality

$$|m + \frac{1}{z + m}| \leq \delta$$  \hfill (7.31)

for some $\delta \leq 1$. Then

$$\min \left\{ |m - m_{sc}(z)|, \left| m - \frac{1}{m_{sc}(z)} \right| \right\} \leq \frac{C\delta}{\sqrt{\kappa + \eta + \delta}} \leq C\sqrt{\delta}.$$  \hfill (7.32)

**Proof.** For $\delta \leq 1$ and $|z| \leq 20$, (7.31) implies that $|m| \leq 22$. Write (7.31) as

$$m + \frac{1}{z + m} =: \Delta, \quad |\Delta| \leq \delta,$$

and subtract this from the equation $m_{sc} + (z + m_{sc})^{-1} = 0$. After some simple algebra we get

$$(m - m_{sc})\left[ m - \frac{1}{m_{sc}} \right] = \Delta(m + z),$$  \hfill (7.33)

i.e.,

$$|m - m_{sc}| \left| m - \frac{1}{m_{sc}} \right| \leq C\delta,$$  \hfill (7.34)
for some fixed constant $C$, where we have used (6.11).

We separate two cases. If $|1 - m_{sc}^2| \leq C' \sqrt{\delta}$ for some large $C'$, then we write the above inequality as

$$|m - m_{sc}| \left| m - m_{sc} - \frac{1 - m_{sc}^2}{m_{sc}} \right| \leq C\delta.$$  \hspace{1cm} (7.35)

We claim that this implies $|m - m_{sc}| \leq 2C' \sqrt{\delta}/|m_{sc}|$. Indeed, if $|m - m_{sc}| \geq 2C' \sqrt{\delta}/|m_{sc}|$ were true, then the second factor in (7.35) were at least $C' \sqrt{\delta}/|m_{sc}|$, so the left hand side of (7.35) were at least $2(C'/|m_{sc}|)^2 \delta$. Since $m_{sc} \gg 1$, see (6.11), we would get a contradiction if $C'$ is large enough. Thus we proved $|m - m_{sc}| \leq 2C' \sqrt{\delta}/|m_{sc}| \leq C\sqrt{\delta}$ in this case, which is in agreement with the first inequality in (7.32) since the condition $|1 - m_{sc}^2| \leq C' \sqrt{\delta}$ also implies $\sqrt{\kappa + \eta} \leq \sqrt{\delta}$ by (6.12).

In the second case we have $|1 - m_{sc}^2| \gg C' \sqrt{\delta}$, i.e., $\sqrt{\kappa + \eta} \gg \sqrt{\delta}$, so for (7.32) we need to prove that $|m - m_{sc}| \leq \delta/\sqrt{\kappa + \eta}$ or $|m - m_{sc}^{-1}| \leq \delta/\sqrt{\kappa + \eta}$. If $|m - m_{sc}| \leq \frac{1}{2}|1 - m_{sc}^2|/|m_{sc}|$, then the second factor in (7.35) were at least $\frac{1}{2}|1 - m_{sc}^2|/|m_{sc}| \gg \sqrt{\kappa + \eta}$, so we would immediately get $|m - m_{sc}| \leq \delta/\sqrt{\kappa + \eta}$. We are left with the case $|m - m_{sc}| \gg \frac{1}{2}|1 - m_{sc}^2|/|m_{sc}|$, i.e., $|m - m_{sc}| \gg \sqrt{\kappa + \eta}$. Rewrite (7.35) as

$$|m - \frac{1}{m_{sc}} + \frac{1 - m_{sc}^2}{m_{sc}}| \left| m - \frac{1}{m_{sc}} \right| \leq C\delta,$$  \hspace{1cm} (7.36)

and repeat the previous argument with interchanging the role of $m_{sc}$ and $1/m_{sc}$. We conclude that either $|m - m_{sc}^{-1}| \leq \frac{1}{2}|1 - m_{sc}^2|/|m_{sc}|$, in which case we immediately get $|m - m_{sc}^{-1}| \leq \delta/\sqrt{\kappa + \eta}$, or $|m - m_{sc}^{-1}| \gg \sqrt{\kappa + \eta}$. In the latter case, combining it with $|m - m_{sc}| \gg \sqrt{\kappa + \eta}$ from the previous argument, we would get $|m - m_{sc}| \gg \sqrt{\kappa + \eta}$. Since $\kappa + \eta \gg |1 - m_{sc}^2| \gg C'\delta$, this would contradict (7.34) if $C'$ is large enough. This completes the proof of the lemma. \hfill \square

Applying Lemma 7.6 in (7.29), we see that

$$\min \left\{ |m_N(z) - m_{sc}(z)|, \left| m_N(z) - \frac{1}{z + m_N(z)} \right| \right\} \leq \frac{1}{|z + m_N(z)|} \min \left\{ \left( \max_i |\Omega_i| \right)^{1/2}, \max_i \frac{|\Omega_i|}{\sqrt{\kappa}} \right\},$$  \hspace{1cm} (7.37)

i.e., a good bound on $\Omega_i$ directly yields an estimate on $m_N - m_{sc}$ provided a lower bound on $|z + m_N(z)|$ is given and if the possibility that $|m_N(z) - m_{sc}(z)|$ is small can be excluded.

Using this scheme, we now derive the initial estimate, which is (7.1) for any $\eta \geq N^{-1/16}$. To check (7.30), we start with the trivial bound $\Im m_N \leq \eta^{-1}$. By (7.27), we have $Z_i \sim N^{-15/32}$ for $\eta \geq N^{-1/16}$. Hence we have

$$\max_i |\Omega_i| \leq \max_i |Z_i| + |h_{ii}| + C N^{-15/16} \sim N^{-15/32}.$$  \hspace{1cm} (7.38)

Since $\Im(z + m_N(z)) \geq \eta \geq N^{-1/16}$, (7.30) is satisfied for $\eta \geq N^{-1/16}$. Thus (7.29) implies that

$$\left| m_N(z) + \frac{1}{z + m_N(z)} \right| \leq \frac{1}{|z + m_N(z)|^2} O(N^{-15/32}) \ll N^{-11/32}$$  \hspace{1cm} (7.39)

for any $\eta \geq N^{-1/16}$. The precise exponents do not matter here, our goal is to find some $\eta = N^{-c}$, such that the last equation holds with an error $N^{-c'}$ for some $c, c' > 0$.

Applying Lemma 7.6 we obtain, for any $\eta \geq N^{-1/16}$, that either

$$|m_N(z) - m_{sc}(z)| \ll N^{-11/64}, \quad \text{or} \quad \left| m_N(z) - \frac{1}{m_{sc}(z)} \right| \ll N^{-11/64}.$$  \hspace{1cm} (7.40)

However, the second option is excluded since $\Im m_N \geq 0$ and thus

$$\left| m_N(z) - \frac{1}{m_{sc}(z)} \right| \gg \Im m_N - \Im \frac{1}{m_{sc}(z)} \gg c \Im m_{sc}(z) \gg c\eta \gg cN^{-1/16},$$  \hspace{1cm} (7.41)

where we used (6.13). This proves that

$$|m_N(z) - m_{sc}(z)| \ll N^{-11/64}.$$  \hspace{1cm} (7.42)
Once we know that \(|m_N(z) - m_{sc}(z)|\) is small, we can simply perturb the relation
\[
\left| \frac{1}{z + m_{sc}(z)} \right| = |m_{sc}(z)| \leq 1
\]
from Lemma 6.2 to obtain
\[
|m_N(z)| \leq C, \quad \text{and} \quad \left| \frac{1}{z + m_N(z)} \right| \leq C. \tag{7.43}
\]
Thus we can bound \(\Omega_i\) from (7.27) and \(h_{ii} \prec N^{-1/2}\) by
\[
\max_i |\Omega_i| \prec \frac{1}{\sqrt{N\eta}}, \tag{7.44}
\]
and, instead of (7.39), we get the stronger bound
\[
\left| m_N(z) + \frac{1}{z + m_N(z)} \right| \prec \frac{1}{\sqrt{N\eta}}. \tag{7.45}
\]
Applying (7.32) once again, with \(\delta = (N\eta)^{-1/2}\), and using (7.41) to exclude the possibility that \(m_N\) is close to \(1/m_{sc}\), we obtain the better bound
\[
|m_N(z) - m_{sc}(z)| \prec \min \left( \frac{1}{(N\eta)^{1/4}}, \frac{1}{\sqrt{N\eta\kappa}} \right), \quad \text{for any } \eta \geq N^{-1/16}, \tag{7.45}
\]
which is exactly (7.1) for \(\eta \geq N^{-1/16}\).

7.1.5 Step 5. Continuity argument and completion of the proof

With (7.1) proven for any \(\eta \geq N^{-1/16}\), we now proceed to reduce the scale of \(\eta\), while \(E\), the real part of \(z\), is kept fixed. To do this, choose another scale \(\eta_1 = \eta - N^{-4}\) slightly smaller than \(\eta\) (but still \(\eta_1 \geq N^{-1}\)). Since
\[
|m'_{sc}(z)| \leq N^{-1} \sum_{\alpha} \frac{1}{|\lambda_{\alpha} - z|^2} \leq \eta^{-2}, \tag{7.46}
\]
we have the deterministic relation
\[
|m_N(z_1) - m_N(z)| \leq N^{-2}, \quad z = E + i\eta, \quad z_1 = E + i\eta_1. \tag{7.47}
\]
Hence (7.1) holds as well for \(\eta_1\) instead of \(\eta\) since \(N^{-2}\) is smaller than all our error terms. But we cannot rely on this idea for the obvious reason that we would need to apply this procedure at least \(N^4\) times and the many small errors would accumulate. We need to show that the estimate in (7.1) does not deteriorate even by the small amount \(N^{-2}\). Since the definition of \(\prec\) includes additional factors \(N^\varepsilon\), it is not sufficiently sensitive to track such precision. For the continuity argument we will now abandon the formalism of stochastic domination and we go back to tracking exceptional sets precisely. In Section 8 we will setup a continuity scheme fully within the framework of stochastic domination, but here we complete the proof in a more elementary way.

The idea is to show that for any small exponent \(\varepsilon \in (0, \frac{1}{10})\) there is a constant \(C_1\) large enough such that if
\[
S(z) := \min \left\{ |m_N(z) - m_{sc}(z)|, |m_N(z) - m_{sc}^{-1}(z)| \right\} \leq C_1 N^\varepsilon \min \left( \frac{1}{(N\eta)^{1/4}}, \frac{1}{\sqrt{N\eta\kappa}} \right) \tag{7.48}
\]
in a set \(A\) in the probability space, then we not only have the bound
\[
S(z_1) \leq C_1 N^\varepsilon \min \left( \frac{1}{(N\eta)^{1/4}}, \frac{1}{\sqrt{N\eta\kappa}} \right) + 2N^{-2} \tag{7.49}
\]
43
that trivially follows from (7.47), but we also have that

$$S(z_1) \leq C_1 N^\varepsilon \min \left( \frac{1}{(N \eta_1)^{1/4}}, \frac{1}{\sqrt{N \eta \kappa}} \right)$$

(7.50)

holds with the same $C_1$ and $\varepsilon$ in a set $A_1 \subset A$ with

$$P(A \setminus A_1) \leq N^{-D} \quad \text{for any } D > 0.$$ 

(7.51)

Thus the deterioration of the estimate from (7.47) to (7.50) can be avoided at least with a very high probability. To see this, we note that in the regime $\eta \geq N^{-1+\delta}$ the bound (7.49) implies that in the set $A$ we have

$$|\text{Im } m_N(z_1)| \leq |m_N(z_1)| \leq |m_{sc}(z_1)| + o(1) \leq 2,$$

(7.52)

and

$$\frac{1}{|z_1 + m_N(z_1)|} \leq 2.$$ 

(7.53)

The last estimate is a perturbative consequence of the bounds $|z + m_{sc}(z)|^{-1} \leq 1$ and $|z + m_{sc}^{-1}(z)|^{-1} \leq C$, where we used either

$$|(z_1 + m_N(z_1)) - (z + m_{sc}(z))| \leq |m_N(z_1) - m_N(z)| + |z_1 - z| + |m_N(z) - m_{sc}(z)| = o(1)$$

(7.54)

or

$$|(z_1 + m_N(z_1)) - (z + m_{sc}^{-1}(z))| \leq |m_N(z_1) - m_N(z)| + |z_1 - z| + |m_N(z) - m_{sc}^{-1}(z)| = o(1);$$

one of which follows from (7.47) and (7.48). Let $A_1$ be the intersection of three sets: $A$, the set on which $|h_i| \leq N^{-\frac{1}{4}+\varepsilon}$ holds and the set that (7.27) holds at the scale $\eta_1$, i.e., at $z = E + i \eta_1$. Hence, together with (7.53), the condition (7.30) holds in $A_1$. Now we can use (7.37) together with (7.52) and (7.53) to prove that (7.50) holds if $C_1$ is chosen large enough.

At each step, we lose a set of probability $N^{-D}$ due to checking (7.27) at that scale. The estimate from the previous scale is used only to check (7.52), (7.53) and (7.30). Finally, the constant $C_1$ in the final bound of $S(z_1)$ comes from the constant in (7.32) which is uniform in $\eta$. Therefore $C_1$ does not deteriorate when passing to the next scale. The only price to pay is the loss of an exceptional set of probability $N^{-D}$. Since the number of steps are $N^4$, while $D$ can be arbitrary large, this loss is affordable. Since the exponent $\varepsilon > 0$ was arbitrary, this proves

$$S(z) = \min \left\{ |m_N(z) - m_{sc}(z)|, |m_N(z) - m_{sc}^{-1}(z)| \right\} \prec \min \left\{ \frac{1}{\sqrt{N \eta \kappa}}, \frac{1}{(N \eta)^{1/4}} \right\}$$

(7.55)

for any fixed $z \in D_\kappa$. Using this relation for a discrete net $D_\kappa := D_\gamma \cap N^{-4}(Z + iZ)$ and a union bound, we obtain that (7.55) holds simultaneously for all $z \in D_\gamma$. Using the Lipschitz continuity of $S(z)$ (with a Lipschitz constant at most $N^2$), we get (7.55) simultaneously for all $z \in D_\kappa$. Finally, we need to show that the estimate holds not only for the minimum, but for $|m_N(z) - m_{sc}(z)|$. We have already seen this for any $\eta \geq N^{-1/16}$ in (7.45). Fixing $E = \text{Re } z$, we consider $m_N(E + i \eta)$, $m_{sc}(E + i \eta)$ and $S(E + i \eta)$ as functions of $\eta$. Since these are continuous functions, by reducing $\eta$ we obtain that $m_N(E + i \eta)$ remains close to $m_{sc}(E + i \eta)$ as long as the right hand side (7.55) is larger than the difference

$$|m_{sc}(E + i \eta) - m_{sc}^{-1}(E + i \eta)| \leq \sqrt{\kappa_E + \eta}.$$ 

Since the right hand side of (7.55) increases as $\eta$ decreases, while the separation bound $\sqrt{\kappa_E + \eta}$ decreases, once $\eta$ is so small that the separation bound $\sqrt{\kappa_E + \eta}$ becomes smaller than the right hand side of (7.55), the difference between $m_{sc}$ and $m_{sc}^{-1}$ remains irrelevant for any smaller $\eta$. Thus (7.1) holds on the entire $D_\kappa$. This proves the weak local semicircle law, i.e., Theorem 7.1 except the large deviation estimate (7.27) which we will prove in the next subsection.
7.2 Large deviation estimates

Finally, in order to estimate large sums of independent random variables as in (7.8) and later in (8.2), we will need a large deviation estimate for linear and quadratic functionals of independent random variables. The case of the linear functionals is standard. Quadratic functionals were considered in [62,69]; the current formulation is taken from [52].

Theorem 7.7 (Large deviation bounds). Let \(X_i^{(N)}\) and \(Y_i^{(N)}\) be independent families of random variables and \((a_{ij}^{(N)})\) and \((b_i^{(N)})\) be deterministic; here \(N \in \mathbb{N}\) and \(i,j = 1, \ldots, N\). Suppose that all entries \(X_i^{(N)}\) and \(Y_i^{(N)}\) are independent and satisfy
\[
\mathbb{E}X = 0, \quad \mathbb{E}|X|^2 = 1, \quad \|X\|_p := (\mathbb{E}|X|^p)^{1/p} \leq \mu_p
\] (7.56)
for all \(p \in \mathbb{N}\) and some constants \(\mu_p\). Then we have the bounds
\[
\sum_i b_i X_i \prec \left( \sum_i |b_i|^2 \right)^{1/2},
\] (7.57)
\[
\sum_{i,j} a_{ij} X_i Y_j \prec \left( \sum_{i,j} |a_{ij}|^2 \right)^{1/2},
\] (7.58)
\[
\sum_{i \neq j} a_{ij} X_i X_j \prec \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2}.
\] (7.59)

Our proof in fact generalizes trivially to arbitrary multilinear estimates for quantities of the form \(\sum_{i_1,\ldots,i_k} a_{i_1\ldots i_k}(u)X_{i_1}(u)\cdots X_{i_k}(u)\), where the star indicates that the summation indices are constrained to be distinct.

For our purposes the most important inequality is the last one (7.59). It confirms the intuition from Section 7.1.3 that for sums of the form \(\sum_{i \neq j} a_{ij} X_i X_j\) the second moment calculation gives the correct order of magnitude and it can be improved to a high probability statement in large deviation sense. Indeed, the second moment calculation gives
\[
\mathbb{E} \left| \sum_{i \neq j} a_{ij} X_i X_j \right|^2 = \sum_{i \neq j, i' \neq j'} a_{ij} \bar{a}_{i'j'} X_i X_j \bar{X}_{i'} \bar{X}_{j'} = \sum_{i \neq j} \left[ |a_{ij}|^2 + a_{ij} \bar{a}_{ji} \right] \leq 2 \sum_{i \neq j} |a_{ij}|^2,
\]

since only the pairings \(i = i', j = j'\) or \(i = j', j = i'\) give nonzero contribution.

To prepare the proof of Theorem 7.7, we first recall the following version of the Marcinkiewicz-Zygmund inequality.

Lemma 7.8. Let \(X_1, \ldots, X_N\) be a family of independent random variables each satisfying (7.56) and suppose that the family \((b_i)\) is deterministic. Then
\[
\left\| \sum_i b_i X_i \right\|_p \leq (Cp)^{1/2} \mu_p \left( \sum_i |b_i|^2 \right)^{1/2}.
\] (7.60)

Proof. The proof is a simple application of Jensen’s inequality. Writing \(B^2 := \sum_j |b_i|^2\), we get, by the
classical Marcinkiewicz-Zygmund inequality \[126\] in the first line, that
\[
\left\| \sum_i b_i X_i \right\|_p^{p} \leq \left( C_p \right)^{p/2} \left\| \left( \sum_i |b_i|^2 |X_i|^2 \right)^{1/2} \right\|_p^{p}
\]
\[
= \left( C_p \right)^{p/2} B_p \mathbb{E} \left[ \left( \sum_i |b_i|^2 |X_i|^2 \right)^{p/2} \right]
\]
\[
\leq \left( C_p \right)^{p/2} B_p \mathbb{E} \left[ \sum_i |b_i|^2 |X_i|^2 \right]
\]
\[
\leq \left( C_p \right)^{p/2} B_p \mu_p^p.
\]

Next, we prove the following intermediate result.

**Lemma 7.9.** Let \(X_1, \ldots, X_N, Y_1, \ldots, Y_N\) be independent random variables each satisfying (7.56), and suppose that the family \((a_{ij})\) is deterministic. Then for all \(p \geq 2\) we have
\[
\left\| \sum_{i,j} a_{ij} X_i Y_j \right\|_p \leq C_p \mu_p^2 \left( \sum_{i,j} |a_{ij}|^2 \right)^{1/2}.
\]

**Proof.** Write
\[
\sum_{i,j} a_{ij} X_i Y_j = \sum_j b_j Y_j, \quad b_j := \sum_i a_{ij} X_i.
\]
Note that \((b_j)\) and \((Y_j)\) are independent families. By conditioning on the family \((b_j)\), we therefore get from Lemma 7.8 and the triangle inequality that
\[
\left\| \sum_j b_j Y_j \right\|_p \leq \left( C_p \right)^{1/2} \mu_p \left\| \sum_j |b_j|^2 \right\|_{p/2}^{1/2} \leq \left( C_p \right)^{1/2} \mu_p \left( \sum_j \|b_j\|_p^2 \right)^{1/2}.
\]
Using Lemma 7.8 again, we have
\[
\|b_j\|_p \leq \left( C_p \right)^{1/2} \mu_p \left( \sum_i |a_{ij}|^2 \right)^{1/2}.
\]
This concludes the proof.

**Lemma 7.10.** Let \(X_1, \ldots, X_N\) be independent random variables each satisfying (7.56), and suppose that the family \((a_{ij})\) is deterministic. Then we have
\[
\left\| \sum_{i \neq j} a_{ij} X_i X_j \right\|_p \leq C_p \mu_p^2 \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2}.
\]

**Proof.** The proof relies on the identity (valid for \(i \neq j\))
\[
1 = \frac{1}{2^N} \sum_{I \cup J = \mathbb{N}} 1(i \in I)1(j \in J), \quad (7.61)
\]
where the sum ranges over all partitions of \(\mathbb{N}_N = \{1, \ldots, N\}\) into two sets \(I\) and \(J\), and \(Z_N := 2^{N-2}\) is independent of \(i\) and \(j\). Moreover, we have
\[
\sum_{I \cup J = \mathbb{N}} 1 = 2^{N} - 2, \quad (7.62)
\]

46
where the sum ranges over nonempty subsets \( I \) and \( J \). Now we may estimate

\[
\left\| \sum_{i \neq j} a_{ij} X_i X_j \right\|_p \leq \frac{1}{Z_N} \sum_{I \cup J = N} \left\| \sum_{i \in I} \sum_{j \in J} a_{ij} X_i X_j \right\|_p \leq \frac{1}{Z_N} \sum_{I \cup J = N} Cp \mu_p^2 \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2},
\]

where we used that, for any partition \( I \cup J = N \), the families \((X_i)_{i \in I}\) and \((X_j)_{j \in J}\) are independent, and hence the Lemma 7.9 is applicable. The claim now follows from (7.62).

As remarked above, the proof of Lemma 7.10 may be easily extended to multilinear expressions of the form \( \sum_{i_1, \ldots, i_k} a_{i_1 \ldots i_k} X_{i_1} \cdots X_{i_k} \).

We may now complete the proof of Theorem 7.7.

**Proof of Theorem 7.7.** The proof is a simple application of Chebyshev’s inequality. Part (i) follows from Lemma 7.8, part (ii) from Lemma 7.10, and part (iii) from Lemma 7.9. We give the details for part (iii).

For \( \varepsilon > 0 \) and \( D > 0 \) we have

\[
P \left[ \sum_{i \neq j} a_{ij} X_i X_j \geq N^\varepsilon \Psi \right] \leq P \left[ \sum_{i \neq j} a_{ij} X_i X_j \geq N^\varepsilon \Psi, \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2} \leq N^{\varepsilon/2} \Psi \right]
+ P \left( \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2} \geq N^{\varepsilon/2} \Psi \right)
\leq P \left[ \sum_{i \neq j} a_{ij} X_i X_j \geq N^{\varepsilon/2} \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2} \right] + N^{-D-1}
\leq \left( \frac{Cp \mu_p^2}{N^{\varepsilon/2}} \right)^p + N^{-D-1}
\]

for arbitrary \( D \). In the second step we used the definition of \( \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2} < \Psi \) with parameters \( \varepsilon/2 \) and \( D + 1 \). In the last step we used Lemma 7.10 by conditioning on \((a_{ij})\). Given \( \varepsilon \) and \( D \), there is a large enough \( p \) such that the first term on the last line is bounded by \( N^{-D-1} \). Since \( \varepsilon \) and \( D \) were arbitrary, the proof is complete.

The claimed uniformity in \( u \) in the case that \( a_{ij} \) and \( X_i \) depend on an index \( u \) also follows from the above estimate.  \( \square \)
8 Proof of the local semicircle law

In this section we start the proof of the local semicircle law, Theorem 6.7. This section can be read independently of the previous Section 7, so some basic definitions and facts are repeated for convenience. For those readers who may wish to compare this argument with the proof the weak law, Theorem 7.1, we mention that the basic strategy is similar except that we use an additional mechanism that we call fluctuation averaging. We point out that the first estimate in (7.29) was not optimal; here we estimated the average of \( \Omega_i \) by its maximum. Since \( \Omega_i \)'s are almost centred random variables with weak correlation, a more precise estimate of their average leads to a considerable improvement. We now recall that the basic steps to prove the weak law were

1) self consistent equation for \( m_N \);  
2) interlacing of eigenvalues to compare \( m_N \) and \( m^{(i)}_N \);  
3) quadratic large deviation estimate for the error term \( Z_i \);  
4) initial estimate for large \( \eta \);  
5) extend the estimate to small \( \eta \) by the continuity argument.

Exploiting the fluctuation averaging mechanism requires a control on the individual matrix elements of the resolvent \( G \) instead of just its normalized trace, \( m_N = N^{-1} \text{Tr} G \). Therefore, instead of considering the scalar equation for \( m_N \), we will consider the vector self-consistent equation for the diagonal elements \( G_{ii} \) of the Green function and investigate the stability of this equation. There is no direct analogue of the interlacing property for \( G_{ii} \), thus we introduce new resolvent decoupling identities to compare the resolvents of the original matrix and its minors. We will still use the quadratic large deviation estimate to bound the error term. We also use a continuity argument similar to the one given in the weak law to derive a crude estimate on \( G_{ii} \) (formulated in terms of a certain dichotomy), but instead of making many small steps and keeping track of the exceptional sets, we follow a genuinely continuous approach within the framework of the stochastic domination. Finally, we use the fluctuation averaging lemma (Lemma 8.9) to boost the error estimate by one order and an iteration argument to prove the local semicircle law, Theorem 6.7. We now start the rigorous proof. We will largely follow the presentation in [55].

8.1 Tools

In this subsection we collect some basic definitions and facts. First we repeat the Definition 7.3 of the partial expectation:

**Definition 8.1** (Partial expectation and independence). Let \( X \equiv X(H) \) be a random variable. For \( i \in \{1, \ldots, N\} \) define the operations \( P_i \) and \( Q_i \) through

\[
P_i X := \mathbb{E}(X|H^{(i)}), \quad Q_i X := X - P_i X.
\]

We call \( P_i \) partial expectation in the index \( i \). Moreover, we say that \( X \) is independent of a set \( T \subset \{1, \ldots, N\} \) if \( X = P_i X \) for all \( i \in T \).

Next, we define matrices with certain columns and rows zeroed out.

**Definition 8.2.** For \( T \subset \{1, \ldots, N\} \) we set \( H^{(T)} \) to be the \( N \times N \) matrix defined by

\[ (H^{(T)})_{ij} := 1(i \notin T)1(j \notin T)h_{ij}, \quad i, j = 1, 2, \ldots, N. \]

Moreover, we define the resolvent of \( H^{(T)} \) and its normalized trace through

\[
G_{ij}^{(T)}(z) := (H^{(T)} - z)^{-1}, \quad m^{(T)}(z) := \frac{1}{N} \text{Tr} G^{(T)}(z).
\]
We also set the notation

\[ \sum_{i}^{(T)} := \sum_{i : i \not\in T}. \]

These definitions are the natural generalizations of \( H^{(i)} \) and \( G^{(i)} \) introduced in Section 7.1.1. In particular, notice that \( H^{(T)} \) is the matrix obtained by setting all rows and columns in \( T \) to be zero. This is different from considering the minors by removing the columns and rows in \( T \). Similarly, \( G^{(T)} \) is still an \( N \times N \) matrix with \( G^{(T)}_{ii} = -z^{-1} \) for \( i \in T \) and \( G^{(T)}_{ij} = 0 \) if \( i \in T \) and \( j \not\in T \). We will denote \( G^{(i)} \) simply by \( G^{(i)} \) and similarly for a few more indices. This is consistent with the notations we used in the earlier chapters.

The following resolvent decoupling identities form the backbone of all of our calculations. The idea behind them is that a resolvent matrix element \( G_{ij} \) depends strongly on the \( i \)-th and \( j \)-th columns of \( H \), but weakly on all other columns. The first identity determines how to make a resolvent matrix element \( G_{ij} \) behind them is that a resolvent matrix element \( G_{ij} \) is still an \( ij \)-th or in the \( i \)-th column of \( H \). We added a third identity that relates sums of off-diagonal resolvent entries with a diagonal one. The proofs are elementary.

**Lemma 8.3** (Resolvent decoupling identities). For any real or complex Hermitian matrix \( H \) and \( T \subset \{1, \ldots, N\} \) the following identities hold.

\[ i) \text{ First resolvent decoupling identity [69]: If } i, j, k \not\in T \text{ and } i, j \neq k \text{ then} \]

\[ G^{(T)}_{ij} = G^{(T)}_{ik} G^{(T)}_{kj} G^{(T)}_{kk}. \quad (8.1) \]

\[ ii) \text{ Second resolvent decoupling identity [52]: If } i, j \not\in T \text{ satisfy } i \neq j \text{ then} \]

\[ G^{(T)}_{ij} = -G^{(T)}_{ii} \sum_{k} h_{ik} G^{(T)}_{kj} = -G^{(T)}_{jj} \sum_{k} G^{(T)}_{ik} h_{kj}, \quad (8.2) \]

where the superscript in the summation means omission, e.g., the summation in the first sum runs over all \( k \not\in T \cup \{i\} \).

\[ iii) \text{ [Ward identity] For any } T \subset \{1, \ldots, N\} \text{ we have} \]

\[ \sum_{j} |G^{(T)}_{ij}|^2 = \frac{1}{\eta} \text{Im} G^{(T)}_{ii}. \quad (8.3) \]

**Proof.** We will prove this lemma only for \( \eta = 0 \), the general case is a straightforward modification. We first consider the (8.2). Recall the resolvent expansion stating that for any two matrices \( A \) and \( B \),

\[ \frac{1}{A + B} = \frac{1}{A} - \frac{1}{A + B} B \frac{1}{A} = \frac{1}{A} - B \frac{1}{A + B}, \quad (8.4) \]

provided that all the matrix inverses exist.

To obtain the first formula in (8.2), we use the first resolvent identity (8.4) at the \( (ij) \)-th matrix element with \( A = H^{(i)} - z \) and \( B = H - H^{(i)} \). Since \( G^{(i)}_{ij} = (A^{-1})_{ij} = 0 \) if \( i \neq j \), we immediately have

\[ G_{ij} = -\sum_{k} G_{ii} h_{ik} G_{kj}^{(i)} - \sum_{k} G_{ik} h_{ki} G_{ij}^{(i)} = -G_{ii} \sum_{k \neq i} h_{ik} G_{kj}^{(i)}, \quad i \neq j. \quad (8.5) \]

The second identity in (8.2) follows in the same way by using the second identity in (8.4). To prove the identity (8.1), we let \( A = H^{(k)} - z \) and \( B = H - H^{(k)} \). Then from the first formula in (8.4) we have

\[ G_{ij} = G_{ij}^{(k)} - G_{ik} \sum_{\ell \neq k} h_{\ell j} G_{ij}^{(k)} = G_{ij}^{(k)} + \frac{G_{ik} G_{kj}}{G_{kk}}, \quad (8.6) \]
and in the second step we used (8.2).
Finally, the Ward identity (8.3), already used in (7.23), is well known. It follows from the spectral decomposition for $H$ with eigenvalues $\lambda_\alpha$ and eigenvectors $u_\alpha$, namely
\[
\sum_j |G_{ij}|^2 = \sum_j G_{ij} G_{ji}^* = |[G|^2]_{ii} = \sum_\alpha \left| u_\alpha(i) \right|^2 = \frac{1}{\eta} \text{Im} \sum_\alpha \frac{|u_\alpha(i)|^2}{\lambda_\alpha - z} = \frac{1}{\eta} \text{Im} G_{ii}.
\]

8.2 Self-consistent equations on two levels
Using the notation from the previous section, the Schur formula (Lemma 7.2) can be written as
\[
\frac{1}{G_{ii}^{(T)}} = h_{ii} - z - \sum_{k,l}^{(T_i)} h_{ik} G_{kl}^{(Ti)} h_{li},
\]
where $i \notin T \subset \{1, \ldots, N\}$. We can take $T = \emptyset$ to have
\[
G_{ii} = \frac{1}{h_{ii} - z - \sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li}}.
\]
The partial expectation with respect to the index $i$ in the denominator can be computed explicitly since $h_{ik} h_{li}$ is independent of $G_{kl}^{(i)}$ and its expectation is nonzero only if $k = l$. We thus get
\[
P_i \left[ \sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li} \right] = \sum_k s_{ik} G_{kk}^{(i)} = \sum_k s_{ik} G_{kk} - \sum_k s_{ik} G_{ik} G_{ki} G_{ii} = \sum_k s_{ik} G_{kk} - \sum_k s_{ik} G_{ik} G_{ki} G_{ii},
\]
where in the second step we used (8.1). We will compare (8.8) with the defining equation of $m_{sc}$:
\[
m_{sc} = 1 - z - m_{sc},
\]
so we introduce the notation for the difference
\[
v_i := G_{ii} - m_{sc}.
\]
Recalling (6.1), we get the following system of self-consistent equations for $v_i$
\[
v_i = \frac{1}{-z - m_{sc} - \left( \sum_k s_{ik} v_k - \Upsilon_i \right)} - m_{sc},
\]
where
\[
\Upsilon_i := A_i + h_{ii} - Z_i, \quad A_i := \sum_k s_{ik} G_{ik} G_{ki} G_{ii}, \quad Z_i := Q_i \left[ \sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li} \right].
\]
All these quantities depend on $z$, but we omit it in the notation. We will show that $\Upsilon$ is a small error term. This is clear about $h_{ii}$ by (6.27). The term $A_i$ will be small since off-diagonal resolvent entries are small. Finally, $Z_i$ will be small by a large deviation estimate (7.59) from Theorem 7.7.
Before we present more details, we heuristically show the power of this new system of self-consistent equations (8.10) and compare it with the single self-consistent equation used in Section 7 and, in fact, used in all previous literature on the resolvent method for Wigner matrices.
8.2.1 A scalar self-consistent equation

Introduce the notation

\[ \{a\} = \frac{1}{N} \sum_i a_i \] (8.12)

for the average of a vector \((a_i)_{i=1}^N\). Consider the standard Wigner case, \(s_{ij} = \frac{1}{N}\). Then

\[ \sum_k s_{ik} v_k = \frac{1}{N} \sum_k v_k = [v] \quad (= m_N - m_{sc}) \].

Neglecting \(\Upsilon_i\) in (8.10) and taking the average of this relation for each \(i = 1, 2, \ldots, N\), we get

\[ [v] \approx \frac{1}{-z - m_{sc} - m_{sc}} \approx \frac{1}{-z - m_{sc} - m_{sc}}. \] (8.13)

Recall the (8.9), the defining equation of \(m_{sc}\):

\[ m_{sc} = \frac{1}{-z - m_{sc}}. \]

Using Lemma 7.6, which asserts that this equation is stable under small perturbations, at least away from the spectral edges \(z = \pm 2\), we conclude \([v] \approx 0\) from (8.13). This means that \(m_N \approx m_{sc}\) and hence for the empirical density \(\varrho_N \approx \varrho_{sc}\), i.e., we obtained the Wigner’s original semicircle law. This is exactly the idea we used in Section 7 except that the interlacing argument is replaced by Lemma 8.3 this time.

8.2.2 A vector self-consistent equation

If we are interested in individual resolvent matrix elements \(G_{ii}\) instead of their average, \(\frac{1}{N} \text{Tr} G\), then the scalar equation (8.13) discussed in the previous section is not sufficient. We have to consider (8.10) as a system of equations for the components of the vector \(v = (v_1, \ldots, v_N)\). In order to analyze it, we will linearize this system of equations. There are two possible linearizations depending on how we expand the error terms.

Linearization I. If we know that \((\sum_k s_{ik} v_k - \Upsilon_i)\) is small, we can expand the denominator in (8.10) to have

\[ v_i = m_{sc}^2 (\sum_k s_{ik} v_k - \Upsilon_i) + m_{sc}^3 (\sum_k s_{ik} v_k - \Upsilon_i)^2 + O\left(\left(\sum_k s_{ik} v_k - \Upsilon_i\right)^3\right), \] (8.14)

where we have used the defining equation (8.9) for \(m_{sc}\), and used that \(|m_{sc}^4| \lesssim 1\) in the error term. After rearranging (8.14) we get

\[ [(1 - m_{sc}^2 S) v_i] = E_i := -m_{sc}^2 \Upsilon_i + m_{sc}^3 (\sum_k s_{ik} v_k - \Upsilon_i)^2 + O\left(\left(\sum_k s_{ik} v_k - \Upsilon_i\right)^3\right), \] (8.15)

Linearization II. If we know that \(v_i\) is small, then it is useful to rewrite (8.10) into the following form:

\[ -\sum_k s_{ik} v_k + \Upsilon_i = \frac{1}{m_{sc} + v_i} - \frac{1}{m_{sc}}, \] (8.16)

where we have also used the defining equation of \(m_{sc}\). If \(|v_i| = o(1)\), then we expand \(v_i\) to the second order and multiply both sides by \(m_{sc}^2\) to obtain

\[ [(1 - m_{sc}^2 S) v_i] = E_i := -m_{sc}^2 \Upsilon_i + \frac{1}{m_{sc}} v_i^2 + O(|v_i|^3), \] (8.17)

where we have estimated \(|m_{sc}^4| \geq c\) in the last term (see Lemma 6.2).
Notice that the definitions of $\mathcal{E}_i$ in (8.15) and (8.18) are different, although their leading behavior $-m_{sc}^2 \mathcal{Y}_i$ is the same. In both cases we can continue the analysis by inverting the operator $(1 - m_{sc}^2 S)$ to obtain

$$v = \frac{1}{1 - m_{sc}^2 S} \mathcal{E}, \quad \text{hence} \quad \|v\|_\infty \leq \left\| \frac{1}{1 - m_{sc}^2 S} \right\|_{\infty} \|\mathcal{E}\|_\infty,$$  \quad (8.18)

and this relation shows how the quantity $\Gamma$, defined in (6.17), emerges. If the error term is indeed small and $\Gamma$ is bounded, then we obtain that $\|v\|_\infty = \max |G_{ii} - m_{sc}|$ is small.

While the expansion logic behind equations (8.14) and (8.17) is the same and the resulting formulae are very similar, the structure of the main proof depends on which linearization of the self-consistent equation is used. In both cases we need to derive an a-priori bound to ensure that the expansion is valid. Intuitively, the smallness of $\sum_k s_{ik} v_k - \mathcal{Y}_i$ seems easier than that of $v_i$ since both terms are averaged quantities and extra averaging typically helps. However, these are random objects and every estimate comes with an exceptional set in the probability space where it does not hold. It turns out that on the technical level it is worth minimizing the bookkeeping of these events and this reason favors the second version of the linearization which operates with controlling a single quantity, $\max |v_i|$. In this book, therefore, we will follow the linearization (8.17).

We remark that the other option was used in [69, 70] which required first proving the weak semicircle law, Theorem 7.1, to provide the necessary a priori bound. The linearization (8.17) circumvents this step and the a priori bound on $v_i$ will be proved directly.

### 8.3 Proof of the local semicircle law without using the spectral gap

In this section we prove a restricted version of Theorem 6.7, namely we replace threshold $\eta$ with a larger threshold $\eta_E$ defined as

$$\eta_E := \min \left\{ \eta : \frac{1}{M \xi} \leq \min \left\{ \frac{M^{-\gamma}}{\Gamma(E + i \xi)^3}, \frac{M^{-2\gamma}}{\Gamma(E + i \xi)^4 \Im m_{sc}(E + i \xi)} \right\} \right\} \text{ holds for all } \xi \geq \eta \right\}. \quad (8.19)$$

This definition is exactly the same as (6.28), but $\Gamma$ is replaced with the larger quantity $\Gamma$, in other words we do not make use of the spectral gap in $S$. This will pedagogically simplify the presentation but it will prove the estimates in Theorem 6.7 only for the $\eta \geq \eta_E$ regime. In Section 9 we will give the proof for the entire $\eta \geq \eta_E$ regime. We recall Lemma 6.3 showing that there is no difference between $\Gamma$ and $\tilde{\Gamma}$ for generalized Wigner matrices away from the edges (both are of order 1), so readers interested in the local semicircle law only in the bulk should be content with the simpler proof. Near the spectral edges, however, there is a substantial difference. Note that even in the Wigner case (see (6.22)), $\eta_E$ is much larger near the spectral edges than the optimal threshold $\eta_E$. For generalized Wigner matrices, while $\eta_E \gg 1/N$, the threshold $\eta_E$ is determined by the relation $N \eta_E (\kappa_E + \eta_E)^{3/2} \gg 1$, where $\kappa_E = |E| - 2$ is the distance of $E$ from the spectral edges.

We stress, however, that the proof given below does not use any model-specific upper bound on $\Gamma$, such as (6.22) or (6.24), only the trivial lower and upper bounds, (6.19) and (6.20), are needed. The actual size of $\Gamma$ enters only implicitly by determining the threshold $\eta_E$. This makes the argument applicable to a wide class of problems beyond generalized Wigner matrices, including band matrices, see [55].

**Definition 8.4.** We call a deterministic nonnegative function $\Psi \equiv \psi^{(N)}(z)$ an admissible control parameter if we have

$$c M^{-1/2} \leq \Psi \leq M^{-c} \quad (8.20)$$

for some constant $c > 0$ and large enough $N$. Moreover, after fixing a $\gamma > 0$, we call any (possibly $N$-dependent) subset

$$D = D^{(N)} \subset \left\{ z : |E| \leq 10, M^{-1} \leq \eta \leq 10 \right\}$$

a spectral domain.
In this section we will mostly use the spectral domain
\[ S := \left\{ z : |E| \leq 10, \eta \in [\eta_E, 10] \right\}, \] (8.21)

where we note that
\[ \eta_E \geq \frac{1}{8} M^{-1+\gamma} \] (8.22)

using the lower bound \( \Gamma \geq c \), from (6.19), in the definition (8.19). Define the random control parameters
\[ \Lambda_o := \max_{i \neq j} |G_{ij}|, \quad \Lambda_d := \max_i |G_{ii} - m_{sc}|, \quad \Lambda := \max(\Lambda_o, \Lambda_d). \] (8.23)

where the letters \( d \) and \( o \) refer to diagonal and off-diagonal elements. In the typical regime we will work, all these quantities are small. The key quantity is \( \Lambda \) and we will develop an iterative argument to control it. We first derive an estimate of \( \Lambda_o + |Z_i| + |Y_i| \) in terms of \( \Lambda \). This will be possible only in the event when \( \Lambda \) is already small, so we will need to introduce an indicator function \( \phi = 1(\Lambda \leq M - c) \) with some small \( c \).

More generally, we will consider any indicator function \( \phi \) so that \( \phi \Lambda \prec M - c \). Notice that this is a somewhat weaker concept than \( \phi \Lambda \leq M - c \) (even if the exponent \( c \) is slightly adjusted), but it turns out to be more flexible since algebraic manipulations involving \( \prec \) (Proposition 6.5) can be directly used.

8.3.1 Large deviation bounds on \( \Lambda_o \) and \( Y_i \)

**Lemma 8.5.** The following statements hold for any spectral domain \( D \). Let \( \phi \) be the indicator function of some (possibly \( z \)-dependent) event. If \( \phi \lambda \prec M^{-c} \) for some \( c > 0 \) holds uniformly in \( z \in D \), then
\[ \phi(\Lambda_o + |Z_i| + |Y_i|) \prec \sqrt{\frac{\Im m_{sc} + \Lambda}{M\eta}}, \] (8.24)

uniformly in \( z \in D \). Moreover, for any fixed (\( N \)-independent) \( \eta > 0 \) we have
\[ \Lambda_o + |Z_i| + |Y_i| \prec M^{-1/2} \] (8.25)

uniformly in \( z \in \{ w \in D : \Im w = \eta \} \).

In other words, (8.24) means that
\[ \Lambda_o + |Z_i| + |Y_i| \prec \sqrt{\frac{\Im m_{sc} + \Lambda}{M\eta}}, \] (8.26)
on the event where \( \Lambda \prec M^{-c} \) has been apriori established.

**Proof of Lemma 8.5.** We first observe that \( \phi \lambda \prec M^{-c} \ll 1 \) and the positive lower bound \( |m_{sc}(z)| \geq c \) implies that
\[ \frac{\phi}{|G_{ii}|} \prec 1. \] (8.27)

A simple iteration of the expansion formula (8.1) concludes that
\[ \phi|G_{ij}^{(T)}| \prec M^{-c}, \quad \text{for } i \neq j, \quad \phi|G_{ii}^{(T)}| \prec 1, \quad \frac{\phi}{|G_{ii}^{(T)}|} \prec 1 \] (8.28)

for any subset \( T \) of fixed cardinality.

We begin with the first statement in Lemma 8.5. First we estimate \( Z_i \), which we split as
\[ \phi|Z_i| \leq \phi \left| \sum_k^{(i)} (|h_{ik}|^2 - s_{ik})G_{kk}^{(i)} \right| + \phi \left| \sum_{k \neq l}^{(i)} h_{ik}G_{kl}^{(i)}h_{li} \right|. \] (8.29)
We estimate each term using Theorem 7.7 by conditioning on \( G^{(i)} \) and using the fact that the family \((h_{ik})_{k=1}^N\) is independent of \( G^{(i)} \). By (7.57) the first term of (8.29) is stochastically dominated by

\[
\phi \left[ \sum_k s_{ik} |G_{kk}^{(i)}|^2 \right]^{1/2} < M^{-1/2},
\]

where (8.28), (6.3) and (6.1) were used. For the second term of (8.29) we apply (7.59) from Theorem 7.7 with \( a_{kl} = s_{ik}^{1/2} G_{kl} G_{lj} s_{lj}^{1/2} \) and \( X_k = s_{ik}^{-1/2} h_{ik} \). We find

\[
\phi \sum_{k,l} s_{ik} |G_{kl}^{(i)}|^2 s_{lj} \leq \frac{\phi}{M} \sum_{k,l} s_{ik} |G_{kl}^{(i)}|^2 = \frac{\phi}{M \eta} \sum_{k} s_{ik} \text{Im} G_{kk}^{(i)} < \frac{\text{Im} m_{sc} + \Lambda}{M \eta},
\]

where in the last step we used (8.1) and the estimate \( 1/G_{ii} < 1 \). Thus we get

\[
\phi |Z_i| < \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M \eta}},
\]

where we absorbed the bound \( M^{-1/2} \) on the first term of (8.29) into the right-hand side of (8.31). Here we only needed to use \( \text{Im} m_{sc}(z) \geq c \eta \) as follows from an explicit estimate, see (6.13).

Next, we estimate \( \Lambda_o \). We can iterate (8.2) once to get, for \( i \neq j \),

\[
G_{ij} = -G_{ii} \sum_k h_{ik} G_{kj}^{(i)} = -G_{ii} G_{jj}^{(i)} \left( h_{ij} - \sum_k h_{ik} G_{kj}^{(ij)} h_{lj} \right).
\]

The term \( h_{ij} \) is trivially \( O_{<}(M^{-1/2}) \). In order to estimate the other term, we invoke (7.58) from Theorem 7.7 with \( a_{kl} = s_{ik}^{1/2} G_{kl} G_{lj} s_{lj}^{1/2} \), \( X_k = s_{ik}^{-1/2} h_{ik} \), and \( Y_l = s_{lj}^{-1/2} h_{lj} \). As in (8.30), we find

\[
\phi \sum_{k,l} s_{ik} |G_{kl}^{(i)}|^2 s_{lj} \leq \frac{\phi}{M \eta} \sum_{k} s_{ik} \text{Im} G_{kk}^{(i)} < \frac{\text{Im} m_{sc} + \Lambda}{M \eta},
\]

and thus

\[
\phi \Lambda_o < \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M \eta}},
\]

where we again absorbed the term \( h_{ij} \prec M^{-1/2} \) into the right-hand side.

In order to estimate \( A_i \) and \( h_{ii} \) in the definition of \( Y_i \), we use (8.28) to get

\[
\phi |[A_i] + [h_{ii}]| \prec \phi \Lambda_o^2 + M^{-1/2} \leq \phi \Lambda_o + C \sqrt{\text{Im} m_{sc} M \eta} < \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M \eta}},
\]

where the second step follows from \( \text{Im} m_{sc} \geq c \eta \). Collecting (8.31), (8.33), this completes the proof of (8.24).

The proof of (8.25) is almost identical to that of (8.24). The quantities \( |G_{kk}^{(i)}| \) and \( |G_{kk}^{(ij)}| \) are estimated by the trivial deterministic bound \( \eta^{-1} = O(1) \). We omit the details. \( \square \)

**8.3.2 Initial bound on \( \Lambda \)**

In this subsection, we prove an initial bound asserting that \( \Lambda \prec M^{-1/2} \) holds for \( z \) with a large imaginary part \( \eta \). This bound is rather easy to get after we have proved in (8.25) that the error term \( Y_i \) in the self-consistent equation (8.16) is small.
Lemma 8.6. We have \( \Lambda \prec M^{-1/2} \) uniformly in \( z \in [-10, 10] + 2i \).

Proof. We shall make use of the trivial bounds
\[
|G_{ij}^{(T)}| \leq \frac{1}{\eta} = \frac{1}{2}, \quad |m_{sc}| \leq \frac{1}{\eta} = \frac{1}{2}, \tag{8.34}
\]
where the last inequality follows from the fact that \( m_{sc} \) is the Stieltjes transform of a probability measure. From (8.25) we get
\[
\Lambda_o + |Z_i| \prec M^{-1/2}. \tag{8.35}
\]
Moreover, we use (8.1) and (8.32) to estimate
\[
|A_i| \leq \sum_j s_{ij} |G_{ij}G_{ji}| \leq M^{-1} + \sum_j s_{ij} |G_{ji}G_{jj}| \left| h_{ij} - \sum_{k,l} h_{ik}G_{ij}^{(i)}h_{lj} \right| \prec M^{-1/2},
\]
where the last step follows by using (7.58), exactly as the estimate of the right-hand side of (8.32) in the proof of Lemma 8.5. We conclude that \( |Y_i| \prec M^{-1/2} \).

Next, we write (8.16) as
\[
v_i = \frac{m_{sc}(\sum_k s_{ik}v_k - Y_i)}{m_{sc}^{-1} - \sum_k s_{ik}v_k + Y_i}. \tag{8.36}
\]
Using \( |m_{sc}^{-1}| \geq 2 \) and \( |v_k| \leq 1 \) as follows from (8.34), we find
\[
\left| m_{sc}^{-1} + \sum_k s_{ik}v_k - Y_i \right| \geq 1 + O_\prec(M^{-1/2}).
\]
Using \( |m_{sc}| \leq 1/2 \) and taking the maximum over all \( i \) in (8.36), we therefore conclude that
\[
\Lambda_d \leq \frac{\Lambda_o + O_\prec(M^{-1/2})}{2 + O_\prec(M^{-1/2})} = \frac{\Lambda_o}{2} + O_\prec(M^{-1/2}), \tag{8.37}
\]
from which the claim follows together with the estimate on \( \Lambda_o \) from (8.35).

8.3.3 A rough bound on \( \Lambda \) and a continuity argument

The next step is to get a rough bound to \( \Lambda \) by a continuity argument.

Proposition 8.7. We have \( \Lambda \prec M^{-\gamma/3}\Gamma^{-1} \) uniformly in the domain \( S \) defined in (8.21).

Proof. The core of the proof is a continuity argument. The first task is to establish a gap in the range of \( \Lambda \) by establishing a dichotomy. Roughly speaking, the following lemma asserts that, for all \( z \in S \), with high probability either \( \Lambda \leq M^{-\gamma/2}\Gamma^{-1} \) or \( \Lambda \geq M^{-\gamma/4}\Gamma^{-1} \), i.e., there is a gap or forbidden region in the range of \( \Lambda \) with very high probability.

Lemma 8.8. We have the bound
\[
1(\Lambda \leq M^{-\gamma/4}\Gamma^{-1})\Lambda \prec M^{-\gamma/2}\Gamma^{-1}
\]
uniformly in \( S \).
Proof. Set
\[ \phi := 1(\Lambda \leq M^{-\gamma/4} \Gamma^{-1}). \]
Then by definition we have \( \phi \Lambda \leq M^{-\gamma/4} \Gamma^{-1} \leq CM^{-\gamma/4} \), where in the last step we have used that \( \Gamma \) is bounded below (6.19). Hence we may invoke (8.24) to estimate \( \Lambda_0 \) and \( \Upsilon_i \) by \( \sqrt{\frac{\text{Im } m_{sc} + \Lambda}{M \eta}} \). In order to estimate \( \Lambda_d \), we use (8.18) to get
\[ \phi \Lambda_d = \phi \max_i |v_i| \prec \Gamma \left( \Lambda^2 + \sqrt{\frac{\text{Im } m_{sc} + \Lambda}{M \eta}} \right). \] (8.38)
Recalling (6.19) and (8.24), we therefore get
\[ \phi \Lambda \prec \phi \Gamma \left( \Lambda^2 + \sqrt{\frac{\text{Im } m_{sc} + \Lambda}{M \eta}} \right). \] (8.39)

Next, by definition of \( \phi \) we may estimate
\[ \phi \Gamma \Lambda \leq M^{-\gamma/2} \Gamma^{-1}. \]
Moreover, by definition of \( S \), we have
\[ \frac{1}{M \eta} \leq \min \left\{ \frac{M^{-\gamma}}{\Gamma^3}, \frac{M^{-2\gamma}}{\Gamma^4 \text{Im } m_{sc}} \right\}. \] (8.40)
Together with the definition of \( \phi \), we have
\[ \phi \Gamma \sqrt{\frac{\text{Im } m_{sc} + \Lambda}{M \eta}} \leq \Gamma \sqrt{\frac{\text{Im } m_{sc}}{M \eta}} + \Gamma \sqrt{\frac{\Gamma^{-1}}{M \eta}} \leq M^{-\gamma \Gamma^{-1}} + M^{-\gamma/2 \Gamma^{-1}} \leq 2M^{-\gamma/2 \Gamma^{-1}}. \]
(Notice that the middle inequality is the crucial place where the definition of \( \eta_E \) and the restriction \( \eta \geq \eta_E \) are used.) Plugging this into (8.39) yields \( \phi \Lambda \prec M^{-\gamma/2 \Gamma^{-1}} \), which is the claim.

If we knew that \( \Lambda \) is excluded from the interval \([M^{-\gamma/2 \Gamma^{-1}}, M^{-\gamma/4 \Gamma^{-1}}]\), then we could immediately finish the proof of Proposition 8.7. We could argue that \( \Lambda = \Lambda(E + i\eta) \) is continuous in \( \eta = \text{Im } z \) and hence cannot jump from one side of the gap to the other; moreover, for \( \eta = 2 \) it is below the gap by Lemma 8.6, so \( \Lambda \) is below the gap for all \( z \in S \) with high probability. For a pictorial illustration of this argument, see Figure 1.

However, Lemma 8.8 guarantees a gap in the range of \( \Lambda \) only with a very high probability for each fixed \( z \). We need to use a fine discrete grid in the space of \( z \) to upgrade this statement to all \( z \) with high probability. In the next step we explain the details.

The continuity argument

Fix \( D > 10 \). Lemma 8.8 implies that for each \( z \in S \) the probability that \( \Lambda \) falls into the gap (or the forbidden region) is very small, i.e., we have
\[ \mathbb{P} \left( M^{-\gamma/3} \Gamma(z)^{-1} \leq \Lambda(z) \leq M^{-\gamma/4} \Gamma(z)^{-1} \right) \leq N^{-D} \] (8.41)
for \( N \geq N_0 \), where \( N_0 = N_0(\gamma, D) \) does not depend on \( z \). All argument below is valid for \( N \geq N_0 \).

Next, take a lattice \( \Delta \subset S \) such that \( |\Delta| \leq N^{10} \) and for each \( z \in S \) there exists a \( w \in \Delta \) such that \( |z - w| \leq N^{-4} \). Then (8.41) combined with a union bounds gives
\[ \mathbb{P} \left( \exists w \in \Delta : M^{-\gamma/3} \Gamma(w)^{-1} \leq \Lambda(w) \leq M^{-\gamma/4} \Gamma(w)^{-1} \right) \leq N^{-D+10}. \] (8.42)
From the definitions of $\Lambda(z)$, $\Gamma(z)$, and $S$ (recall (6.19)), we immediately find that $\Lambda$ and $\Gamma$ are Lipschitz continuous on $S$, with Lipschitz constant at most $M^2$. Hence (8.42) implies that

$$P\left( \exists z \in S : 2M^{-\gamma/3}\Gamma(z)^{-1} \leq \Lambda(z) \leq 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1} \right) \leq N^{-D+10},$$

i.e., a slightly smaller gap is present simultaneously for all $z \in S$ and not only for the discrete lattice $\Delta$. We conclude that there is an event $\Xi$ satisfying $P(\Xi) \geq 1 - N^{-D+10}$ such that, for each $z \in S$, either $1(\Xi)\Lambda(z) \leq 2M^{-\gamma/3}\Gamma(z)^{-1}$ or $1(\Xi)\Lambda(z) \geq 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1}$. Since $\Lambda$ is continuous and $S$ is by definition connected, we conclude that either

$$\forall z \in S : 1(\Xi)\Lambda(z) \leq 2M^{-\gamma/3}\Gamma(z)^{-1}$$

or

$$\forall z \in S : 1(\Xi)\Lambda(z) \geq 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1}.$$  

(8.43) or  

(8.44)

Here the bounds (8.43) and (8.44) each hold surely, i.e., for every realization of $\Lambda(z)$.

It remains to show that (8.44) is impossible. In order to do so, it suffices to show that there exists a $z \in S$ such that $\Lambda(z) < 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1}$ with probability greater than $1/2$. But this holds for any $z$ with $\text{Im} z = 2$, as follows from Lemma 8.6 and the bound $\Gamma \leq C\eta^{-1}$ (6.20). This concludes the proof of Proposition 8.7.

8.3.4 Fluctuation averaging lemma

Recall the definition of the small control parameter $\Pi$ from (6.30) and the definition of the average $[a]$ of a vector $(a_i)_{i=1}^N$ from (8.12). We have shown in Lemma 8.5 that $|\Upsilon_i| < \Pi$, but in fact the average $[\Upsilon]$ is one order better; it is $\Pi^2$. This is due to the fluctuation averaging phenomenon which we state as the following lemma. We will explain its proof and related earlier results in Section 10.

We shall perform the averaging with respect to a family of complex weights $T = (t_{ik})$ satisfying

$$0 \leq |t_{ik}| \leq M^{-1}, \quad \sum_k |t_{ik}| \leq 1.$$  

(8.45)
Typical example weights are \( t_{ik} = s_{ik} \) and \( t_{ik} = N^{-1} \). Note that in both of these cases \( T \) commutes with \( S \).

**Lemma 8.9** (Fluctuation averaging). Fix a spectral domain \( D \) and a deterministic control parameter \( \Psi \) satisfying (8.20). Let the weights \( T = (t_{ik}) \) satisfy (8.45).

(i) If \( \Lambda \prec \Psi \), then we have
\[
\sum_k t_{ik} Q_k G_{kk} = O_\prec(\Psi^2) .
\] (8.46)

(ii) If \( \Lambda_d \prec M^{-c} \) for some \( c > 0 \) and \( \Lambda_o \prec \Psi_o \) (with \( \Psi_o \) also satisfying (8.20)), then we have
\[
\sum_k t_{ik} Q_k \frac{1}{G_{kk}} = O_\prec(\Psi_o^2) .
\] (8.47)

(iii) Assume that \( T \) commutes with \( S \) and \( \Lambda \prec \Psi \). Then we have
\[
\sum_k t_{ik} v_k = O_\prec(\Gamma^2) .
\] (8.48)

If, additionally, we have
\[
\sum_k t_{ik} = 1, \quad \text{for all } i,
\] (8.49)

then
\[
\sum_k t_{ik} (v_k - [v]) = O_\prec((\bar{\Gamma}^2) \Psi^2) , \quad \text{for all } i,
\] (8.50)

where we defined \( v_i := G_{ii} - m \).

The estimates (8.46)–(8.48), and (8.50) are uniform in the index \( i \).

Notice that the last statement (8.50) involves the parameter \( \bar{\Gamma} \) instead of \( \Gamma \) indicating that the spectral gap of \( S \) is used in its proof. We now prove a simple corollary of the bound (8.47).

**Corollary 8.10.** Suppose that \( \Lambda_d \prec M^{-c} \) for some \( c > 0 \) and \( \Lambda_o \prec \Psi_o \) for some deterministic control parameter \( \Psi_o \) satisfying (8.20). Suppose that the weights \( T = (t_{ik}) \) satisfy (8.45). Then we have \( \sum_k t_{ak} Y_k = O_\prec(\Psi_o^2) \).

**Proof.** The claim easily follows from the Schur complement formula (8.7) written in the form
\[
Y_i = A_i + Q_i \frac{1}{G_{ii}} .
\]

We may therefore estimate \( \sum_k t_{ak} Y_k \) using the trivial bound \( |A_i| \prec \Psi_o^2 \) as well as the fluctuation averaging bound from (8.47).

### 8.3.5 The final iteration scheme

First note that Proposition 8.7 guarantees that \( \phi \equiv 1 \) may be chosen in Lemma 8.5, since the condition \( \phi \Lambda \prec M^{-c} \) is satisfied. Therefore Lemma 8.5 asserts that \( \Lambda_o \) is stochastically dominated by
\[
\sqrt{\frac{\text{Im} \, m_{ac} + \Lambda}{M \eta}} \leq M^{-\varepsilon} \Lambda + \sqrt{\frac{\text{Im} \, m_{ac}}{M \eta}} + M^\varepsilon \frac{M \eta}{M \eta} ,
\]
where we have used the Schwarz inequality. The next lemma, the main estimate behind the proof of Theorem 6.7, extends this estimate to bound \( \Lambda_o \) with the same quantity. Thus, roughly speaking, we can estimate \( \Lambda \) by \( M^{-\varepsilon} \Lambda \) plus a deterministic error term. This gives a recursive relation on the upper bound for \( \Lambda \) which will be the basic step of our iteration scheme.
Proposition 8.11. Let $\Psi$ be a control parameter satisfying

$$cM^{-1/2} \leq \Psi \leq M^{-\gamma/3}\Gamma^{-1} \quad (8.51)$$

and fix $\varepsilon \in (0, \gamma/3)$. Then on the domain $S$ we have the implication

$$\Lambda < \Psi \implies \Lambda < F(\Psi), \quad (8.52)$$

where we defined

$$F(\Psi) := M^{-\varepsilon} + \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M\eta}} + \frac{M^\varepsilon}{M\eta}. \quad (8.53)$$

Proof. Suppose that $\Lambda < \Psi$ for some deterministic control parameter $\Psi$ satisfying (8.51). We invoke Lemma 8.5 with $\phi = 1$ (recall the bound (6.19)) to get

$$\Lambda_0 + |Z_i| + |Y_i| < \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M\eta}} < \sqrt{\frac{\text{Im} m_{sc} + \Psi}{M\eta}}. \quad (8.53)$$

Next, we estimate $\Lambda_d$. Define the $z$-dependent indicator function

$$\psi := 1(\Lambda \leq M^{-\gamma/4}).$$

By (8.51), (6.19), and the assumption $\Lambda < \Psi$, we have $1 - \psi \prec 0$. On the event $\{\psi = 1\}$, (8.17) is rigorous and we get the bound

$$\psi|v_i| \leq C\psi \left| \sum_k s_{ik}v_k - Y_i \right| + C\psi\Lambda^2.$$ 

Using the fluctuation averaging estimate (8.48) to bound $\sum_k s_{ik}v_k$ and (8.53) to bound $Y_i$, we find

$$\psi|v_i| < \Gamma\Psi^2 + \sqrt{\frac{\text{Im} m_{sc} + \Psi}{M\eta}}, \quad (8.54)$$

where we used the assumption $\Lambda < \Psi$ and the lower bound from (6.19) so that $C\psi\Lambda^2 \leq \Gamma\Psi^2$. Since the set $\{\psi = 0\}$ has very small probability (using our notation, this is expressed by $1 - \psi \prec 0$), we conclude

$$\Lambda_d < \Gamma\Psi^2 + \sqrt{\frac{\text{Im} m_{sc} + \Psi}{M\eta}}, \quad (8.55)$$

which, combined with (8.53), yields

$$\Lambda < \Gamma\Psi^2 + \sqrt{\frac{\text{Im} m_{sc} + \Psi}{M\eta}}. \quad (8.56)$$

Using Schwarz inequality and the assumption $\Psi \leq M^{-\gamma/3}\Gamma^{-1}$ we conclude the proof. \hfill \Box

Finally, we complete the proof of Theorem 6.7. It is easy to check that, on the domain $S$, if $\Psi$ satisfies (8.51) then so does $F(\Psi)$. In fact this step is the origin of the somewhat complicated definition of $\eta_E$. We may therefore iterate (8.52). This yields a bound on $\Lambda$ that is essentially the fixed point of the map $\Psi \mapsto F(\Psi)$, which is given by $\Pi$, defined in (6.30), (up to the factor $M^\varepsilon$). More precisely, the iteration is started with $\Psi_0 := M^{-\gamma/3}\Gamma^{-1}$; the initial hypothesis $\Lambda \prec \Psi_0$ is provided by Proposition 8.7. For $k \geq 1$ we set $\Psi_{k+1} := F(\Psi_k)$. Hence from (8.52) we conclude that $\Lambda \prec \Psi_k$ for all $k$. Choosing $k := \lceil \varepsilon^{-1} \rceil$ yields

$$\Lambda < \sqrt{\frac{\text{Im} m_{sc} + \Psi}{M\eta}} + \frac{M^\varepsilon}{M\eta}.$$

59
Since $\varepsilon$ was arbitrary, we have proved that
\[ \Lambda \prec \Pi = \sqrt{\frac{\text{Im} m_{sc}(z)}{M\eta}} + \frac{1}{M\eta}, \tag{8.57} \]
which is (6.31).

What remains is to prove (6.32). On the set $\{\psi = 1\}$, we once again use (8.17) to get
\[ \psi m_{sc}^2 \left( - \sum_k s_k v_k + \Upsilon_i \right) = -\psi v_i + O(\psi \Lambda^2). \tag{8.58} \]

Averaging in (8.58) yields
\[ \psi m_{sc}^2 (-[v] + [\Upsilon]) = -\psi[v] + O(\psi \Lambda^2). \tag{8.59} \]

By (8.57) and (8.53) with $\Psi = \Pi$, we have $\Lambda + |\Upsilon_i| \prec \Pi$. Moreover, by Corollary 8.10 (with the choice of $t_{ik} = \frac{1}{N}$) we have $|[\Upsilon]| \prec \Pi$. Thus we get
\[ \psi[v] = m_{sc}^2 \psi[v] + O(\psi \Lambda^2). \]

Since $1 - \psi \prec 0$, we conclude that $[v] = m_{sc}^2[v] + O_\prec(\Pi^2)$. Therefore
\[ |[v]| \prec \frac{\Pi^2}{1 - m_{sc}^2} \leq \left( \frac{\text{Im} m_{sc}}{1 - m_{sc}^2} + \frac{1}{1 - m_{sc}^2 M\eta} \right) \frac{2}{M\eta} \leq \left( C + \frac{\Gamma}{M\eta} \right) \frac{2}{M\eta} \leq \frac{C}{M\eta}. \tag{8.60} \]

Here in the third step we used the elementary explicit bound $\text{Im} m_{sc} \leq C|1 - m_{sc}^2|$ from (6.12)–(6.13) and the bound $\Gamma \geq 1 - m_{sc}^2$, and from (6.21). Since $|m_{ni} - m_{sc}| = ||v||$, this concludes the proof of (6.32). The proof of (6.33) is exactly the same, just in the third inequality of (8.60) we may use the stronger bound $\text{Im} m_{sc} = \frac{1}{\sqrt{\kappa + \eta}}$ from (6.13) in the regime $|E| \geq 2$. This completes the proof of Theorem 6.7 in the entire regime $S$, i.e., for $\eta \geq \eta_E$.

8.3.6 Summary of the proof: a recapitulation

In order to highlight the main ideas again, we summarize the key steps leading to the proof of Theorem 6.7 restricted to the regime $\eta \geq \eta_E$. As a guiding principle we stress that the main control parameter in the proof is $\Lambda = \max_{ij} \{|G_{ii} - m_{sc}|, |G_{ij}|\}$. The main goal is to get a closed, self-improving inequality for $\Lambda$, at least with a very high probability. We needed the following ingredients:

1) A self consistent system of equations for the diagonal elements of the resolvent, written as
\[ -(Sv)_i + \Upsilon_i = \frac{1}{m_{sc} + v_i} - \frac{1}{m_{sc}}, \quad \text{where} \quad v_i = G_{ii} - m_{sc}. \tag{8.61} \]

Since $\Upsilon_i$ contains the resolvent $G^{(i)}$, we also need perturbation formulas connecting $G$ and $G^{(i)}$ to make these equations self-consistent.

2) A large deviation bound on the off-diagonal elements and on the fluctuation, i.e., on $\Lambda_\circ + |Z_i|$, as
\[ \Lambda_\circ + |Z_i| + |\Upsilon_i| \prec \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M\eta}}. \tag{8.62} \]

3) An initial estimate on $\Lambda$ for large $\eta$, where the a priori bounds $|G_{ij}| \leq \eta^{-1} \leq C$ are effective;

4) A dichotomy, showing that there is a forbidden region for $\Lambda$:
\[ 1(\Lambda \leq M^{-\gamma/4} \Gamma^{-1}) \Lambda \prec M^{-\gamma/2} \Gamma^{-1}; \]
5) A crude bound on $\Lambda$ down to small $\eta$ obtained from the dichotomy and the initial estimate via a continuity argument;

6) Application of the fluctuation averaging lemma, i.e., Lemma 8.9, to estimate $Sv$ in the self-consistent equation. Thus $Sv$ becomes of order $\Lambda^2$, i.e., one order higher than the trivial bound $|v_i| \leq \Lambda$. This “boost” exploits the cancellations in averages of weakly correlated centered random variables and it constitutes the crucial improvement over Section 7. Thus we can use (8.17) to have

$$v_i \leq \Upsilon_i + O(\Lambda^2). \quad (8.63)$$

7) Combination of the large deviation bound (8.62) for $\Upsilon_i$ with (8.63) provides an estimate on $\Lambda_d = \max_i |v_i|$ in terms of $\Lambda$ that is better than the trivial bound $\Lambda_d \leq \Lambda$. Together with the large deviation bound on $\Lambda_o$ in (8.62), we have a closed inequality for $\Lambda$:

$$\Lambda \leq M^{-\varepsilon} \Lambda + \sqrt{\frac{\text{Im} m_{ec}}{M \eta}} + \frac{M^\varepsilon}{M \eta}.$$

The error term $M^{-\varepsilon} \Lambda$ can be absorbed into the left side and we have proved the estimate on $\Lambda$ asserted in Theorem 6.7. In practice, the last inequality was formulated in terms of a control parameter and we used an iteration scheme.

8) Averaging the self-consistent equation (8.17) in order to obtain a stronger estimate on the average quantity $[v]$. Fluctuation averaging is used again, this time to the average of $\Upsilon$, so that it becomes one order higher. This concludes the proof of Theorem 6.7 for $\eta \geq \eta_E$. 

61
9 Sketch of the proof of the local semicircle law using the spectral gap

In Section 8.3 we proved the local semicircle law, Theorem 6.7, uniformly for $\eta \geq \eta_E$ instead of the larger regime $\eta \geq \tilde{\eta}_E$. Recall that for generalized Wigner matrices the two thresholds $\eta_E$ and $\tilde{\eta}_E$ are determined by the relation $N(\eta_E + \eta_E)^{3/2} \gg 1$ and $\tilde{\eta}_E \gg 1/N$. Hence these two thresholds coincide in the bulk but substantially differ near the edge. In this section we sketch the proof of Theorem 6.7 for any $\eta \geq \tilde{\eta}_E$, i.e., for the optimal domain for $\eta$. For the complete proof we refer to Section 6 of [55]. This section can be read independently of Section 8.3.

We point out that the difference between these two thresholds stems from the difference between $\Gamma$ and $\tilde{\Gamma}$, see (6.28) and (8.19). The bound $\Gamma$ on the norm of $(1 - m^2S)^{-1}$ entered the proof when the self-consistent equation (8.17) were solved. The key idea in this section is that we can use $\tilde{\Gamma}$ to solve the self-consistent equation (8.17) separately on the subspace of constants (the span of the vector $e$) and on its orthogonal complement $e^\perp$.

Step 1. We bound the control parameter $\Lambda = \Lambda(z)$ from (8.23) in terms of $\Theta := |m_N - m_{sc}|$. This is the content of the following lemma. From now on, we assume that $c \leq \tilde{\Gamma} \leq C$, which is valid for generalized Wigner matrices, see (6.24). This will simplify several estimates in the argument that follows; for the general case, we refer to [55].

Lemma 9.1. Define the $z$-dependent indicator function
\[
\phi := 1(\Lambda \leq M^{-\gamma/4})
\]
and the random control parameter
\[
q(\Theta) := \sqrt{\frac{\text{Im} m_{sc} + \Theta}{M\eta}} + \frac{1}{M\eta} , \quad \Theta = |m_N - m_{sc}|.
\]
Then we have
\[
\phi \Lambda \prec \Theta + q(\Theta).
\]

Proof. For the whole proof we work on the event $\{\phi = 1\}$, i.e., every quantity is multiplied by $\phi$. We consistently drop these factors $\phi$ from our notation in order to avoid cluttered expressions. In particular, we set $\Lambda \leq M^{-\gamma/4}$ throughout the proof.

We begin by estimating $\Lambda_o$ and $\Lambda_d$ in terms of $\Theta$. Recalling (6.19), we find that $\phi$ satisfies the hypotheses of Lemma 8.5, from which we get
\[
\Lambda_o + |\Upsilon_i| \prec r(\Lambda), \quad r(\Lambda) := \sqrt{\frac{\text{Im} m_{sc} + \Lambda}{M\eta}}.
\]
In order to estimate $\Lambda_d$, from (8.17), we have, on the event $\{\phi = 1\}$, that
\[
v_i - m_{sc}^2 \sum_k s_{ik}v_k = O_{\prec}(\Lambda^2 + r(\Lambda)) ;
\]
here we used the bound (9.4) on $|\Upsilon_i|$. Next, we subtract the average $N^{-1} \sum_i$ from each side to get
\[
(v_i - \langle v \rangle) - m_{sc}^2 \sum_k s_{ik}(v_k - \langle v \rangle) = O_{\prec}(\Lambda^2 + r(\Lambda)),
\]
where we used $\sum_k s_{ik} = 1$. Note that the average over $i$ of the left-hand side vanishes, so that the average of the right-hand side also vanishes. Hence the right-hand side is perpendicular to $e$. Inverting the operator $1 - m_{sc}^2S$ on the subspace $e^\perp$ and using $\tilde{\Gamma} \leq C$, therefore yields
\[
|v_i - \langle v \rangle| \prec \Lambda^2 + r(\Lambda) .
\]
Combining with the bound $\Lambda_o \prec r(\Lambda)$ from (9.4) and recalling that $\Theta = ||v||$, we therefore get

$$\Lambda \prec \Theta + \Lambda^2 + r(\Lambda).$$

(9.7)

By definition of $\phi$ we have $\Lambda^2 \leq M^{-\gamma/4}\Lambda$, so that the second term on the right-hand side of (9.7) may be absorbed into the left-hand side:

$$\Lambda \prec \Theta + r(\Lambda),$$

(9.8)

where we used the cancellation property (6.25). Using (9.8) and the Cauchy-Schwarz inequality, we get

$$r(\Lambda) \leq \sqrt{\text{Im} m_{sc} + \Lambda} \prec \sqrt{\text{Im} m_{sc} + \Theta} \leq \sqrt{\text{Im} m_{sc} + \Theta} + M^{-\varepsilon} r(\Lambda) + M^\varepsilon \frac{1}{M\eta},$$

for any $\varepsilon > 0$. Since $\varepsilon > 0$ can be arbitrarily small, we conclude that

$$r(\Lambda) \prec q(\Theta).$$

(9.9)

Clearly, (9.3) follows from (9.8) and (9.9).

\[\Box\]

**Step 2.** Having estimated $\Lambda$ in terms of $\Theta$, we can derive a closed relation involving only $\Theta$. More precisely, we will show that the following self-consistent equation holds:

$$\phi \left( (1 - m_{sc}^2)[v] - m_{sc}^{-1}[v]^2 \right) = \phi O_{\prec} (q(\Theta)^2 + M^{-\gamma/4} \Theta^2).$$

(9.10)

Considering the right hand side as a small error, we will view (9.10) as a small perturbation of a quadratic equation for $[v]$. Recalling that $\Theta = ||v||$, the error term can be determined self-consistently. Thus, up to the accuracy of the error terms, we can solve (9.10) for $[v]$.

We now give a heuristic proof for (9.10). There are two main issues which we ignore in this sketch. First, we work only on the event $\{\phi = 1\}$, i.e., we assume that an a-priori bound $\Lambda \ll 1$ has already been proved uniformly for all $\eta \geq \tilde{\eta}_E$. It will require a separate argument to show that the complement event $\{\phi = 0\}$ is negligible and in some sense this constitutes the essential part to extend the proof of Theorem 6.7 from $\eta \geq \eta_E$ to the larger regime $\eta \geq \tilde{\eta}_E$. Second, we will neglect certain subtleties of the $\prec$ relation. While most arithmetics involving $\prec$ works in the same way as the usual inequality relation $\leq$, the cancellation rule (6.25) requires the coefficient of $X$ on the right hand side to be small. We will disregard this requirement below and we will treat $\prec$ as the usual $\leq$.

Since we work on the event $\{\phi = 1\}$, it is understood that every quantity below is multiplied by $\phi$, but we will ignore this fact in the formulas. Recall from (8.17) that

$$v_i - m_{sc}^2 \sum_k s_{ik} v_k + m_{sc}^2 Y_i = m_{sc}^{-1} v_i^2 + O(\Lambda^3).$$

(9.11)

In order to take the average over $i$ and get a closed equation for $[v]$, we write, using (9.6),

$$v_i^2 = ([v] + v_i - [v])^2 = [v]^2 + 2[v](v_i - [v]) + O_{\prec} \left( (\Lambda^2 + r(\Lambda))^2 \right).$$

Plugging this back into (9.11) and taking the average over $i$ gives

$$1 - m_{sc}^2 [v] + m_{sc}^2 [Y] = m_{sc}^{-1} [v]^2 + O_{\prec} \left( \Lambda^3 + r(\Lambda)^2 \right).$$

By Corollary 8.10 with $t_{ik} = \frac{1}{N}$, we can estimate

$$[Y] \prec \Lambda_o^2 \prec r(\Lambda)^2 \prec q(\Theta)^2,$$

(9.12)
where we have used (8.24) and (9.9). Hence we have

\[(1 - m_{\text{sc}}^2)[v] - m_{\text{sc}}^{-1}[v]^2 = O\left(\Theta^3 + q(\Theta)^2\right),\]

where we have used (9.3). This shows (9.10).

**Step 3.** Finally, we need to solve the approximately quadratic relation (9.10) for \([v]\), keeping in mind that \(\Theta = |v|\). This will give the main estimates (6.32) and (6.33) in Theorem 6.7. The entry-wise version of the local semicircle law, (6.31), will then directly follow from (9.3) on the event \(\{\phi = 1\}\).

Recall the definition of \(q\) from (9.2) so that

\[q(\Theta)^2 \asymp \frac{\text{Im} m_{\text{sc}} + \Theta}{M\eta} + \frac{1}{(M\eta)^2},\]

and recall from Lemma 6.2 that

\[c \leq |m_{\text{sc}}(z)| \leq 1 - cn\eta, \quad |1 - m_{\text{sc}}^2(z)| \asymp \sqrt{\kappa + \eta}, \quad \text{Im} m_{\text{sc}}(z) = \begin{cases} \sqrt{\kappa + \eta} & \text{if } |E| \leq 2 \\ \eta & \text{if } |E| > 2. \end{cases} \tag{9.14}\]

In particular, the coefficient of \(\Theta^2 = |v|^2\) in the left hand side of (9.10) is a non vanishing constant, in the right hand side it is at most \(M^{-\gamma/4} \ll 1\), so the latter can be absorbed into the former. We can thus rewrite (9.10) as

\[[v]^2 - B|v| = D, \tag{9.15}\]

where

\[B = \frac{1 - m_{\text{sc}}^2 + O((M\eta)^{-1})}{m_{\text{sc}}^{-1} + O(M^{-\gamma/4})} \asymp 1 - m_{\text{sc}}^2 + O((M\eta)^{-1}), \quad |D| \leq \frac{\text{Im} m_{\text{sc}} + 1}{m_{\text{sc}}^{-1} + O(M^{-\gamma/4})} \asymp \frac{\text{Im} m_{\text{sc}}}{M\eta} + \frac{1}{(M\eta)^2}.\]

This quadratic equation has two solutions

\[[v] = \frac{B \pm \sqrt{B^2 + 4D}}{2}, \tag{9.16}\]

the correct one is selected by a continuity argument. Fix an energy \(|E| \leq 2\), consider \([v] = [v(E + i\eta)]\) as a function of \(\eta\) and look at the two solution branches (9.16) as continuous functions of \(\eta\).

It is easy to check that for \(\eta \gg 1\) large we have \(|B| \asymp 1\) and \(|D| \ll 1\), so the two solution are well separated, one of them is close to 0, the other one close to \(B\). The a priori bound \(|[v]| \leq \eta^{-1} \ll 1\) guarantees that the correct branch is the one near 0. A more careful analysis of (9.16) with the given coefficient functions \(B(\eta)\) and \(D(\eta)\) shows that on this branch the following inequality holds:

\[\Theta = |[v]| \lesssim (M\eta)^{-1}. \tag{9.17}\]

To see this more precisely, we distinguish two cases by reducing \(\eta\) from a very large value to 0. In the first regime, where \(M\eta\sqrt{\kappa + \eta} \gg 1\), we have \(|B| \asymp \sqrt{\kappa + \eta}\) and \(|D| \lesssim \sqrt{\kappa + \eta}/M\eta \ll |B|^2\), so the two branches remain separated and the relevant branch satisfies \(\Theta \lesssim |D/B| \lesssim (M\eta)^{-1}\). As we decrease \(\eta\) and enter the regime \(M\eta\sqrt{\kappa + \eta} \lesssim 1\) both solutions satisfy \(\Theta \lesssim |B| + |D|\). In this regime \(|B| \lesssim (M\eta)^{-1}\) and \(|D| \lesssim (M\eta)^{-2}\), so we also have (9.17). In both cases we conclude the final bound (6.32). The proof of (6.33) for the regime \(|E| > 2\) is similar. The improvement comes from the fact that in this regime (9.14) gives a better bound for \(\text{Im} m_{\text{sc}}(z)\).
10 Fluctuation averaging mechanism

10.1 Intuition behind the fluctuation averaging

Before starting the proof of Lemma 8.9 we explain why it is so important in our context. We also give a heuristic second moment calculation to show the essential mechanism.

The leading error in the self-consistent equation (8.58) for \( v_i \) is \( \Upsilon_i \). Among the three summands of \( \Upsilon_i \), see (8.11), typically \( Z_i \) is the largest, thus we typically have

\[
v_i \approx Z_i
\]

in the regime where \( \Gamma \) is bounded. The large deviation bounds in Theorem 7.7 show that \( Z_i \approx \Lambda_o \) and a simple second moment calculation shows that this bound is essentially optimal. On the other hand, (8.3) shows that the typical size of the off-diagonal resolvent matrix elements is at least of order \( (N\eta)^{-1/2} \), thus the estimate

\[
Z_i \approx \Lambda_o \lesssim \frac{1}{\sqrt{N\eta}}
\]

is essentially optimal in the generalized Wigner case \((M \asymp N)\). Together with (10.1) this shows that the natural bound for \( \Lambda \) is \((N\eta)^{-1/2}\), which is also reflected in the bound (6.31).

However, the bound (6.32) for the average, \( m_N - m = [v] \), is of order \( \Lambda^2 \approx (N\eta)^{-1} \), i.e., it is one order better than the bound for \( v_i \). For the purpose of estimating \([v]\), it is the average \([\Upsilon] \) of the leading errors \( \Upsilon_i \) that matters, see (8.59). Since \( Z_i \), the leading term in \( \Upsilon_i \), is a fluctuating quantity with zero expectation, the improvement comes from the fact that fluctuations cancel out in the average. The basic example of this phenomenon is the central limit theorem; the average of \( N \) independent centered random variables is typically of order \( N^{-1/2} \), much smaller than the size of each individual variable. In our case, however, \( Z_i \) are not independent. In fact, their correlations do not decay, at least in the Wigner case where the column of \( H \) with an index other than \( i \) plays the same role in the function \( Z_i \). Thus standard results on central limit theorems for weakly correlated random variables do not apply here and a new argument is needed. We first give an idea how the fluctuation averaging mechanism works by a second moment calculation.

Second moment calculation. First we claim that

\[
\left| Q_k G_{kk}^{-1} \right| \lesssim \Psi_o.
\]

Indeed, from the Schur complement formula (8.7) we get \(|Q_k(G_{kk})^{-1}| \lesssim |h_{kk}| + |Z_k|\). The first term is estimated by \(|h_{kk}| \lesssim M^{-1/2} \lesssim \Psi_o\). The second term is estimated exactly as in (8.29) and (8.30), giving \(|Z_k| \lesssim \Psi_o\). In fact, the same bound holds if \( G_{kk} \) is replaced with \( G_{kk}^{(T)} \) as long as the cardinality of \( \mathcal{T} \) is bounded, see (10.10) later.

Abbreviate \( X_k := Q_k(G_{kk}^{-1}) \) and compute the variance

\[
\mathbb{E} \left| \sum_k t_{ik} X_k \right|^2 = \sum_{k,l} t_{ik} t_{il} \mathbb{E} X_k X_l = \sum_k t_{ik}^2 \mathbb{E} X_k X_k + \sum_{k \neq l} t_{ik} t_{il} \mathbb{E} X_k X_l.
\]

Using the bounds (8.45) on \( t_{ik} \) and (10.2), we find that the first term on the right-hand side of (10.3) is \( O_{\omega}(M^{-1}\Psi_o^2) = O_{\omega}(\Psi_o^2) \), where we used that \( \Psi_o \) is admissible, recalling (8.20). Let us therefore focus on the second term of (10.3). Using the fact that \( k \neq l \), we apply the first resolvent decoupling formula (8.1) to \( X_k \) and \( X_l \) to get

\[
\mathbb{E} X_k X_l = \mathbb{E} Q_k \left( \frac{1}{G_{kk}} \right) Q_l \left( \frac{1}{G_{ll}} \right) = \mathbb{E} Q_k \left( \frac{1}{G_{kk}} - \frac{G_{kl} G_{lk}}{G_{kk} G_{kl}} \right) Q_l \left( \frac{1}{G_{ll}} - \frac{G_{lk} G_{kl}}{G_{ll} G_{kk}} \right).
\]

Notice that we used (8.1) in the form

\[
\frac{1}{G_{kk}^{(T)}} = \frac{1}{G_{kk}^{(T)} G_{kk}^{(T)}} - \frac{G_{kl}^{(T)} G_{lk}^{(T)}}{G_{kk}^{(T)} G_{kk}^{(T)} G_{ll}^{(T)}} \quad \text{for any } k \neq l, k, l \notin \mathcal{T},
\]

65
Suppose that the deterministic control parameter $\Psi$ satisfies $\Psi \geq N^{-C}$, and that for all $p$ there is a constant $C_p$ such that the nonnegative random variable $X$ satisfies $\mathbb{E}X^p \leq N^{C_p}$. Suppose moreover that $X \prec \Psi$. Then for any fixed $n \in \mathbb{N}$ we have

$$\mathbb{E}X^n \prec \Psi^n. \quad (10.6)$$

(Note that this estimate involves deterministic quantities only, i.e., it means that $\mathbb{E}X^n \leq N^\varepsilon \Psi^n$ for any $\varepsilon > 0$ if $N \geq N_0(n, \varepsilon)$.) Conversely, if (10.6) holds for any $n \in \mathbb{N}$, then $X \prec \Psi$. Moreover, we have

$$P_iX^n \prec \Psi^n, \quad Q_iX^n \prec \Psi^n \quad (10.7)$$

uniformly in $i$. If $X = X(u)$ and $\Psi = \Psi(u)$ depend on some parameter $u$ and the above assumptions are uniform in $u$, then so are the conclusions.

This lemma is a generalization of (6.26), but notice that the majoring quantity $\Psi$ has to be deterministic. For general random variables, $X \prec Y$ may not imply the analogous relation $\mathbb{E}[X|\mathcal{F}] \prec \mathbb{E}[Y|\mathcal{F}]$ for conditional expectations with respect to a $\sigma$-algebra $\mathcal{F}$.

**Proof of Lemma 10.1.** To prove (10.6), it is enough to consider the case $n = 1$; the case of larger $n$ follows immediately from the case $n = 1$, using the basic property of the stochastic domination that $X \prec \Psi$ implies $X^n \prec \Psi^n$.

For the first claim, pick $\varepsilon > 0$. Then

$$\mathbb{E}X = \mathbb{E}X \cdot 1(X \leq N^\varepsilon \Psi) + \mathbb{E}X \cdot 1(X > N^\varepsilon \Psi) \leq N^\varepsilon \Psi + \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{P}(X > N^\varepsilon \Psi)} \leq N^\varepsilon \Psi + N^{C_2/2-D/2},$$

for arbitrary $D > 0$. The first claim therefore follows by choosing $D$ large enough. For the converse statement, we use Chebyshev inequality: for any $\varepsilon > 0$ and $D > 0$,

$$\mathbb{P}(X > N^\varepsilon \Psi) \leq \frac{\mathbb{E}X^n}{N^{\varepsilon n} \Psi^n} \leq N^{\varepsilon-n} \leq N^{-D},$$

by choosing $n$ large enough.

Finally, the claim (10.7) follows from Chebyshev’s inequality, using a high-moment estimate combined with Jensen’s inequality for partial expectation. We omit the details, which are similar to those of the first claim.

We remark that it is tempting to generalize the converse of (10.6) and try to verify stochastic domination $X \prec Y$ between two random variables by proving that $\mathbb{E}X^n \leq \mathbb{E}Y^n$. This implication in general is wrong, but it is correct if $Y$ is deterministic. This subtlety is the reason why many of the following theorems about
two random variables $A$ and $B$ are formulated in such a way that “if $A$ satisfies $A \prec \Psi$ for some deterministic $\Psi$ then, $B \prec \Psi$”, instead of directly saying the more natural (but sometimes wrong) assertion that $A \prec B$.

We shall apply Lemma 10.1 to the resolvent entries of $G$. In order to verify its assumptions, we record the following bounds.

**Lemma 10.2.** Suppose that $\Lambda \prec \Psi$ and $\Lambda_o \prec \Psi_o$ for some deterministic control parameters $\Psi$ and $\Psi_o$ both satisfying (8.20). Fix $p \in \mathbb{N}$. Then for any $i \neq j$ and $T \subset \{1, \ldots, N\}$ satisfying $|T| \leq p$ and $i, j \notin T$ we have

$$G^{(T)}_{ij} = O_{\prec}(\Psi_o), \quad \frac{1}{G^{(T)}_{ii}} = O_{\prec}(1).$$

(10.8)

Moreover, we have the rough bounds $|G^{(T)}_{ij}| \leq M$ and

$$E \left| \frac{1}{G^{(T)}_{ii}} \right|^n \leq N^\varepsilon$$

(10.9)

for any $\varepsilon > 0$ and $N \geq N_0(n, \varepsilon)$.

**Proof.** The bounds (10.8) follow easily by a repeated application of (8.1), the assumption $\Lambda \prec M^{-c}$, and the lower bound in (6.11). The deterministic bound $|G^{(T)}_{ij}| \leq M$ follows immediately from $\eta \geq M^{-1}$ by definition of a spectral domain.

In order to prove (10.9), we use the Schur complement formula (8.7) applied to $1/G^{(T)}_{ii}$, where the expectation is estimated using (5.6) and $|G^{(T)}_{ij}| \leq M$. (Recall (6.4).) This gives

$$E \left| \frac{1}{G^{(T)}_{ii}} \right|^p \prec N^{C_p}$$

for all $p \in \mathbb{N}$. Since $1/G^{(T)}_{ii} \prec 1$, (10.9) therefore follows from (10.6).

We now start the proof of Lemma 8.9. The main statement is (8.47), the other bounds will be relatively simple consequences. Finally, we present an alternative argument for (8.47) that organizes the stopping rule in the expansion somewhat differently.

**Proof of Lemma 8.9. Part I: Proof of (8.47).** First we claim that, for any fixed $p \in \mathbb{N}$, we have

$$\left| Q_k \frac{1}{G^{(k)}_{kk}} \right| \prec \Psi_o$$

(10.10)

uniformly for $T \subset \{1, \ldots, N\}$, $|T| \leq p$, and $k \notin T$. To simplify notation, for the proof we set $T = \emptyset$; the proof for nonempty $T$ is the same. From the Schur complement formula (8.7) we get $|Q_k(G_{kk})^{-1}| \leq |h_{kk}| + |Z_k|$. The first term is estimated by $|h_{kk}| \prec M^{-1/2} \prec \Psi_o$. The second term is estimated exactly as in (8.29) and (8.30):

$$|Z_k| \prec \left( \sum_{x \neq y} s_{xk} |G^{(k)}_{xy}|^2 s_{yk} \right)^{1/2} \prec \Psi_o,$$

where in the last step we used that $|G^{(k)}_{xy}| \prec \Psi_o$ as follows from (10.8), and the bound $1/|G_{kk}| \prec 1$ (recall that $\Lambda \prec \Psi \leq M^{-c}$). This concludes the proof of (10.10).

Abbreviate $X_k := Q_k(G_{kk})^{-1}$. We shall estimate $\sum_k t_k X_k$ in probability by estimating its $p$-th moment by $\Psi_o^p$, from which the claim (8.47) will easily follow using Chebyshev’s inequality.

After this preparation, the rest of the proof for (8.47) can be divided into four steps.
Step 1: Coincidence structure in the expansion of the $L^p$ norm. Fix some even integer $p$ and write

$$\mathbb{E}\left| \sum_k t_{ik} X_k \right|^p = \sum_{k_1, \ldots, k_p} t_{i k_1} \cdots t_{i k_p} \overline{t}_{i k_p / z} \cdots \overline{t}_{i k_p / z + 1} \mathbb{E} X_{k_1} \cdots X_{k_p / z} \overline{X}_{k_p / z + 1} \cdots \overline{X}_{k_p}.$$ 

Next, we regroup the terms in the sum over $k := (k_1, \ldots, k_p)$ according to the coincidence structure in $k$ as follows: given a sequence of indices $k$, define the partition $\mathcal{P}(k)$ of $\{1, \ldots, p\}$ by the equivalence relation $r \sim s$ if and only if $k_r = k_s$. Denote the set of all partitions of $\{1, \ldots, p\}$ by $\mathcal{P}_p$. Then we write

$$\mathbb{E}\left| \sum_k t_{ik} X_k \right|^p = \sum_{\mathcal{P} \in \mathcal{P}_p} \sum_k t_{i k_1} \cdots t_{i k_p} \overline{t}_{i k_p / z} \cdots \overline{t}_{i k_p / z + 1} \mathbb{E} X_{k_1} \cdots X_{k_p / z} \overline{X}_{k_p / z + 1} \cdots \overline{X}_{k_p} 1(\mathcal{P}(k) = P)V(k),$$

where we defined

$$V(k) := \mathbb{E} X_{k_1} \cdots X_{k_p / z} \overline{X}_{k_p / z + 1} \cdots \overline{X}_{k_p}.$$ 

Given a partition $P$, for any $r \in \{1, \ldots, p\}$, we denote by $[r]$ the block of $r$ in $P$, i.e., the set of all indices in the same block of the partition as $r$. Let $L \equiv L(P) := \{r : [r] = \{r\}\} \subset \{1, \ldots, p\}$ be the set of lone labels. Denote by $k_L := (k_r)_{r \in L}$ the summation indices associated with lone labels.

Step 2: Resolution of dependence in weakly dependent random variables. The resolvent entry $G_{kk}$ depends strongly on the randomness in the $k$-column of $H$, but only weakly on the randomness in the other columns. We conclude that if $r$ is a lone label then all factors $X_{k_r}$ with $s \neq r$ in $V(k)$ depend weakly on the randomness in the $k_r$-th column of $H$ (if $r$ is not a lone label then this statement holds only for “all factors $X_{k_r}$ with $k_s \neq k_r$”). Thus, the idea is to make all resolvent entries inside the expectation of $V(k)$ as independent of the indices $k_L$ as possible (see Definition 7.3), using the first decoupling resolvent identity (8.1): for $x, y, u \notin T$ and $x, y \neq u$ ($x$ can be equal to $y$),

$$G_{xy}^{(T)} = G_{xy}^{(\gamma u)} + \frac{G_{xx}^{(T)} G_{xy}^{(T)}}{G_{ux}^{(T)}}$$

and for $x, u \notin T$ and $x \neq u$

$$\frac{1}{G_{xx}^{(T)}} = \frac{1}{G_{xx}^{(\gamma u)}} \frac{G_{xu}^{(T)} G_{uu}^{(T)}}{G_{xx}^{(T)} G_{xx}^{(\gamma u)} G_{uu}^{(T)}}.$$ 

Definition 10.3. A resolvent entry $G_{xy}^{(T)}$ with $x, y \notin T$ is maximally expanded with respect to a set $B \subset \{1, \ldots, N\}$ if $B \subset \mathbb{T} \cup \{x, y\}$.

Given the set $k_L$ of lone indices, we say that a resolvent entry $G_{xy}^{(T)}$ is maximally expanded if it is maximally expanded with respect to the set $B = k_L$. The motivation behind this definition is that using (8.1) we cannot add upper indices from the set $k_L$ to a maximally expanded resolvent entry. We shall apply (8.1) to all resolvent entries in $V(k)$. In this manner we generate a sum of monomials consisting of off-diagonal resolvent entries and inverses of diagonal resolvent entries. We can now repeatedly apply (8.1) to each factor until either they are all maximally expanded or a sufficiently large number of off-diagonal resolvent entries has been generated. The cap on the number of off-diagonal entries is introduced to ensure that this procedure terminates after a finite number of steps.

In order to define the precise algorithm, let $\mathcal{A}$ denote the set of monomials in the off-diagonal entries $G_{xy}^{(T)}$, with $\mathcal{T} \subset k_L$, $x \neq y$, and $x, y \in k \setminus \mathcal{T}$, as well as the inverse diagonal entries $1/G_{xx}^{(T)}$, with $\mathcal{T} \subset k_L$ and $x \in k \setminus \mathcal{T}$. Starting from $V(k)$, the algorithm will recursively generate sums of monomials in $\mathcal{A}$. Let $\mathcal{d}(A)$ denote the number of off-diagonal entries in $A \in \mathcal{A}$. For $A \in \mathcal{A}$ we shall define $w_0(A)$, $w_1(A) \in \mathcal{A}$ satisfying

$$A = w_0(A) + w_1(A), \quad \mathcal{d}(w_0(A)) = \mathcal{d}(A), \quad \mathcal{d}(w_1(A)) \geq \max\{2, \mathcal{d}(A) + 1\}. \quad (10.14)$$ 

The idea behind this splitting is to use (8.1) on one entry of $A$; the first term on the right-hand side of (8.1) gives rise to $w_0(A)$ and the second to $w_1(A)$. The precise definition of the algorithm applied to $A \in \mathcal{A}$ is as follows.
(1) If all factors of \( A \) are maximally expanded or \( d(A) \geq p + 1 \) then stop the expansion of \( A \).

(2) Otherwise choose some (arbitrary) factor of \( A \) that is not maximally expanded. If this entry is off-diagonal, \( G_{xy}^{(T)} \), we use (10.12) to decompose \( G_{xy}^{(T)} \) into a sum of two terms with \( u \) the smallest element in \( k \setminus (T \cup \{x, y\}) \). If the chosen entry is diagonal, \( 1/G_{xx}^{(T)} \), we use (10.13) to decompose \( G_{xx}^{(T)} \) into two terms with \( u \) the smallest element in \( k \setminus (T \cup \{x\}) \). The choice of \( u \) to be the smallest element is not important, we just chose it for definiteness of the algorithm. From the splitting of the factor \( G_{xy}^{(T)} \) or \( 1/G_{xx}^{(T)} \) in the monomial \( A \), we obtain a splitting \( A = w_0(A) + w_1(A) \) (i.e., we replace \( G_{xy}^{(T)} \) or \( 1/G_{xx}^{(T)} \) by the right-hand sides of (10.12) or (10.13)).

It is clear that (10.14) holds with the algorithm just defined. In fact, in most cases the last inequality in (10.14) is an equality, \( d(w_1(A)) = \max\{2, d(A) + 1\} \). The only exception is when (10.12) is used for \( x = y \), then two new off-diagonal entries are added, i.e., \( d(w_1(A)) = d(A) + 2 \). Notice also that this algorithm contains some arbitrariness in the choice of the factor of \( A \) to be expanded but any choice will work for the proof.

We now apply this algorithm recursively to each entry \( A' := 1/G_{k,k} \) in the definition of \( V(k) \). More precisely, we start with \( A' \) and define \( A'_0 := w_0(A') \) and \( A'_1 := w_1(A') \). In the second step of the algorithm we define four monomials

\[
A'_{00} := w_0(A'_0), \quad A'_{01} := w_0(A'_1), \quad A'_{10} := w_1(A'_0), \quad A'_{11} := w_1(A'_1),
\]

and so on, at each iteration performing the steps (1) and (2) on each new monomial independently of the others. Note that the lower indices are binary sequences that describe the recursive application of the operations \( w_0 \) and \( w_1 \). In this manner we generate a binary tree whose vertices are given by finite binary strings \( \sigma \). The associated monomials satisfy \( A'_{\sigma i} := w_i(A'_{\sigma'}) \) for \( i = 0, 1 \), where \( \sigma i \) denotes the binary string obtained by appending \( i \) to the right end of \( \sigma \). See Figure 2 for an illustration of the tree. Notice that any monomial that is obtained along this recursion is of the form

\[
G_{x_1y_1}^{(T_1)}G_{x_2y_2}^{(T_2)}\cdots \frac{G_{x_1y_i}^{(T_i)}}{G_{x_2y_i}^{(T_2)}}G_{x_3y_i}^{(T_3)}\cdots,
\]

where all factors in the numerators are off-diagonal entries \( x_i \neq y_i \) and all factors in the denominators are diagonal entries. For later usage, we shall refer to the sum

\[
|T_1| + |T_2| + \ldots + |T_1'| + |T_2'| + \ldots
\]

as the \textit{total number of upper indices} in the monomial (10.15). (The absolute value denotes the cardinality of the set.)

We stop the recursion of a tree vertex whenever the associated monomial satisfies the stopping rule of step (1). In other words, the set of leaves of the tree is the set of binary strings \( \sigma \) such that either all factors of \( A'_{\sigma} \) are maximally expanded or \( d(A'_{\sigma}) \geq p + 1 \).

We list a few obvious facts of this algorithm:

(i) \( d(A'_{\sigma}) \leq p + 1 \) for any vertex \( \sigma \) of the tree (by the stopping rule in (1)).

(ii) the number of ones in any \( \sigma \) is at most \( p \) (since each application of \( w_1 \) increases \( d(\cdot) \) by at least one).

(iii) the number of resolvent entries (in the numerator and denominator) in \( A'_{\sigma} \) is bounded by \( 4p + 1 \) (since each application of \( w_1 \) increases the number of resolvent entries by at most four, and the application of \( w_0 \) does not change this number).

(iv) The maximal total number of upper indices in \( A'_{\sigma} \) for any tree vertex \( \sigma \) is \( (4p + 1)p \) (since \( |T_i| \leq p \) and \( |T'_i| \leq p \) in the formula (10.16)).

(v) \( \sigma \) contains at most \( (4p + 1)p \) zeros (since each application of \( w_0 \) increases the total number of upper indices by one).
Figure 2: The binary tree generated by applying the algorithm (1)–(2) to a monomial $A^r$. Each vertex of the tree is indexed by a binary string $\sigma$, and encodes a monomial $A^r_\sigma$. An arrow towards the left represents the action of $w_0$ and an arrow towards the right the action of $w_1$. The monomial $A^r_{11}$ satisfies the assumptions of step (1), and hence its expansion is stopped, so that the tree vertex 11 has no children.

(vi) The maximal length of the string $\sigma$ (i.e., the depth of the tree) is at most $(4p^2 + 1)p + 4p^2 + 2p$.

(vii) The number of leaves of a tree is bounded by $(Cp^2)^p$ (since this number is bounded by $\sum_{k=0}^{p} (4p^2 + 2p) \leq (Cp^2)^p$ where $k$ is the number of ones in the string encoding the leaf). Denoting by $L_r$ the set of leaves of the binary tree generated from $A^r$, we have $|L_r| \leq (Cp^2)^p$. In particular, the algorithm terminates after at most $(Cp^2)^p$ steps since each step generates a new leaf.

By definition of the tree and $w_0$ and $w_1$, we have the following resolution of dependence decomposition

$$X_k^r = Q_k^r \sum_{\sigma \in L_r} A^r_\sigma,$$

where each monomial $A^r_\sigma$ for $\sigma \in L_r$ either consists entirely of maximally expanded resolvent entries or satisfies $d(A^r_\sigma) = p + 1$. (This is an immediate consequence of the stopping rule in (1)). Using (10.11) and (10.17) we have the representation

$$V(k) = \sum_{\sigma_1 \in L_1} \cdots \sum_{\sigma_p \in L_p} \mathbb{E}(Q_k^1 A^r_{\sigma_1}) \cdots (Q_k^p A^r_{\sigma_p}).$$

(10.18)

Step 3. Bounding the individual terms in the expansion (10.18). We now claim that any nonzero term on the right-hand side of (10.18) satisfies

$$\mathbb{E}(Q_k^1 A^r_{\sigma_1}) \cdots (Q_k^p A^r_{\sigma_p}) = O_{<}(\Psi_o^{p+|L|}).$$

(10.19)

Proof of (10.19). Before embarking on the proof, we explain its idea. First notice that for any string $\sigma$ we have

$$A^r_\sigma = O_{<}(\Psi_o^{b(\sigma)+1}),$$

(10.20)

where $b(\sigma)$ is the number ones in the string $\sigma$. Indeed, if $b(\sigma) = 0$ then this follows from (10.10); if $b(\sigma) \geq 1$ this follows from the last statement in (10.14) which guarantees that every one in the string $\sigma$ increases the exponent of $\Psi_o$ by at least one (we also use (10.8)). In particular, each $A^r_\sigma$ is bounded by at least $\Psi_o$.

If we used only the trivial bound $A^r_\sigma = O_{<}(\Psi_o)$ for each factor in (10.19), i.e., we did not exploit the gain from the $w_1(A)$ type terms in the expansion, then the naive size of the left-hand side of (10.19) would only be $\Psi_o^{|L|}$. The key observation behind (10.19) is that each lone label $s \in L$ yields one extra factor $\Psi_o$ to the estimate. This is because the expectation in (10.18) would vanish if all other factors $(Q_k^r A^r_{\sigma_r}), r \neq s$, etc.
were independent of \( k_s \). The expansion of the binary tree makes this dependence explicit by expressing \( k_s \) as a lower index. But this requires performing an operation \( w_1 \) with the choice \( u = k_s \) in (10.12) or (10.13). However, \( w_1 \) increases the number of off-diagonal element by at least one. In other words, every index associated with a lone label must have a “partner” index in a different resolvent entry which arose by application of \( w_1 \). Such a partner index may only be obtained through the creation of at least one off-diagonal resolvent entry. The actual proof below shows that this effect applies cumulatively for all lone labels.

In order to give the rigorous proof of (10.19), we consider two cases. Consider first the case where for some \( r = 1, \ldots, p \) the monomial \( A_r^{\sigma_r} \) on the left-hand side of (10.19) is not maximally expanded. Then \( d(A_r^{\sigma_r}) = p + 1 \), so that (10.8) yields \( A_r^{\sigma_r} \prec \Psi_o^{p+1} \). Therefore the observation that \( A_r^{\sigma_r} \prec \Psi_o \) for all \( s \neq r \), together with (10.7) implies that the left-hand side of (10.19) is \( O_{\prec}(\Psi_o^{2p}) \). Since \( |L| \leq p \), (10.19) follows.

Consider now the case where \( A_r^{\sigma_r} \) on the left-hand side of (10.19) is maximally expanded for all \( r = 1, \ldots, p \). The key observation is the following claim about the left-hand side of (10.19) with a nonzero lower index \( k \).

For each \( s \in L \) there exists \( r := \tau(s) \in \{1, \ldots, p\} \setminus \{s\} \) such that the monomial \( A_r^{\sigma_r} \) contains a resolvent entry with lower index \( k_s \).

To prove (\ast), suppose by contradiction that there exists an \( s \in L \) such that for all \( r \in \{1, \ldots, p\} \setminus \{s\} \) the lower index \( k_s \) does not appear in the monomial \( A_r^{\sigma_r} \). To simplify notation, we assume that \( s = 1 \). Then, for all \( r = 2, \ldots, p \), since \( A_r^{\sigma_r} \) is maximally expanded, we find that \( A_r^{\sigma_r} \) is independent of \( k_1 \). Therefore, we have
\[
\mathbb{E}(Q_1, A_1^{\sigma_1}) \cdots (Q_p, A_p^{\sigma_p}) = \mathbb{E}Q_1 \cdots (A_2^{\sigma_2}) \cdots (Q_p, A_p^{\sigma_p}) = 0,
\]
where in the last step we used that \( \mathbb{E}Q_i(X)Y = \mathbb{E}Q_i(XY) = 0 \) if \( Y \) is independent of \( i \). This concludes the proof of (\ast).

The statement (\ast) can be reformulated as asserting that, after expansion, every label \( s \) has a “partner” label \( r = \tau(s) \), such that the index \( k_s \) appears also as a lower index in the expansion of \( A_r^{\tau(r)} \) (note that there may be several such partner labels \( r \), we can choose \( \tau(s) \) to be any one of them).

For \( r \in \{1, \ldots, p\} \) we define \( \ell(r) := \sum_{s \in L} 1(\tau(s) = r) \), the number of times that the label \( r \) was chosen as a partner to some lone label \( s \). We now claim that
\[
A_r^{\sigma_r} = O_{\prec}(\Psi_o^{1+\ell(r)}).
\]

To prove (10.21), fix \( r \in \{1, \ldots, p\} \). By definition, for each \( s \in \tau^{-1(\{r\})} \) the index \( k_s \) appears as a lower index in the monomial \( A_r^{\sigma_r} \). Since \( s \in L \) is by definition a lone label and \( s \neq r \), we know that \( k_s \) does not appear as an index in \( A_r^{\tau(r)} = 1/G_{k_1, k_2}, \) i.e., \( k_r \neq k_s \). By definition of the monomials associated with the tree vertex \( \sigma_r \), it follows that \( b(\sigma_r) \), the number of ones in \( \sigma_r \), is at least \( |\tau^{-1(\{r\})}| = \ell(r) \). Therefore, we conclude that that \( \tau(s) \) is a lone label. Recalling (10.20), we therefore get (10.21).

Using (10.21) and Lemma 10.1 we find
\[
\left| (Q_1^1, A_1^{\sigma_1}) \cdots (Q_p^p, A_p^{\sigma_p}) \right| \prec \prod_{r=1}^p \Psi_o^{1+\ell(r)} = \Psi_o^{p+|L|}.
\]

This concludes the proof of (10.19).

Summing over the binary trees in (10.18) and using Lemma 10.1, we get from (10.19)
\[
V(k) = O_{\prec}(\Psi_o^{p+|L|}).
\]

**Step 4. Summing over the expansion.** We now return to the sum (10.11). We perform the summation by first fixing \( P \in \Psi_o \), with associated lone labels \( L = L(P) \). Using \( |t_{ij}| \leq M^{-1} \), (8.45), we have
\[
\sum_{k} 1(P(k) = P) t_{ik_1} \cdots t_{ik_p/2} t_{ik_p/2+1} \cdots t_{ik_p} \leq M^{-p} \sum_{k} 1(P(k) = P).
\]
From (10.11) and (10.22) we get
\[ C \text{ constant} \]
where in the last step we used the lower bound from (8.20) and estimated the summation over \( P \) that is not contained in \( L \).

Now the number of \( k \) with \(|L|\) lone labels can be bounded easily by \( M^{|L|+(p-|L|)/2} \) since each block of \( P \) that is not contained in \( L \) consists of at least two labels. Thus we can bound the last displayed equation by
\[
\left| \sum_k 1(P(k) = P) t_{ik_1} \cdots t_{ik_p/2} \bar{t}_{ik_p/2+1} \cdots \bar{t}_{ik_p} \right| \leq M^{-p} M^{p+|L|} = (M^{-1/2})^{p-|L|}.
\]

From (10.11) and (10.22) we get
\[
E \left| \sum_k t_{ik} X_k \right|^p \leq \sum_{p \in \Psi_p} (M^{-1/2})^{p-|L|} \Psi_o^{p+|L|} \leq C_p \Psi_o^{2p},
\]
where in the last step we used the lower bound from (8.20) and estimated the summation over \( \Psi_p \) with a constant \( C_p \) (which is bounded by \((Cp^2)^p\)). Summarizing, we have proved that
\[
E \left| \sum_k t_{ik} X_k \right|^p \leq \Psi_o^{2p} \quad \text{(10.23)}
\]
for any \( p \in 2\mathbb{N} \).

We conclude the proof of (8.47) with a simple application of Chebyshev’s inequality. Fix \( \varepsilon > 0 \) and \( D > 0 \). Using (10.23) and Chebyshev’s inequality we find
\[
\mathbb{P} \left( \left| \sum_k t_{ik} X_k \right| > N^\varepsilon \Psi_o^2 \right) \leq N N^{-\varepsilon p}
\]
for large enough \( N \geq N_0(\varepsilon, p) \). Choosing \( p \geq \varepsilon^{-1}(1 + D) \), we conclude the proof of (8.47).

**Remark 10.4.** The first decoupling formula (8.1) is the only identity about the entries of \( G \) that is needed in the proof of Lemma 8.9. In particular, the second decoupling formula (8.2) is never used, and the actual entries of \( H \) never appear in the argument.

**Part II. Proof of (8.46).** The bound (8.46) may be proved by following the proof of (8.47) verbatim; the only modification is the bound
\[
|Q_k G^{(T)}_{kk}| = |Q_k (G^{(T)}_{kk} - m)| \prec \Psi,
\]
which replaces (10.10). Here we again use the same upper bound \( \Psi_o = \Psi \) for \( \Lambda \) and \( \Lambda_o \).

**Part III. Proof of (8.48) and (8.50).** In order to prove (8.48), we note that the last term in (8.17) is of order \( O_o(\Psi^2) \). Hence we have
\[
w_a := \sum_i t_{ai} v_i = m_{sc}^2 \sum_{i,k} t_{ai} s_{ik} v_k - m_{sc}^2 \sum_i t_{ai} Y_i + O_o(\Psi^2) = m_{sc}^2 \sum_{i,k} s_{ai} t_{ik} v_k + O_o(\Psi^2),
\]
where in the last step we used Corollary 8.10 (recall that the proof of this corollary used only (8.47) which was already proved in Part I) to bound \( \sum_i t_{ai} Y_i \) by \( O_o(\Psi^2) \) and that the matrices \( T \) and \( S \) commute by assumption. Introducing the vector \( w = \{w_a\}_{a=1}^N \) we therefore have the equation
\[
w = m_{sc}^2 \bar{S} w + O_o(\Psi^2),
\]
where the error term is in the sense of the \( \ell^\infty \)-norm (uniform in the components of the vector \( w \)). Inverting the matrix \( 1 - m_{sc}^2 \bar{S} \) and recalling the definition (6.17) yields (8.48).

The proof of (8.50) is similar, except that we have to treat the subspace \( e_1^\perp \) separately. Using (8.49) we write
\[
\sum_i t_{ai} (v_i - [v]) = \sum_i t_{ai} v_i - \sum_i \frac{1}{N} v_i,
\]

72
and apply the above argument to each term separately. This yields
\[ \sum_i t_{ai}(v_i - [v]) = m_{sc}^2 \sum_i t_{ai} \sum_k s_{ik} v_k - m_{sc}^2 \sum_i \frac{1}{N} \sum_k s_{ik} v_k + O_(\Psi^2) \]
\[ = m_{sc}^2 \sum_{i,k} s_{ai} t_{ik}(v_k - [v]) + O_(\Psi^2), \]
where we used (6.1) in the second step. Note that the error term on the right-hand side is perpendicular to \( \mathbf{e} \) when regarded as a vector indexed by \( a \), since all other terms in the equation are. Hence we may invert the matrix \((1 - m^2 S)\) on the subspace \( \mathbf{e}^\perp \), as above, to get (8.50). This completes the proof of Lemma 8.9. □

### 10.3 Alternative proof of (8.47) of Lemma 8.9

We conclude this section with an alternative proof of the key statement (8.47) of Lemma 8.9. While the underlying argument remains similar, the following proof makes use of an additional decomposition of the space of random variables, which avoids the use of the stopping rule from Step (1) in the above proof of Lemma 8.9. This decomposition may be regarded as an abstract reformulation of the stopping rule.

**Alternative proof of (8.47) of Lemma 8.9.** As before, we set \( X_k := Q_k(G_{kk})^{-1} \). For simplicity of presentation, we set \( t_{ik} = N^{-1} \). The decomposition is defined using the operations \( P_i \) and \( Q_i \), introduced in Definition 7.3. It is immediate that \( P_i \) and \( Q_i \) are projections, that \( P_i + Q_i = 1 \), and that all of these projections commute with each other. For a set \( A \subset \{1, \ldots, N\} \) we use the notations \( P_A := \prod_{i \in A} P_i \) and \( Q_A := \prod_{i \in A} Q_i \).

Let \( p \) be even and introduce the shorthand \( \tilde{X}_{ks} := X_{ks} \) for \( s \leq p/2 \) and \( \tilde{X}_{ks} := X_{ks} \) for \( s > p/2 \). Then we get
\[ \mathbb{E} \left[ \frac{1}{N} \sum_k X_k \right]^p = \frac{1}{N^p} \sum_{k_1, \ldots, k_p} \mathbb{E} \prod_{s=1}^p \tilde{X}_{k_s} = \frac{1}{N^p} \sum_{k_1, \ldots, k_p} \mathbb{E} \prod_{s=1}^p \left( \prod_{r=1}^p (P_{k_r} + Q_{k_r}) \tilde{X}_{k_s} \right). \]
Introducing the notations \( k = (k_1, \ldots, k_p) \) and \( |k| = \{k_1, \ldots, k_p\} \), we therefore get by multiplying out the parentheses
\[ \mathbb{E} \left[ \frac{1}{N} \sum_k X_k \right]^p = \frac{1}{N^p} \sum_k \sum_{A_1, \ldots, A_p \subset |k|} \mathbb{E} \prod_{s=1}^p (P_{A_s} Q_{A_s} \tilde{X}_{k_s}). \quad (10.25) \]

Next, by definition of \( \tilde{X}_{k_s} \), we have that \( \tilde{X}_{k_s} = Q_{k_s} \tilde{X}_{k_s} \), which implies that \( P_{A_s} \tilde{X}_{k_s} = 0 \) if \( k_s \notin A_s \). Hence may restrict the summation to \( A_s \) satisfying
\[ k_s \in A_s \quad (10.26) \]
for all \( s \). Moreover, we claim that the right-hand side of (10.25) vanishes unless
\[ k_s \in \bigcup_{q \neq s} A_q \quad (10.27) \]
for all \( s \). Indeed, suppose that \( k_s \in \bigcap_{q \neq s} A_q^c \) for some \( s \), say \( s = 1 \). In this case, for each \( s = 2, \ldots, p \), the factor \( P_{A_s} Q_{A_s} \tilde{X}_{k_s} \) is independent of \( k_1 \) (see Definition 7.3). Thus we get
\[ \mathbb{E} \prod_{s=1}^p (P_{A_s} Q_{A_s} \tilde{X}_{k_s}) = \mathbb{E} (P_{A_1} Q_{A_1} Q_{k_1} \tilde{X}_{k_1}) \prod_{s=2}^p (P_{A_s} Q_{A_s} \tilde{X}_{k_s}) \]
\[ = \mathbb{E} Q_{k_1} \left( (P_{A_1} Q_{A_1} \tilde{X}_{k_1}) \prod_{s=2}^p (P_{A_s} Q_{A_s} \tilde{X}_{k_s}) \right) = 0, \]
73
where in the last step we used that $\mathbb{E}Q_i(X) = 0$ for any $i$ and random variable $X$.

We conclude that the summation on the right-hand side of (10.25) is restricted to indices satisfying (10.26) and (10.27). Under these two conditions we have

$$\sum_{s=1}^{p}|A_s| \geq 2||k||,$$

(10.28) since each index $k_s$ must belong to at least two different sets $A_q$: to $A_s$ (by (10.26)) as well as to some $A_q$ with $q \neq s$ (by (10.27)).

Next, we claim that for $k \in A$ we have

$$|Q_AX_k| < \Psi_o^{[A]}.$$

(Note that if we were doing the case $X_k = Q_kG_{kk}$ instead of $X_k = Q_k(G_{kk})^{-1}$, then (10.29) would have to be weakened to $|Q_AX_k| < \Psi^{[A]}$, in accordance with (8.46). Indeed, in that case and for $A = \{k\}$, we only have the bound $|Q_kG_{kk}| < \Psi$ and not $|Q_kG_{kk}| < \Psi_o$.)

Before proving (10.29), we show how it may be used to complete the proof. Using (10.25), (10.29), and Lemma 10.1, we find

$$\mathbb{E}\left[\frac{1}{N}\sum_k X_k\right]^p < C_p \frac{1}{N^p} \sum_k \Psi_o^{2[p]}\sum_{u=1}^{p} \Psi_o^{2u} \frac{1}{N^p} \sum_k 1(||k|| = u)$$

$$\leq C_p \sum_{u=1}^{p} \Psi_o^{2u} N^{u-p} \leq C_p (\Psi_o + N^{-1/2})^{2p} \leq C_p \Psi_o^{2p},$$

where in the first step we estimated the summation over the sets $A_1, \ldots, A_p$ by a combinatorial factor $C_p$ depending on $p$, in the forth step we used the elementary inequality $a^n b^m \leq (a+b)^{n+m}$ for positive $a, b$, and in the last step we used (8.20) and the bound $M \leq N$. Thus we have proved (10.23), from which the claim follows exactly as in the first proof of (8.47).

What remains is the proof of (10.29). The case $|A| = 1$ (corresponding to $A = \{k\}$) follows from (10.10), exactly as in the first proof of (8.47). To simplify notation, for the case $|A| \geq 2$ we assume that $k = 1$ and $A = \{1, 2, \ldots, t\}$ with $t \geq 2$. It suffices to prove that

$$\left|Q_1 \cdots Q_2 \frac{1}{G_{11}}\right| < \Psi_o^2.$$

(10.30)

We start by writing, using the first decoupling formula (8.1),

$$Q_2 \frac{1}{G_{11}} = Q_2 \frac{1}{G_{11}^{(2)}} + Q_2 \frac{G_{12}G_{21}}{G_{11}^{(2)}G_{22}} = Q_2 \frac{G_{12}G_{21}}{G_{11}^{(2)}G_{22}},$$

where the first term vanishes since $G_{11}^{(2)}$ is independent of 2 (see Definition 7.3). We now consider

$$Q_3Q_2 \frac{1}{G_{11}} = Q_3Q_2 \frac{G_{12}G_{21}}{G_{11}^{(2)}G_{22}},$$

and apply again the first decoupling formula (8.1) with $k = 3$ to each resolvent entry on the right-hand side, and multiply everything out. The result is a sum of fractions of entries of $G$, whereby all entries in the numerator are diagonal and all entries in the denominator are diagonal. The leading order term vanishes,

$$Q_2Q_3 \frac{G_{12}^{(3)}G_{21}^{(3)}}{G_{11}^{(3)}G_{11}^{(2)}G_{22}^{(3)}} = 0,$$
so that the surviving terms have at least three (off-diagonal) resolvent entries in the numerator. We may now continue in this manner; at each step the number of (off-diagonal) resolvent entries in the numerator increases by at least one.

More formally, we obtain a sequence $A_2, A_3, \ldots, A_t$, where $A_2 := Q_2 G_{12} G_{21} G_{11} G_{22}^{(2)}$ and $A_i$ is obtained by applying (8.1) with $k = i$ to each entry of $Q_i A_{i-1}$, and keeping only the nonvanishing terms. The following properties are easy to check by induction.

(i) $A_t = Q_t A_{t-1}$.

(ii) $A_t$ consists of the projection $Q_2 \cdots Q_t$ applied to a sum of fractions such that all entries in the numerator are off-diagonal and all entries in the denominator are diagonal.

(iii) The number of (off-diagonal) entries in the numerator of each term of $A_t$ is at least $i$.

By Lemma 10.1 combined with (ii) and (iii) we conclude that $|A_t| \prec \Psi^t$. From (i) we therefore get

$$Q_t \cdots Q_2 \frac{1}{G_{11}} = A_t = O_\prec (\Psi^t).$$

This is (10.30). Hence the proof is complete.

10.3.1 History of the fluctuation averaging

Lemma 8.9 is a version of the fluctuation averaging mechanism, taken from [55], that is the most useful for our purpose in proving Theorem 6.7. We now briefly comment on its history.

The first version of the fluctuation averaging mechanism appeared in [68] for the Wigner case, where $[Z] = N^{-1} \sum_k Z_k$ was bounded by $A_2^2$ (recall $Z_i$ from (8.11)). Since $Q_k [G_{kk}]^{-1}$ is essentially $Z_k$, see (8.7), this corresponds to the first bound in (8.46). A different proof (with a better bound on the constants) was given in [70]. A conceptually streamlined version of the original proof was extended to sparse matrices [56] and to sample covariance matrices [113]. Finally, an extensive analysis in [52] treated the fluctuation averaging of general polynomials of resolvent entries and identified the order of cancellations depending on the algebraic structure of the polynomial. Moreover, in [52] an additional cancellation effect was found for the quantity $Q_i |G_{ij}|^2$. All proofs of the fluctuation averaging lemmas rely on computing expectations of high moments of the averages and carefully estimating various terms of different combinatorial structure. In [52] we have developed a Feynman diagrammatic representation for bookkeeping the terms, but this is necessary only for the case of general polynomials. For the special cases stated in Lemma 8.9, the proof presented here is relatively simple.
11 Eigenvalue location: the rigidity phenomenon

The local semicircle law in Theorem 6.7 was proven for universal Wigner matrices, characterized by the upper bound $s_{ij} \leq 1/M$ on the variances of $h_{ij}$. This result implies rigidity estimates on the location of the eigenvalues with a precision depending on $M$; e.g. in the bulk spectrum the eigenvalues can typically be located with a precision slightly above $1/M$. For the sake of a simplicity, from now on we restrict our presentation to the special case of generalized Wigner matrices, characterized by $C_{inf}/N \leq s_{ij} \leq C_{sup}/N$ with two positive constants $C_{inf}$ and $C_{sup}$, see Definition 2.1. So from now on the parameter $M$ will be replaced with $N$. For rigidity results in the general case, see Theorem 7.6 in [55].

In this section we will show that eigenvalues for generalized Wigner matrices are quite rigid; they may fluctuate only on a scale slightly above $1/N$. This is a manifestation that the eigenvalues are strongly correlated; the typical fluctuation scale of $N$ independent, say Poisson, points in a finite interval would be $N^{-1/2}$.

11.1 Extreme eigenvalues

We first state a corollary to Theorem 6.7 on the extreme eigenvalues.

**Corollary 11.1.** For any generalized Wigner matrix, the largest eigenvalue of $H$ is bounded by $2 + N^{-2/3+\varepsilon}$ for any $\varepsilon > 0$ in the sense that for any $\varepsilon > 0$ and any $D \geq 1$ we have

$$\mathbb{P}\left( \max_{\alpha=1,...,N} |\lambda_\alpha| \geq 2 + N^{-2/3+\varepsilon} \right) \leq N^{-D}$$

(11.1)

for any $N \geq N_0(\varepsilon, D)$.

**Proof.** Set $\eta = N^{-2/3}$ and choose an energy $E = 2 + \kappa$ outside of the spectrum with some $\kappa \geq N^{-2/3+\varepsilon} \gg N^{\varepsilon/2} \eta$. From (6.33) with $z = E + i\eta$ we have

$$|\text{Im} m_N(z) - \text{Im} m_{sc}(z)| \leq \frac{N^{\varepsilon/2}}{N\kappa} \ll \frac{1}{N\eta}$$

(11.2)

with very high probability. On the other hand, if there is an eigenvalue $\lambda$ with $|\lambda - E| \leq \eta$ then we have

$$\text{Im} m_N(z) \geq \frac{1}{2N\eta} \gg \text{Im} m_{sc}(z) + \frac{N^{\varepsilon/2}}{N\kappa}.\) (11.3)

Here the first inequality comes from

$$\text{Im} m_N(z) = \frac{1}{N} \sum_{\alpha} \frac{\eta}{|\lambda_\alpha - E|^2 + \eta^2} \geq \frac{1}{N} \frac{\eta}{|\lambda - E|^2 + \eta^2},$$

while the second follows from $\text{Im} m_{sc}(z) \approx \eta/\sqrt{\kappa}$ from (6.13).

The equations (11.2) and (11.3) however contradict each other, showing that with very high probability there is no eigenvalue with $|\lambda - E| \leq N^{-2/3}$. We can repeat this argument for a grid of energies $E = E_k = 2 + N^{-2/3+\varepsilon} + kN^{-2/3}$ for $k = 0, 1, \ldots, CN^{2/3+D}$ for any $D$ finite. We can then use the union bound for the exceptional sets of small probabilities to exclude the existence of eigenvalues between $N^{-2/3+\varepsilon}$ and $N^D$. Finally, for a large enough $D$, it is trivial to prove that the bound $\|H\| \leq \sum_{ij} |h_{ij}| \leq N^D$ holds with very high probability. This excludes eigenvalues $|\lambda| \geq N^D$ and proves the corollary. 

11.2 Stieltjes transform and regularized counting function

For any signed measure $\tilde{\rho}$ on the real line, define its Stieltjes transform by

$$S(z) = \int_{\mathbb{R}} \frac{\tilde{\rho}(\lambda)}{\lambda - z} \, d\lambda, \quad z \in \mathbb{C} \setminus \mathbb{R}.$$  

(11.4)
The Helffer-Sjöstrand formula (that was originally used to develop an alternative functional calculus for self-adjoint operators, see e.g., [36]) yields an expression for \( \int \lambda f(\lambda) \tilde{\rho}(\lambda) d\lambda \) for a large class of function \( f(\lambda) \) on the real line in terms of the Stieltjes transform of \( \tilde{\rho} \). More precisely, let \( f \in C^1(\mathbb{R}) \) with compact support and let \( \chi(y) \) be a smooth cutoff function with support in \([-1,1] \), with \( \chi(y) = 1 \) for \( |y| \leq 1/2 \) and with bounded derivatives. Define

\[
\tilde{f}(x+iy) := (f(x) + iyf'(x))\chi(y)
\]

to be an almost-analytic extension of \( f \). With the standard convention \( z = x + iy \) and \( \partial_z = [\partial_x + i\partial_y]/2 \), we have

\[
f(\lambda) = \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{\partial_z \tilde{f}(x+iy)}{\lambda - x - iy} dxdy.
\]

(11.5)

To see this identity, we use \( \partial_z(\lambda - z)^{-1} = 0 \) to write

\[
\frac{1}{\pi} \int_{\mathbb{R}^2} \frac{\partial_z \tilde{f}(x+iy)}{\lambda - x - iy} dxdy = \frac{1}{2\pi i} \int_{\mathbb{R}^2} \partial_z \left[ \frac{\tilde{f}(x+iy)}{\lambda - x - iy} \right] dzd\bar{z}.
\]

(11.6)

We can rewrite the last term as

\[
\frac{1}{2\pi i} \int_{\mathbb{R}^2} d\left[ \frac{\tilde{f}(x+iy)}{\lambda - x - iy} \right],
\]

(11.7)

where the operator \( d \) is the differential in the sense of one form. From the Green theorem and the compact support of \( \tilde{f} \), we can integrate by parts to a small circle \( C_\varepsilon \) of radius \( \varepsilon \) around \( \lambda \). The contribution of the two dimensional integration inside the circle will vanish in the limit \( \varepsilon \to 0 \). Hence the last term is equal to

\[
\lim_{\varepsilon \to 0} \frac{1}{2\pi i} \int_{C_\varepsilon} \tilde{f}(x+iy) \frac{d}{\lambda - x - iy} dz = f(\lambda)
\]

(11.8)

and this proves the (11.5). Computing the \( \partial_z \) in (11.5) explicitly, we have

\[
f(\lambda) = \frac{1}{\pi} \int_{\mathbb{R}^2} iyg''(x)\chi(y) + i(f(x) + iyf'(x))\chi'(y) dxdy.
\]

(11.9)

We remark that the same formula holds if we simply defined \( \tilde{f}(x+iy) := f(x)\chi(y) \). The addition of the term \( iyf'(x)\chi(y) \) is to make \( \partial_z \tilde{f}(x+iy) = O(y) \) for \( |y| \) small which will be needed in the following estimates. In fact, the concept of almost-analytic function can be extended to arbitrary order \( n \). We could have defined \( \tilde{f} \) such that \( \partial_z \tilde{f}(x+iy) = O(|y|^n) \), but we will not need this result as it would not improve our estimate.

The following lemma in a slightly less general forms appeared in [59,68,69]. It shows how to translate estimates on the Stieltjes transform \( S(z) \) in the regime \( \text{Im } z \geq \tilde{\eta} \) to the regularized distribution function of \( \tilde{\rho} \). Given two real numbers, \( E_1 < E_2 \), we wish to estimate \( \tilde{\rho}([E_1, E_2]) \), the \( \tilde{\rho} \)-measure of the interval \([E_1, E_2]\). Since the integral of the sharp indicator function cannot be directly controlled, we need to smooth it out; the function \( f_2 - f_1 \) in the lemma below plays this role. The scale of the smoothing, denoted by \( \tilde{\eta} \) below, must be larger than \( \tilde{\eta} \). Another feature of this lemma is that the condition on \( S(z) = S(E+i\eta) \) is local; only \( E \)-values in a \( \tilde{\eta} \)-neighborhood of the endpoints \( E_1, E_2 \) are needed.

**Lemma 11.2.** Let \( \tilde{\rho} \) be a signed measure on \( \mathbb{R} \) and let \( S \) be its Stieltjes transform. Fix two energies \( E_1 < E_2 \). Suppose that for some \( 0 < \tilde{\eta} \leq 1/2 \) and \( 0 < U_1 \leq U_2 \) we have

\[
|S(x+iy)| \leq \frac{U_1}{y} \text{ for any } x \in [E_1, E_1 + \tilde{\eta}] \cup [E_2, E_2 + \tilde{\eta}] \text{ and } \tilde{\eta} \leq y \leq 1,
\]

(11.10)

\[
|\text{Im } S(x+iy)| \leq \frac{U_1}{y} \text{ for any } x \in [E_1, E_2 + \tilde{\eta}] \text{ and } 1/2 \leq y \leq 1,
\]

(11.11)

\[
|\text{Im } S(x+iy)| \leq \frac{U_2}{y} \text{ for any } x \in [E_1, E_1 + \tilde{\eta}] \cup [E_2, E_2 + \tilde{\eta}] \text{ and } 0 < y < \tilde{\eta}.
\]

(11.12)
For \( \nu > \eta \), define two functions \( f_j = f_{E_j, \nu} : \mathbb{R} \to \mathbb{R} \) such that \( f_j(x) = 1 \) for \( x \in (-\infty, E_j] \), \( f_j(x) \) vanishes for \( x \in [E_j + \nu, \infty) \); moreover \( |f_j'(x)| \leq C \eta^{-1} \) and \( |f_j''(x)| \leq C \eta^{-2} \) for some constant \( C \). Then for some other constant \( C > 0 \) independent of \( U_1, U_2 \) and \( \nu \), we have

\[
\left| \int (f_2 - f_1)(\nu) \varphi(\lambda) d\lambda \right| \leq C \| \varphi \|_{TV} \left[ |U_1| \log \eta + U_2 \eta^2 \right],
\]

where \( \| \cdot \|_{TV} \) denotes the total variation norm of signed measures. In addition, if we assume that \( \varphi \) has compact support, then we also have

\[
\left| \int f_2(\nu) \varphi(\lambda) d\lambda \right| \leq C \| \varphi \|_{TV} \left[ |U_1| \log \eta + U_2 \eta^2 \right],
\]

where \( C \) depends on the size of the support of \( \varphi \).

**Proof.** In this proof, we will consider only the case \( \nu = \eta \) and \( U_1 = U_2 \). We will denote these common values by \( \eta \) and \( U \). We may also assume \( \| \varphi \|_{TV} = 1 \). These simplifications only streamline some notation, the general case is proven exactly in the same way.

Let \( f = f_2 - f_1 \) that is smooth and compactly supported. From the representation (11.9) and since \( f \) is real, we have

\[
\int_{-\infty}^{\infty} f(\nu) \varphi(\lambda) d\lambda = \text{Re} \int_{-\infty}^{\infty} f(\nu) \varphi(\lambda) d\lambda \leq C \int \int y f''(x) \chi(y) \text{Im} S(x + iy) dxdy + C \int \int \left( |f(x)| |\chi'(y)| |\text{Im} S(x + iy)| + |y||f'(x)||\chi'(y)||\text{Re} S(x + iy)| \right) dxdy,
\]

for some universal \( C > 0 \), and where \( \chi \) is a smooth cutoff function with support in \([-1, 1]\), with \( \chi(y) = 1 \) for \(|y| \leq 1/2\) and with bounded derivatives. Recall that \( f' \) is \( O(\eta^{-1}) \) on two intervals of size \( O(\eta) \) each and \( \chi' \) is supported in \( 1/2 \leq |y| \leq 1 \). By (11.10), we can bound

\[
\int \int |y||f'(x)||\chi'(y)||\text{Re} S(x + iy)| dxdy \leq CU
\]

(note that due to \( \tilde{S}(x + iy) = S(x - iy) \), the bounds analogous to (11.10)–(11.11) also hold for negative \( y \)'s). Using that \( f \) is bounded with compact support, we have, by (11.11),

\[
\int \int |f(x)||\chi'(y)||\text{Im} S(x + iy)| dxdy \leq CU.
\]

For the first term on the right hand side of (11.15), we split the integral into two regimes depending on \( 0 < y < \eta \) or \( \eta < y < 1 \). Note that by symmetry we only need to consider positive \( y \). From (11.12), the integral on the first integration regime is easily bounded by

\[
\left| \int \int_{0 < y < \eta} y f''(x) \chi(y) \text{Im} S(x + iy) dxdy \right| = O \left( \int \int_{0 < y < \eta} \left[ 1(E_1 \leq x \leq E_1 + \eta) + 1(E_2 \leq x \leq E_2 + \eta) \right] y \eta^{-2} \frac{U}{y} dxdy \right) = O(U).
\]

For the second integration regime, \( y \in [\eta, 1] \), we can integrate by parts in \( x \), then use \( \partial_x \text{Im} S(x + iy) = -\partial_y \text{Re} S(x + iy) \) and then integrate by parts in \( y \) to get:

\[
\int \int_{y > \eta} y f''(x) \chi(y) \text{Im} S(x + iy) dxdy = - \int \int_{y > \eta} f'(x) \partial_y(y \chi(y)) \text{Re} S(x + iy) dxdy
\]

\[
- \int f'(x) \eta \chi(\eta) \text{Re} S(x + i\eta) dx.
\]

78
Recall that \( f'(x) = 0 \) unless \( |x - E_j| < \eta \) for some \( j = 1, 2 \) and in this regime we have \( f' = O(\eta^{-1}) \). By (11.10), the last integral is easily bounded by \( O(U) \). For the first integral, we use \( \partial_y y \chi(y) = O(1) \) and \( S(x + iy) = O(U/y) \) to have

\[
\left| \int_{y > \eta} \partial_y (y \chi(y)) f'(x) S(x + iy) \, dx \, dy \right| \leq O \left( U \int_{\eta}^{1} \frac{dy}{y} \right) = O \left( U |\log \eta| \right),
\]

which is (11.13). To prove (11.14), we simply choose \( E_1 \) to be any energy on the left side of the support of \( \tilde{\rho} \) and notice that

\[
\left| \int f_2(\lambda) \tilde{\rho}(\lambda) \, d\lambda \right| = \left| \int (f_2 - f_1)(\lambda) \tilde{\rho}(\lambda) \, d\lambda \right| \leq C U |\log \tilde{\eta}|.
\]

This completes the proof of Lemma 11.2.

11.3 Convergence speed of the empirical distribution function

Define the empirical distribution function (or, in physics terminology, integrated density of states)

\[
n_N(E) := \frac{1}{N} |\{\alpha : \lambda_\alpha \leq E\}| = \int_{-\infty}^{E} \varrho_N(x) \, dx,
\]

where \( \varrho_N \) is the empirical eigenvalue distribution

\[
\varrho_N(x) = \frac{1}{N} \sum_{\alpha=1}^{N} \delta(x - \lambda_\alpha).
\]

Similarly, we define the distribution function of the semicircle density

\[
n_{sc}(E) := \int_{-\infty}^{E} \varrho_{sc}(x) \, dx.
\]

We introduce the differences

\[
\varrho^\Delta := \varrho_N - \varrho_{sc}, \quad m^\Delta := m_N - m_{sc}.
\]

and recall \( m_N(z) = \frac{1}{N} \Tr G(z) = \int \varrho_N(x)(x - z)^{-1} \, dx \) is the Stieltjes transform of the empirical density. The following lemma, proven below, shows how local semicircle law implies a convergence rate for the empirical distribution function.

**Lemma 11.3.** Suppose that for some \( |E| \leq 10 \) and \( N^{-1+\varepsilon} \leq \tilde{\eta} \ll 1 \) with some \( \varepsilon > 0 \) we know that

\[
|m_N(x + i\eta) - m_{sc}(x + i\eta)| \prec \frac{1}{N\eta}
\]

holds for any \( |x - E| \leq \tilde{\eta} \) and any \( \eta \in [\tilde{\eta}, 10] \). Then we have

\[
\left| n_N(E) - n_{sc}(E) \right| \prec \tilde{\eta}.
\]

For generalized Wigner matrices, the condition (11.23) holds with the choice \( \tilde{\eta} = N^{-1+\varepsilon} \) uniformly in \( |E| \leq 10 \) (see (6.29) and (6.32)). Thus we have proved the following corollary:

**Corollary 11.4.** For generalized Wigner matrices, we have

\[
\sup_{|E| \leq 10} \left| n_N(E) - n_{sc}(E) \right| \leq N^{-1+\varepsilon}
\]

with probability at least \( 1 - N^{-D} \) for any \( D, \varepsilon > 0 \), if \( N \geq N_0(\varepsilon, D) \) is sufficiently large.
Proof of Lemma 11.3. Since both the semicircle measure and the empirical measure (by Corollary 11.1) have a compact support within \([-3, 3]\) with a very high probability, \((11.24)\) clearly holds for \(E \leq -3\) and for \(E \geq 3\). For an energy \(|E| \leq 3\), we will apply Lemma 11.2 with the choice \(\rho = \rho^\Delta\), \(U_1 = U_2 = N^\epsilon \eta\) and \(E_2 = E\). The estimate \((11.14)\) will immediately imply \((11.24)\), provided the other assumptions on Lemma 11.2 are satisfied.

For this purpose, we need the bounds \((11.10), (11.11)\) and \((11.12)\) on the Stieltjes transform with the choice \(U_1 = U_2 = N^\epsilon \eta\). We denote this common value by \(U := N^\epsilon \eta\). The assumption \((11.23)\) implies \((11.10)\) (with very high probability) if we choose \(U \geq N^{-1+\epsilon}\). From Lemma 6.2 we find

\[
|\text{Im} m_{sc}(x + iy)| \leq C \sqrt{\kappa_x + y},
\]

recalling the definition \(\kappa_x = \min\{|x - 2|, |x + 2|\}\) from \((6.23)\). By spectral decomposition of \(H\), it is easy to see that the function \(y \mapsto y \text{Im} m_N(x + iy)\) is monotonically increasing. Thus we get, using \((11.26)\) and \((11.23)\), that

\[
y \text{Im} m_N(x + iy) \leq \eta \text{Im} m_N(x + i\eta) < \eta \left(\sqrt{\kappa_x + \eta + \frac{1}{N\eta}}\right) < \eta \sqrt{\kappa_x + \eta + \frac{1}{N}},
\]

for \(y \leq \eta\) and \(|x| \leq 10\). Using the notation \(m^\Delta := m_N - m_{sc}\) and recalling \((11.26)\), we therefore get

\[
|y \text{Im} m^\Delta(x + iy)| < \eta \sqrt{\kappa_x + \eta + \frac{1}{N}} \leq C\eta
\]

for \(y \leq \eta\) and \(|x| \leq 10\); here we have used the assumption \(\eta \geq \frac{1}{N}\) in the last step. Therefore, with the choice \(U = \eta N^\epsilon\), the bounds \((11.10), (11.11)\) and \((11.12)\) indeed hold.

Recall from Lemma 11.2 that \(f_{E, \eta}\) denotes a smooth indicator function of \([-\infty, E]\) on scale \(\eta\), namely \(f_{E, \eta}(x) = 1\) for \(x \in [-\infty, E]\) and \(f_{E, \eta}(x) = 0\) for \(x \geq E + \eta\) and having derivative bounded by \(\eta^{-1}\). Using \((11.14)\) from Lemma 11.2, we conclude that

\[
\left| \int f_{E, \eta}(\lambda) \varrho^\Delta(\lambda) \, d\lambda \right| \leq 1 - \eta,
\]

Hence we have

\[
n_N(E) - n_{sc}(E) - \int f_{E, \eta}(\lambda) \varrho^\Delta(\lambda) \, d\lambda = - \int_{E}^{E + \eta} f_{E, \eta}(\lambda) \varrho_N(\lambda) \, d\lambda + \int_{E - \eta}^{E + \eta} f_{E, \eta}(\lambda) \varrho_{sc}(\lambda) \, d\lambda \leq C\eta.
\]

Notice that we only used the positivity of \(\varrho_N\) in the first term and the boundedness of \(f\) and \(\varrho_{sc}\) in the second. Conversely, we have

\[
n_N(E) - n_{sc}(E) - \int f_{E - \eta, \eta}(\lambda) \varrho^\Delta(\lambda) \, d\lambda = \int_{E - \eta}^{E} (1 - f_{E - \eta, \eta})(\lambda) \varrho_N(\lambda) \, d\lambda - \int_{E - \eta}^{E} (1 - f_{E - \eta, \eta})(\lambda) \varrho_{sc}(\lambda) \, d\lambda \geq -C\eta.
\]

Hence we have proved that

\[
-C\eta + \int f_{E - \eta, \eta}(\lambda) \varrho^\Delta(\lambda) \, d\lambda \leq n_N(E) - n_{sc}(E) \leq \int f_{E, \eta}(\lambda) \varrho^\Delta(\lambda) \, d\lambda + C\eta.
\]

The right hand side is directly bounded by \((11.29)\), the left hand side is estimated in the same way, using \((11.29)\) for \(f_{E - \eta, \eta}\) instead of \(f_{E, \eta}\). This proves Lemma 11.3.

\[
\square
\]

11.4 Rigidity of eigenvalues

We now state and prove the rigidity theorem concerning the eigenvalue locations for generalized Wigner matrices. Recall the definition of \(\varrho_N\) from \((11.22)\). Hence we have

\[
\frac{j}{N} = \int_{-\infty}^{\lambda_j} \varrho_N(x) \, dx.
\]
We define the classical location of the \( j \)-th eigenvalue by the equation

\[
\frac{j}{N} = \int_{-\infty}^{\gamma_j} \rho_{sc}(x)dx,
\]

in other words, \( \gamma_j \) is the \( j \)-th \( N \)-quantile of the semicircle distribution.

**Theorem 11.5** (Rigidity of eigenvalues). For any generalized Wigner matrix ensemble, for any constant \( \varepsilon > 0 \) and any constant \( D \geq 1 \) we have

\[
P\left\{ \exists j : |\lambda_j - \gamma_j| \geq N^\varepsilon \left[ \min \left( j, N - j + 1 \right) \right]^{-1/3} N^{-2/3} \right\} \leq N^{-D}
\]

for any sufficiently large \( N \geq N_0(\varepsilon, D) \).

**Proof.** We consider only the case \( j \leq N/2 \); the other case is similar. By (11.30) and (11.31) we have

\[
0 = \int_{-\infty}^{\lambda_j} \rho_N(x)dx - \int_{-\infty}^{\gamma_j} \rho_{sc}(x)dx = \int_{-\infty}^{\lambda_j} (\rho_N(x) - \rho_{sc}(x))dx - \int_{\lambda_j}^{\gamma_j} \rho_{sc}(x)dx,
\]

i.e.,

\[
\left| \int_{\lambda_j}^{\gamma_j} \rho_{sc}(x)dx \right| \leq \left| \int_{-\infty}^{\lambda_j} (\rho_N(x) - \rho_{sc}(x))dx \right| = |n_N(\lambda_j) - n_{sc}(\lambda_j)|,
\]

and thus by the uniform bound (11.25)

\[
\left| \int_{\lambda_j}^{\gamma_j} \rho_{sc}(x)dx \right| \leq \frac{1}{N}.
\]

Since \( \rho_{sc}(x) = \sqrt{\kappa_x} \) for \( x \in [-2, 1] \), we have

\[
n_{sc}(x) \asymp \kappa_x^{3/2}, \quad \rho_{sc}(x) = n'_{sc}(x) \asymp n_{sc}^{1/3}(x), \quad x \in [-2, 1].
\]

This also implies for \( j \leq N/2 \) that

\[
\gamma_j + 2 \asymp \left( \frac{j}{N} \right)^{\varepsilon/3}, \quad \rho_{sc}(\gamma_j) \asymp \left( \frac{j}{N} \right)^{1/3}.
\]

If we knew that \( \rho_{sc}(\gamma_j) \) and \( \rho_{sc}(\lambda_j) \) were comparable, then the monotonicity of \( \rho_{sc} \) and the mean-value theorem (11.34) would immediately imply

\[
|\lambda_j - \gamma_j| \rho_{sc}(\gamma_j) \asymp N^{-1}.
\]

Combining this with the second asymptotics form (11.36), we would have

\[
|\lambda_j - \gamma_j| < N^{-1} \rho_{sc}(\gamma_j)^{-1} \leq CN^{-1}(j/N)^{\varepsilon/3}.
\]

For the complete proof, we first consider the indices \( j \geq j_0 := N^{\varepsilon/2} \). Since \( n_{sc}(\gamma_j) \geq N^{-1+\varepsilon/2} \) and \( n_{sc}(\gamma_j) = n_N(\lambda_j) \), we have

\[
|n_{sc}(\lambda_j) - n_{sc}(\gamma_j)| \leq N^{-\varepsilon/4} n_{sc}(\gamma_j)
\]

with very high probability by (11.25). This shows that \( n_{sc}(\lambda_j) \asymp n_{sc}(\gamma_j) \), but then \( \rho_{sc}(\lambda_j) \asymp \rho_{sc}(\gamma_j) \) by (11.35), as we presumed above.

Finally, for indices \( j \leq j_0 = N^{\varepsilon/2} \) we use monotonicity and rigidity (11.38) for the index \( j_0 \):

\[
\lambda_j \leq \lambda_{j_0} \leq \gamma_{j_0} + N^{-2/3+\varepsilon/2} \leq -2 + N^{-2/3+\varepsilon}
\]

with very high probability. In the last step we used (11.36). For the lower bound on \( \lambda_j \) we refer to (11.1). This shows that \( |\lambda_j + 2| \leq N^{-2/3+\varepsilon} \) with very high probability and we also have \( |\gamma_j + 2| \leq N^{-2/3+\varepsilon} \), so \( |\lambda_j - \gamma_j| \leq 2N^{-2/3+\varepsilon} \). This proves the rigidity estimate (11.32) for all indices \( j \).

□
12 Universality for matrices with Gaussian convolutions

12.1 Dyson Brownian motion

Consider the following matrix valued stochastic differential equation

\[ \frac{dH(t)}{dt} = \frac{1}{\sqrt{N}} dB(t) - \frac{1}{2} H(t) dt, \quad t \geq 0 \]  
\( (12.1) \)

with initial data \( H_0 \), where \( B(t) \) is defined somewhat differently in the two symmetry classes, namely:

(i) in the real symmetric case (indicated by superscript) \( B^{(s)} \) is an \( N \times N \) matrix such that \( b_{ij}^{(s)} \) for \( i < j \) and \( b_{ii}^{(s)}/\sqrt{2} \) are independent standard Brownian motions, and \( b_{ij}^{(s)} = b_{ji}^{(s)} \).

(ii) in the complex Hermitian case \( B^{(h)} \) is an \( N \times N \) matrix such that \( \sqrt{2} \text{Re}(b_{ij}^{(h)}) \), \( \sqrt{2} \text{Im}(b_{ij}^{(h)}) \) for \( i < j \) and \( b_{ii}^{(h)} \) are independent standard Brownian motions, and \( b_{ij}^{(h)} = (b_{ij}^{(h)})^* \).

We will drop the \( s \) or \( h \) superscript, the following formulas hold for both cases. In coordinates, we have

\[ dh_{ij}(t) = \frac{db_{ij}(t)}{\sqrt{N}} - \frac{1}{2} h_{ij}(t) dt, \quad (12.2) \]

where the entries \( h_{ij}(t) \) have variance \( t \) in the Hermitian case and in the symmetric case the off-diagonal entries \( (h_{ij}(t)) \) have variance \( t \), while the diagonal entries have variance \( 2t \).

The equation \((12.1)\) defines a **matrix valued Ornstein-Uhlenbeck (OU) process**. It plays a distinguished role in the theory of random matrices that was discovered by Dyson. Depending on the typographical convenience, the time dependence will sometimes be indicated as an argument, \( H(t) \), and sometimes as an index, \( H_t \), we keep both notations in parallel, i.e., \( H(t) = H_t \). In particular, unlike in some PDE literature, subscripts do not mean derivatives.

Next, we are concerned about the evolution of the eigenvalues \( \lambda(t) = (\lambda_1(t), \lambda_2(t), \ldots, \lambda_N(t)) \) of \( H_t \) along the OU flow. We label the eigenvalues in an increasing order, i.e., we assume that \( \lambda(t) \in \Sigma_N \), where we define the open simplex

\[ \Sigma_N := \{ \lambda \in \mathbb{R}^N : \lambda_1 < \lambda_2 < \ldots < \lambda_N \}. \]  
\( (12.3) \)

A theorem below will guarantee that the eigenvalues are simple and they are continuous functions of \( t \), hence the labelling is consistently preserved along the evolution.

In principle, the eigenvalues \( \{\lambda_i(t)\} \) and eigenvectors \( \{u_i(t)\} \) of \( H_t \) are strongly related and one would expect a coupled system of stochastic differential equations for them. It was Dyson’s fundamental observation [45] that the eigenvalues themselves satisfy an autonomous system of stochastic differential equations (SDE) that do not involve the eigenvectors. This SDE is given in the following definition.

**Definition 12.1.** Given a real parameter \( \beta \geq 1 \), consider the following system of SDE

\[ d\lambda_i = \frac{\sqrt{2}}{\sqrt{\beta N}} dB_i + \left( -\frac{\lambda_i}{2} + \frac{1}{N} \sum_{j}^{(i)} \frac{1}{\lambda_i - \lambda_j} \right) dt, \quad i \in [1, N], \]  
\( (12.4) \)

where \( (B_i) \) is a collection of real-valued, independent standard Brownian motions. The solution of this SDE is called the **Dyson Brownian Motion (DBM)** with parameter \( \beta \).

Notice that we defined the DBM for any \( \beta \geq 1 \) and not only for the classical values \( \beta = 1, 2 \) that will correspond to an OU matrix flow. The following theorem summarizes the main properties of the DBM. For the proof, see Lemma 4.3.3 and Proposition 4.3.5 of [8]. We remark that the authors in [8] considered \((12.4)\) without the drift term \(-\lambda_i/2\). However, any drift term of the form \(-V'(\lambda_i)\) with \( V \in C^2(\mathbb{R}) \) could be added without further complications.
Theorem 12.2. Let \( \beta \geq 1 \) and suppose that the initial data satisfy \( \lambda(0) \in \Sigma_N \). Then there exists a unique (strong) solution to (12.4) in the space of continuous functions \( (\lambda(t))_{t \geq 0} \in C(\mathbb{R}_+, \Sigma_N) \). Furthermore, for any \( t > 0 \) we have \( \lambda(t) \in \Sigma_N \) and \( \lambda(t) \) depends continuously on \( \lambda(0) \). In particular, if \( \lambda(0) \in \Sigma_N \), i.e., the multiplicity of the initial points is one, then \( (\lambda(t))_{t \geq 0} \in C(\mathbb{R}_+, \Sigma_N) \), i.e., this property is preserved for all times along the evolution.

We point out that the solution is considered in the strong sense. This means that despite the singularity in (12.4), the DBM admits a solution for almost all realization of the Brownian motions \( B_i \). For the precise definition we recall the concept of strong solution to a (scalar) SDE of the form

\[
dX_t = a(X_t)dB_t + b(X_t)dt, \quad X_0 = \xi, \quad t \geq 0,
\]

with respect to a fixed realization of a Brownian motion \( B_i \), where \( a \) and \( b \) are given coefficient functions. The strong solution is a process \( (X_t)_{t \geq 0} \) on a filtrated probability space \( (\Omega, F_t) \) with continuous sample paths that satisfies the following properties

- \( X_t \) is adapted to the filtration \( F_t = \sigma(G_t \cup N) \), where
  \[
  G_t = \sigma(B_s, s \leq t, X_0), \quad N := \{ N \subset \Omega, \exists G \subset G_\infty \text{ with } N \subset G, \ P(G) = 0 \},
  \]
  i.e., the filtration of \( X_t \) is the same as that of \( B_t \) after completing it with events of zero measure;
- \( X_0 = \xi \) almost surely;
- For any time \( t \), we have \( \int_0^t [b(X_s)]^2 + |a(X_s)|^2 \) \( ds < \infty \) almost surely;
- The integral version of (12.5), i.e.,
  \[
  X_t = X_0 + \int_0^t a(X_s)dB_s + \int_0^t b(X_s)ds
  \]
holds almost surely for any \( t \geq 0 \).

For simplicity we only presented the definition for the scalar case, \( X_t \in \mathbb{R} \); the extension to the vector valued case, \( X_t \in \mathbb{R}^N \), is straightforward.

The relation between the OU matrix flow (12.1) and the DBM (12.4) is given by the following theorem essentially due to Dyson.

Theorem 12.3. [45] Let \( H_t \) solve the matrix valued SDE (12.1) in a strong sense. Then its eigenvalue process satisfies (12.4), where the parameter \( \beta = 1 \) if the matrix \( B \) is real symmetric and \( \beta = 2 \) if \( B \) is complex Hermitian.

We remark that the Ornstein-Uhlenbeck drift term \(-\frac{1}{2}H_t dt\) in (12.1) is not essential. One may replace it with \(-V'(H)dt\) with any \( V \in C^2(\mathbb{R}) \) potential, then the same theorem holds but the \(-\lambda_i/2 \) term has to be replaced with \(-V'(\lambda_i) \) in (12.4). In particular, the drift term can be removed; for example the monograph [8] defines DBM without the drift term. While there is no fundamental difference between these formulations, (12.1) is technically slightly more convenient since it preserves not only the zero expectation value but also the \( 1/N \) variance of the matrix elements if the variance of the initial data \( H_0 \) is \( 1/N \).

If technical complications related to the singularities \((\lambda_i - \lambda_j)^{-1} \) in (12.4) could be neglected, the proof of Theorem 12.3 would be a relatively straightforward Itô calculus combined with standard perturbation formulas for eigenvalues and eigenvectors of self-adjoint matrices. We will present this calculation in the next section. A rigorous proof is slightly more cumbersome since a priori the \( L^2 \) integrability of the singularity is not guaranteed. The actual proof first regularizes the singular term on a very short scale. Then a stopping time argument shows that the regularization can be removed since the dynamics has an effective level repulsion mechanism that keeps neighboring points separated. We will not present this argument here since it is not particularly instructive for this book. The full details can be found in Section 4.3.1 of [8] for the case when the term \(-\frac{1}{2}H_t dt\) is not present in (12.1). The necessary modifications for (12.1) (or for other drift term \(-V'(\lambda_i) \)) are straightforward.
12.2 Derivation of Dyson Brownian motion and perturbation theory

Let $u_\alpha$, $\alpha = 1, 2, \ldots, N$, be the $\ell^2$-normalized eigenvectors of $H = (h_{ij})$ with (real) eigenvalues $\lambda_\alpha$. For simplicity we assume that all eigenvalues are distinct. Fix an index pair $(i, j)$ and abbreviate the partial derivative as

$$\dot{f} := \frac{\partial f}{\partial h_{ij}}.$$

Differentiating $H u_\alpha = \lambda_\alpha u_\alpha$ and $u^*_\beta u_\beta = \delta_{\alpha\beta}$ yields

$$\dot{H} u_\alpha + H \dot{u}_\alpha = \dot{\lambda}_\alpha u_\alpha + \lambda_\alpha \dot{u}_\alpha,$$  \hspace{1cm} (12.6)

as well as

$$\dot{u}_\alpha^* u_\beta + u_\alpha^* \dot{u}_\beta = 0, \quad \dot{u}_\alpha^* u_\alpha = 0.$$  \hspace{1cm} (12.7)

Taking the inner product with $u_\alpha$ on both side of (12.6), we get

$$\dot{\lambda}_\alpha = u_\alpha^* H u_\alpha.$$  \hspace{1cm} (12.8)

Moreover, multiplying (12.6) by $u_\beta^*$ yields, for $\beta \neq \alpha$,

$$u_\beta^* \dot{H} u_\alpha + u_\beta^* H \dot{u}_\alpha = \lambda_\alpha u_\beta^* \dot{u}_\alpha;$$

i.e.,

$$u_\beta^* \dot{H} u_\alpha + \lambda_\beta u_\beta^* \dot{u}_\alpha = \lambda_\alpha u_\beta^* \ddot{u}_\alpha,$$

where we have used that $H$ is self-adjoint and $\lambda_\beta$ is real in the last equation. Hence,

$$\dot{u}_\alpha = \sum_{\beta \neq \alpha} (u_\beta^* \ddot{u}_\alpha) u_\beta = \sum_{\beta \neq \alpha} \frac{u_\beta^* \dot{H} u_\alpha}{\lambda_\alpha - \lambda_\beta} u_\beta.$$  \hspace{1cm} (12.9)

For simplicity of notations, from now on, we consider only the real symmetric case. The complex Hermitian case can be treated similarly. In the real symmetric case, (12.7) reads as

$$\frac{\partial \lambda_\alpha}{\partial h_{ij}} = u_\alpha(i) u_\alpha(j) [2 - \delta_{ij}],$$  \hspace{1cm} (12.10)

where $u_\alpha(1), u_\alpha(2), \ldots, u_\alpha(N)$ denote the coordinates of the vector $u_\alpha$. We may assume that the eigenvectors are real. From (12.8) we get

$$\frac{\partial u_\alpha(k)}{\partial h_{ij}} = \sum_{\beta \neq \alpha} \frac{u_\beta(i) u_\alpha(j) + u_\beta(j) u_\alpha(i) [1 - \delta_{ij}]}{\lambda_\alpha - \lambda_\beta} u_\beta(k).$$  \hspace{1cm} (12.11)

Combining these last two formulas allows us to compute the second partial derivatives, i.e., for any fixed indices $i, j, k, \ell$ we have

$$\frac{\partial^2 \lambda_\alpha}{\partial h_{ij} \partial h_{k\ell}} = \begin{bmatrix} \frac{\partial u_\alpha(i)}{\partial h_{ij}} u_\alpha(k) + u_\alpha(i) \frac{\partial u_\alpha(k)}{\partial h_{ij}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\lambda_\alpha - \lambda_\beta} \left( (u_\beta(j) u_\alpha(\ell) + u_\beta(\ell) u_\alpha(j) [1 - \delta_{\ell\ell}]) u_\beta(i) u_\alpha(k) \
+ (u_\beta(\ell) u_\alpha(i) + u_\beta(i) u_\alpha(j) [1 - \delta_{i\ell}]) u_\beta(k) u_\alpha(i) \right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\lambda_\alpha - \lambda_\beta} \left( (u_\beta(j) u_\alpha(\ell) + u_\beta(\ell) u_\alpha(j) [1 - \delta_{\ell\ell}]) (u_\beta(i) u_\alpha(k) + u_\beta(k) u_\alpha(i)) \right) \end{bmatrix}.$$
By (12.2), (12.9), (12.11) and using Ito’s formula (neglecting the issue of singularity), we have

\[
d\lambda_\alpha = \sum_{i \leq k} \frac{\partial \lambda_\alpha}{\partial h_{ik}} d h_{ik} + \frac{1}{2} \sum_{i \leq k, j \leq k} \frac{\partial^2 \lambda_\alpha}{\partial h_{ik} \partial h_{lj}} (d h_{ik})(d h_{lj})
\]

\[
= \sum_{i, k} u_\alpha(i) u_\alpha(k) \frac{d b_{ik}}{\sqrt{N}} - \frac{1}{2} \sum_{i, k} \frac{1}{\lambda_\alpha - \lambda_\beta} \left[ |u_\beta(i)|^2 |u_\alpha(k)|^2 + |u_\alpha(i)|^2 |u_\beta(k)|^2 \right] dt
\]

\[
= \frac{1}{\sqrt{N}} \sum_{i, k} u_\alpha(i) u_\alpha(k) d b_{ik} - \frac{\lambda_\alpha}{2} dt + \frac{1}{N} \sum_{\beta \neq \alpha} \frac{1}{\lambda_\alpha - \lambda_\beta} dt. \tag{12.12}
\]

In the second line we used \( b_{ik} = b_{ki} \) for \( i \neq k \) and that \((d h_{ik})(d h_{lj}) = \frac{1}{N} \delta_{ij} \delta_{kl} [1 + \delta_{ik}] dt \). Finally, in the last line we used the equation \( H u_\alpha = \lambda_\alpha u_\alpha \) to get

\[ \sum_{i, k} u_\alpha(i) u_\alpha(k) h_{ik} = \sum_i u_\alpha(i) (H u_\alpha)(i) = \lambda_\alpha \sum_i |u_\alpha(i)|^2 = \lambda_\alpha, \]

which gives the \( -(\lambda_\alpha / 2) dt \) term in (12.4).

In the first term on the right hand side of (12.12) we define a new real Gaussian process

\[ \tilde{B}_i := \sum_{i, k} u_\alpha(i) u_\alpha(k) b_{ik}. \]

Clearly, \( E d \tilde{B}_i = 0 \) and its covariance satisfies

\[ E d \tilde{B}_i d \tilde{B}_{i'} = \mathbb{E} \sum_{i, k} u_\alpha(i) u_\alpha(k) d b_{ik} u_{\alpha'}(i) u_{\alpha'}(j) d b_{lj} = 2 \sum_{i, k} u_\alpha(i) u_{\alpha'}(i) u_\alpha(k) u_{\alpha'}(k) dt = \delta_{\alpha \alpha'} 2 dt. \]

where there are two pairings, \((i, \ell) = (k, j)\) and \((i, \ell) = (j, k)\), that contributed to the contractions. Thus \( \tilde{B}_i = \sqrt{2} \tilde{B}_i \), where \( (B_i)_{i=1}^N \) is the standard (real) Brownian motion in \( \mathbb{R}^N \) and this gives the martingale term in (12.4) with \( \beta = 1 \). A similar formula can be derived for the Hermitian case, where the parameter becomes \( \beta = 2 \); we omit the details.

In the next section we put the Dyson Brownian motion in a more general context which is closer to the interpretation of GUE as an invariant ensemble in the spirit of Section 4. It turns out that the measure on the eigenvalues, explicitly given in (4.3), generates a DBM in a canonical way such that this measure will be invariant under the dynamics. This holds for any values \( \beta \geq 1 \), i.e., even if there is no underlying matrix ensemble.

### 12.3 Strong local ergodicity of the Dyson Brownian motion

One key property of DBM is that it leaves the Gaussian Wigner ensembles invariant. Moreover, the Gaussian measure is the only equilibrium of DBM and the DBM dynamics converges to this equilibrium from any initial condition. In this section we will quantify these ideas.

We start with introducing the concept of the Dirichlet form and the generator associated with any probability measure \( \mu = \mu_N \) on \( \mathbb{R}^N \). For definiteness, the reader may keep the invariant \( \beta \)-ensemble (4.3) in mind, i.e.,

\[ \mu_N(d\lambda) = \frac{e^{-\beta N H_N(\lambda)}}{Z} d\lambda, \quad H_N(\lambda) := \frac{1}{2} \sum_{i=1}^N V(\lambda_i) - \frac{1}{N} \sum_{i < j} \log |\lambda_j - \lambda_i|. \tag{12.13} \]

We will drop the subscript \( N \), but most quantities depend on \( N \).
We define the \textit{Dirichlet form} associated with the measure \( \mu \) on \( \mathbb{R}^N \), which is just the homogeneous \( H^1 \) norm, i.e.,

\[
D_{\mu}(f) := \frac{1}{\beta N} \sum_{i=1}^{N} \int (\partial_i f)^2 \, d\mu = \frac{1}{\beta N} \| \nabla f \|_{L^2(\mu)}^2, \quad (\partial_i \equiv \partial_{\lambda_i}). \tag{12.14}
\]

We remark that in our earlier papers the Dirichlet form was defined with a \( 1/(2N) \) prefactor instead of \( 1/(\beta N) \). The current convention is suitable to consider DBM for general \( \beta \).

The symmetric operator associated with the Dirichlet form is called \textit{generator} and denoted by \( \mathcal{L} = \mathcal{L}_{\mu} \). It satisfies

\[
D_{\mu}(f) = (f, (-\mathcal{L}) f)_{L^2(\mu)} = -\int f \mathcal{L} f \, d\mu.
\]

Notice that we follow the probabilistic convention to define the generator as a negative operator, \( \mathcal{L} \leq 0 \).

Formally, we have \( \mathcal{L} = \frac{1}{\beta N} \Delta - (\nabla \mathcal{H}) \cdot \nabla \), i.e.,

\[
\mathcal{L} = \sum_{i=1}^{N} \frac{1}{\beta N} \partial_i^2 + \sum_{i=1}^{N} \left( -\frac{1}{2} V'(\lambda_i) + \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_i - \lambda_j} \right) \partial_i. \tag{12.15}
\]

In the special Gaussian case, when the confining potential is \( V(\lambda) = \frac{1}{2} \lambda^2 \), the corresponding measure restricted to \( \Sigma_N \subset \mathbb{R}^N \) is denoted by \( \mu_G \). Notice that for \( \beta = 1, 2 \) the measure \( \mu_G \) coincides with the GOE, GUE measures, respectively (see also (4.12)). The generator (12.15) reads as

\[
\mathcal{L}_G = \sum_{i=1}^{N} \frac{1}{\beta N} \partial_i^2 + \sum_{i=1}^{N} \left( -\frac{1}{2} \lambda_i + \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_i - \lambda_j} \right) \partial_i. \tag{12.16}
\]

Now we consider the Dyson Brownian motion (12.4), i.e., dynamics of the eigenvalues \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \Sigma_N \) of \( H_t \) that evolves by (12.1). We write the distribution of \( \lambda \) of \( H_t \) at time \( t \) as \( f_t(\lambda) \mu_G(\lambda) \). Comparing the SDE (12.4) with \( \mathcal{L}_G \), we notice that Kolmorogov’s forward equation for the evolution of the density \( f_t \) takes the form

\[
\partial_t f_t = \mathcal{L}_G f_t, \quad (t \geq 0). \tag{12.17}
\]

We remark that the rigorous definition of \( \mathcal{L} \) as a self-adjoint operator on \( L^2(\mu) \) through the Friedrichs extension and the existence of the corresponding dynamics (12.17) restricted to the simplex \( \Sigma_N \) for any \( \beta \geq 1 \) will be discussed in the separate Section 12.4. In particular, by spectral theorem the operator \( (\mathcal{L})^{1/2} e^{t\mathcal{L}} \) is bounded for \( t > 0 \), thus we see that for any initial condition \( f_0 \in L^2(\mu) \) we have \( f_t \in H^1(\mu) \) for any \( t > 0 \). The restriction \( \beta \geq 1 \) is essential to define the dynamics on the simplex \( \Sigma_N \) without specifying additional boundary conditions; the same restriction was also necessary in Theorem 12.2 to define the strong solution to the DBM as a stochastic differential equation. Indeed, if \( \beta < 1 \), the particles cross each other in the SDE formulation. In this section we thus assume \( \beta \geq 1 \). Nevertheless, some ideas and results presented here are still applicable to any \( \beta > 0 \) with a regularization; we will comment on this in Section 13.7.

By writing the distribution of the eigenvalues as \( f_t \mu_G \), our formulation of the problem has already taken into account Dyson’s observation that the invariant measure for DBM is \( \mu_G \) since as we will see, at \( t \to \infty \) the solution \( f_t \) converges to \( f_\infty = 1 \). A natural question regarding the DBM is how fast the dynamics reaches equilibrium. Dyson had already posed this question in 1962:

\textbf{Dyson’s conjecture [45]:} The global equilibrium of DBM is reached in time of order one and the local equilibrium (in the bulk) is reached in time of order \( 1/N \). Dyson further remarked,

“The picture of the gas coming into equilibrium in two well-separated stages, with microscopic and macroscopic time scales, is suggested with the help of physical intuition. A rigorous proof that this picture is accurate would require a much deeper mathematical analysis.”

86
We will prove that Dyson’s conjecture is correct if the initial data of the DBM is the eigenvalues of a Wigner ensemble, which was Dyson’s original interest. Our result in fact is valid for DBM with much more general initial data that we now survey. Briefly, it will turn out that the global equilibrium is indeed reached within a time of order one, but local equilibrium is achieved much faster if an a priori estimate on the location of the eigenvalues (also called points) is satisfied. Recalling the definition of the classical locations $\gamma_j$ from (11.31), the a priori bound (originally referred to Assumption III in [63,64]) is formulated as follows:

**A priori Estimate:** There exists a $\xi > 0$ such that average rigidity on scale $N^{-1+\xi}$ holds, i.e.,

$$Q = Q_\xi := \sup_{0 \leq t \leq N} \frac{1}{N} \int \sum_{j=1}^{N} (\lambda_j - \gamma_j)^2 f_t(\lambda) \mu_G(d\lambda) \leq CN^{-2+2\xi}$$

(12.18)

with a constant $C$ uniformly in $N$. (We assumed (12.18) with a lower bound $t \geq 0$ in the supremum, but in fact the proof shows that $t \geq N^{-1+2\xi}$ would be sufficient.)

Notice that (12.18) requires that the rigidity of the eigenvalues on scale $N^{-1+\xi}$ holds, at least in an average sense. We recall that Theorem 11.5 guarantees that for Wigner eigenvalues rigidity holds on scale $N^{-1+2\xi}$ for any $\xi > 0$ (in the bulk), so (12.18) holds for any $\xi > 0$ in this case. However, our theory on the local ergodicity of the DBM applies to other models as well, so we wish to keep our formulation general and allowing situations where (12.18) holds only for larger $\xi$.

The main result on the local ergodicity of Dyson Brownian motion (12.17) states that if the a priori estimate (12.18) is satisfied then the local correlation functions of the measure $f_t \mu_G$ are the same as the corresponding ones for the Gaussian measure, $\mu_G = f_\infty \mu_G$, provided that $t$ is larger than $N^{-1+2\xi}$. The $n$-point correlation functions of the (symmetrized) probability measure $\nu$ are defined, similarly to (4.20), by

$$p^{(n)}_{p,\nu}(x_1, x_2, \ldots, x_n) := \int_{\mathbb{R}^{N-n}} \nu(x) dx_{n+1} \cdots dx_N, \quad x = (x_1, x_2, \ldots, x_N).$$

(12.19)

In particular, when $\nu = f_t \mu_G$, we denote the correlation functions by

$$p^{(n)}_{f_t,\mu_G}(x_1, x_2, \ldots, x_n) = p^{(n)}_{f_t,\mu_G}(x_1, x_2, \ldots, x_n).$$

(12.20)

We also use $p^{(n)}_{\mu_G}$ for $p^{(n)}_{\mu_G}$ in general, if $H$ is an $N \times N$ Wigner matrix, then we will also use $p^{(n)}_{H,\mu_G}$ for the correlation functions of the eigenvalues of $H$.

Due to the convention that one can view the locations of eigenvalues as the coordinates of particles (or points), we have used $x$, instead of $\lambda$, in the last equation. From now on, we will use both conventions depending on which viewpoint we wish to emphasize. Notice that the probability distribution of the eigenvalues at the time $t$, $f_t \mu_G$, is the same as that of the Gaussian divisible matrix:

$$H_t = e^{-t/2} H_0 + (1 - e^{-t})^{1/2} H^G,$$

(12.21)

where $H_0$ is the initial Wigner matrix and $H^G$ is an independent standard GUE (or GOE) matrix. This establishes the universality of the Gaussian divisible ensembles. The precise statement is the following theorem (notice that [64] uses somewhat different notation).

**Theorem 12.4.** [64, Theorem 2.1] Suppose that for some exponent $\xi \in (0, \frac{1}{4})$, the average rigidity (12.18) holds for the solution $f_t$ of the forward equation (12.17) on scale $N^{-1+\xi}$. Additionally, suppose that in the bulk the rigidity holds on scale $N^{-1+\xi}$ even without averaging, i.e., for any $\kappa > 0$

$$\sup_{N \geq j \leq (1-\kappa)N} |\lambda_j - \gamma_j| \propto N^{-1+\xi}$$

(12.22)

holds for any $t \in \{N^{-1+2\xi}, N\}$ if $N \geq N_0(\xi, \kappa)$ is large enough. Let $E \in (-2, 2)$ and $b = b_N > 0$ such that $[E - b, E + b] \subset (-2, 2)$. Then for any integer $n \geq 1$ and for any compactly supported smooth test function $O : \mathbb{R}^n \to \mathbb{R}$, we have, for any $t \in \{N^{-1+2\xi}, N\}$,

$$\left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha O(\alpha) \left[ p^{(n)}_{f_t,\mu_G}(x, \alpha) - p^{(n)}_{\mu_G}(x, \alpha) \right] \right| \leq N^{\xi} \left[ \frac{N^{-1+\xi}}{b} + \sqrt{\frac{1}{bNt}} \right] \|O\|_{C^1}$$

(12.23)

for any $N$ sufficiently large, $N \geq N_0(n, \xi, \kappa)$. 

87
The upper limit $t \leq N$ for the time interval in (12.18) and within the theorem is unimportant, one could have replaced it with $N^\varepsilon$ for any $\varepsilon > 0$. Beyond $t \geq N^\varepsilon$ the measure $f_t \mu_G$ is already exponentially close to equilibrium $\mu_G$ in entropy sense, so all correlation functions are also close. Also notice that for simplicity of the notation we replaced $\frac{\alpha}{\sqrt{b}}$ by $a$ in the argument of the correlation functions (compare (12.23) with (5.2)). This can be easily achieved by a change of variables and a redefinition of the test function $O$. This convention will be followed in the rest of the book.

In other words, this theorem states that if we have rigidity on scale $N^{-1+\xi}$ for some $\xi \in (0, \frac{1}{2})$, then the DBM has average energy universality in the bulk for any time $t \geq N^{-1+2\xi}$ on scale $b \gg \max\{N^{-1+\xi}, (Nt)^{-1}\}$. For generalized Wigner matrices we know that rigidity holds on the smallest possible scale, so we have the following corollary:

**Corollary 12.5.** Consider the matrix OU process with initial matrix ensemble $H_0$ being a generalized Wigner ensemble and let $|E| < 2$. Then for any $\varepsilon > 0$, for any $t \in [N^{-1+2\varepsilon}, N^\varepsilon]$ and for any $b \geq (Nt)^{-1}$ with $b < ||E|| - 2$ we have

$$\left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha \, O(\alpha) \left( \rho_{t,N}^{(a)} - \rho_{G,N}^{(a)} \right) \left( E' + \frac{\alpha}{N} \right) \right| \leq \frac{N^{3\varepsilon}}{\sqrt{bNt}} ||O||_{C^1},$$

(12.24)

for any $N \geq N_0(n, \varepsilon)$, where $\rho_{t,N}^{(a)}$ is the correlation function of $H_t$, or equivalently, that of $e^{-t/2}H_0 + \sqrt{1 - e^{-t}}H^G$.

**Proof.** Since $H_0$ is a generalized Wigner ensemble, so is $e^{-t/2}H_0 + \sqrt{1 - e^{-t}}H^G$ for all $t$. Hence the optimal rigidity estimate (11.32) holds for all times. Therefore, (12.18) and (12.22) hold with any $\xi > 0$. Choose $\xi = \varepsilon/2$ and apply Theorem 12.4. The right hand side of (12.23) is bounded by

$$N^\varepsilon \left[ \frac{N^{3\varepsilon}}{N^\varepsilon} + \frac{1}{\sqrt{bNt}} \right] \leq \frac{N^{3\varepsilon}}{\sqrt{bNt}},$$

where we have used that $t \leq N^\varepsilon$ and $bNt \geq 1$. \qed

The estimate (12.24) means that averaged energy universality holds in a window of size slightly bigger than $tN$. A large energy window of size $b = N^{-\delta}$, for some small $\delta > 0$, can already be achieved after a very short times, i.e., for any $t \geq N^{-1+2\delta}$ (by choosing $\varepsilon = \delta/8$). To achieve universality on very short energy scales, $b = N^{-1+\delta}$, we will need relatively large times, $t \geq N^{-\delta/2}$ (by choosing $\varepsilon = \delta/16$). In both cases the right hand side (12.24) is bounded by $CN^{-\varepsilon}$. In the sense of universality of large energy windows of size $b = N^{-\delta}$, our result essentially establishes the Dyson’s conjecture that the time to local equilibrium is $N^{-1}$. A better error estimate can also be achieved if both $b$ and $t$ are large, e.g., with the choice $b = N^{-\delta}$ and $t \geq N^{-\delta}$, we get a convergence of order $N^{-1+2\delta+\varepsilon}$ for any $\varepsilon > 0$.

Going back to Theorem 12.4, we also remark that there is a considerable room in this argument even if optimal rigidity is not available (this is the case for random band matrices, see [55]). In fact, Theorem 12.4 provides universality, albeit only for relatively large times $t \geq N^{-\delta}$ and with a large energy averaging, $b = N^{-\delta}$, even if (12.18) and (12.22) hold only with $\xi = \frac{1+\delta}{2}$. This restriction comes from the requirement that $t \geq N^{-1+2\xi}$ should be implied by $t \geq N^{-\delta}$. This exponent $\xi$ slightly above 1/2 corresponds to a rigidity control on a scale just below $N^{-1/2}$ in the bulk. Clearly, $N^{-1/2}$ is a critical threshold; no local universality can be concluded with this argument unless a rigidity control on scale below $N^{-1/2}$ is established a priori.

### 12.4 Existence and restriction of the dynamics

This technical section was taken from Appendix A of [64]; we include it here for the reader’s convenience.

As in Section 12.3, we consider the Euclidean space $\mathbb{R}^N$ with the normalized measure $\mu = \exp(-\beta NH)/Z$. The Hamiltonian $H$, of the form (12.13), is symmetric with respect to the permutation of the variables $x = (x_1, \ldots, x_N)$, thus the measure can be restricted to the subset $\Sigma_N \subset \mathbb{R}^N$ defined in (12.3). In this appendix we outline how to define the dynamics (12.17) with its generator, formally given by $L = \frac{1}{N} \Delta - \frac{1}{2} (\nabla H) \nabla$, 

88
on ΣN. The condition β ≥ 1 and the specific factors \( \prod_{i<j} |x_j - x_i|^β \) will play a key role in the argument, in particular, we will see that β = 1 is the critical threshold for this method to work.

We first recall the standard definition of the dynamics on \( \mathbb{R}^N \). The quadratic form

\[
E(u, v) := \int_{\mathbb{R}^N} \nabla u \cdot \nabla v \, d\mu
\]

is a closable Markovian symmetric form on \( L^2(\mathbb{R}^N, d\mu) \) with a domain \( C_0^\infty(\mathbb{R}^N) \) (see Example 1.2.1 and Theorem 3.1.3 of [74]). This form can be closed with a form domain \( H^1(\mathbb{R}^N, d\mu) \) defined as the closure of \( C_0^\infty \) in the norm \( \| \cdot \|_2^2 = E(\cdot, \cdot) + \| \cdot \|_2^2 \). The closure is called the Dirichlet form. It generates a strongly continuous Markovian semigroup \( T_t, t > 0 \), on \( L^2 \) (Theorem 1.4.1 [74]) and it can be extended to a contraction semigroup to \( L^1(\mathbb{R}^N, d\mu) \), \( \|T_t f\|_1 \leq \| f\|_1 \) (Section 1.5 [74]). The generator \( L \) of the semigroup, is defined via the Friedrichs extension (Theorem 1.3.1 [74]) and it is a positive self-adjoint operator on its natural domain \( D(L) \) with \( C_0^\infty \) being the core. The generator is given by \( L = \frac{1}{2\mathbb{R}^N} \Delta - \frac{1}{2}(\nabla \mathcal{H}) \nabla \) on its domain (Corollary 1.3.1 [74]). By the spectral theorem, \( T_t \) maps \( L^2 \) into \( D(L) \), thus with the notation \( f_t = T_t f \) for some \( f \in L^2 \), it holds that

\[
\partial_t f_t = L f_t, \quad t > 0, \quad \text{and} \quad \lim_{t \to 0^+} \| f_t - f \|_2 = 0. \tag{12.25}
\]

Moreover, by approximating \( f \) by \( L^2 \) functions and using that \( T_t \) is contraction in \( L^1 \) (Section 1.5 in [74]), the differential equation holds even if the initial condition \( f \) is only in \( L^1 \). In this case the convergence \( f_t \to f \), as \( t \to 0^+ \), holds only in \( L^1 \). We remark that \( T_t \) is also a contraction on \( L^\infty \), by duality.

Now we restrict the dynamics to \( \Sigma = \Sigma_N \) equipped with the probability measure \( \mu_N(dx) := N! \mu(dx) \cdot 1(x \in \Sigma) \). Here \( \mu_N \) is from (12.13) and the factor \( N! \) restores the normalization after restriction to \( \Sigma \). Repeating the general construction with \( \mathbb{R}^N \) replaced by \( \Sigma_N \), we obtain the corresponding generator \( L^{(2)} \) and the semigroup \( T_t^{(2)} \) on the space \( H^1(\Sigma, d\mu_N) \) that is defined as the closure of \( C_0^\infty(\Sigma) \) with respect to the norm \( \| \cdot \|_2^2 = E(\cdot, \cdot) + \| \cdot \|_2^2 \) on \( \Sigma \).

To establish the relation between \( L \) and \( L^{(2)} \), we first define the symmetrized version of \( \Sigma \)

\[
\tilde{\Sigma} := \mathbb{R}^N \setminus \left\{ x : \exists i \neq j \text{ with } x_i = x_j \right\}.
\]

Denote \( X := C_0^\infty(\tilde{\Sigma}) \). The key information is that \( X \) is dense in \( H^1(\mathbb{R}^N, d\mu) \) which is equivalent to the density of \( X \) in \( C_0^\infty(\mathbb{R}^N, d\mu) \). We will check this property below. Then the general argument above directly applies if \( \mathbb{R}^N \) is replaced by \( \Sigma_N \) and it shows that the generator \( L \) is the same (with the same domain) if we start from \( \mathbb{R}^N \) instead of \( C_0^\infty(\mathbb{R}^N, d\mu) \) as a core.

Note that both \( L \) and \( L^{(2)} \) are local operators and \( L \) is symmetric with respect to the permutation of the variables. For any function \( f \) defined on \( \Sigma \), we define its symmetric extension onto \( \tilde{\Sigma} \) by \( \tilde{f} \). Clearly, \( L \tilde{f} = L^{(2)} f \) for any \( f \in C_0^\infty(\Sigma) \). Since the generator is uniquely determined by its action on its core, and the generator uniquely determines the dynamics, we see that for any \( f \in L^1(\Sigma, d\mu) \), one can determine \( T_t^{(2)} f \) by computing \( T_t \tilde{f} \) and restricting it to \( \Sigma \). In other words, the dynamics \((12.17)\) is well defined when restricted to \( \Sigma = \Sigma_N \).

Finally, we have to prove the density of \( X \) in \( C_0^\infty(\mathbb{R}^N, d\mu) \), i.e., to show that \( f \in C_0^\infty(\mathbb{R}^N) \), then there exists a sequence \( f_n \in C_0^\infty(\tilde{\Sigma}) \) such that \( E(f - f_n, f - f_n) \to 0 \). The structure of \( \tilde{\Sigma} \) is complicated since in addition to the one codimensional coalescence hyperplanes \( x_i = x_j \) (and \( x_i = 0 \) in case of \( \Sigma^{(1)} \)), it contains higher order coalescence subspaces with higher codimensions. We will show the approximation argument in a neighborhood of a point \( x \) such that \( x_i = x_j \) but \( x_i \neq x_k \) for any other \( k \neq i, j \). The proof uses the fact that the measure \( d\mu \) vanishes at least to first order, i.e., at least as \( |x_i - x_j| \), around \( x \), thanks to \( \beta \geq 1 \). This is the critical case; the argument near higher order coalescence points is even easier, since they have lower codimension and the measure \( \mu \) vanishes at even higher order.

In a neighborhood of \( x \) we can change to local coordinates such that \( r := x_i - x_j \) remains the only relevant coordinate. Thus the task is equivalent to show that any \( g \in C_0^\infty(\mathbb{R}) \) can be approximated by a sequence \( g_\varepsilon \in C_0^\infty(\mathbb{R} \setminus \{0\}) \) in the sense that

\[
\int_{\mathbb{R}} |g'(r) - g'_\varepsilon(r)|^2 |r|dr \to 0 \tag{12.26}
\]
as \( \varepsilon \to 0 \) over a discrete set, e.g. \( \varepsilon = \varepsilon_n = 1/n \). It is sufficient to consider only the positive semi-axis, i.e., \( r > 0 \). More precisely, define

\[
ge_{\varepsilon}(r) := g(r)\phi_\varepsilon(r),
\]

where \( \phi_\varepsilon \) is a continuous cutoff function defined as \( \phi_\varepsilon(r) = 0 \) for \( r \leq \varepsilon^2 \), \( \phi_\varepsilon(r) = 1 \) for \( r \geq \varepsilon \) and

\[
\phi_\varepsilon(r) = C_\varepsilon (\log r - 2 \log \varepsilon), \quad \varepsilon^2 \leq r \leq \varepsilon, \quad C_\varepsilon = \frac{1}{\log \varepsilon - \log \varepsilon^2} = |\log \varepsilon|^{-1}.
\]

Simple calculation shows that

\[
\int_0^\infty |g'(r) - g'_\varepsilon(r)|^2 r dr \leq C \int_0^\infty |g'(r)|^2 |1 - \phi_\varepsilon(r)|^2 r dr + C \int_0^\infty |g(r)|^2 |\phi'_\varepsilon(r)|^2 r dr.
\]

The first term converges to zero as \( \varepsilon \to 0 \) by dominated convergence since \( g \in H^1(\mathbb{R}^+, r dr) \) and \( \phi_\varepsilon \to 1 \) pointwise. The second term is bounded by

\[
\int_0^\infty |g(r)|^2 |\phi'_\varepsilon(r)|^2 r dr \leq C_\varepsilon^2 \|g\|_{L^2}^2 \int_{\varepsilon^2}^\varepsilon \frac{1}{r} r^2 dr = \frac{\|g\|_{L^2}^2}{|\log \varepsilon|},
\]

which also tends to zero since \( g \in L^\infty \). This shows that \( g_\varepsilon \) converges to \( g \) in \( H^1 \). The function \( g_\varepsilon \) is not yet smooth, but we can smooth it on a scale \( \delta \ll \varepsilon^2 \) by taking the convolution with a smooth function \( \eta_\delta \) compactly supported in \([-\delta, \delta]\). It is an elementary exercise to see that \( \eta_\delta * g_\varepsilon \) converges to \( g_\varepsilon \) in \( H^1 \) as \( \delta \to 0 \). This verifies (12.26). The same proof shows that \( C_0^\infty(\mathbb{R}_+ \setminus \{0\}) \) is dense in \( H^1(\mathbb{R}_+, r dr) \cap L^\infty(\mathbb{R}_+, r dr) \). This completes the proof that for \( \beta \geq 1 \) the dynamics is well defined in the space \( H^1(\Sigma, d\mu_\Sigma) \).

In fact, the boundedness condition \( f \in L^\infty(\mathbb{R}_+, r dr) \) can be removed in the above argument and one can show that \( C_0^\infty(\mathbb{R}_+ \setminus \{0\}) \) is dense in \( H^1(\mathbb{R}_+, r dr) \). This is a type of Meyers-Serrin theorem concerning the equivalence of two possible definitions of Sobolev spaces on the measure space \( (\mathbb{R}_+, r dr) \): the closure of \( C_0^\infty \) functions coincides with the set of functions with finite \( H^1 \)-norm. Another formulation is the following: If we extend the functions on \( \mathbb{R}_+ \) to two dimensional radial functions, i.e., \( G(x) := g(|x|), G_\varepsilon(x) = g_\varepsilon(|x|) \) for \( x \in \mathbb{R}^2 \), then this statement is equivalent to the fact that a point in two dimensions has zero capacity.

To see this stronger claim, we need a different cutoff function in (12.27). Let us now define

\[
\phi_\varepsilon(r) := \log(a + b \log r), \quad \varepsilon^2 \leq r \leq \varepsilon,
\]

with \( b := (1 - \varepsilon)/\log \varepsilon \), \( a := 2\varepsilon - 1 \), and \( \phi_\varepsilon(r) := 0 \) for \( r \leq \varepsilon^2 \), and \( \phi_\varepsilon(r) := 1 \) for \( r \geq \varepsilon \). The first term on the right hand side of (12.29) goes to zero as before. For the last term in (12.29), we have

\[
b^2 \int_{\varepsilon^2}^{\varepsilon} \frac{|g(r)|^2}{r^2(a + b \log r)^2} r dr = b^2 \int_{\varepsilon^2}^{\varepsilon} \frac{|g(r)|^2}{r^2(a + b \log r)^2} r dr.
\]

We view this as an integral on \( \mathbb{R}^2 \), by radially extending the function \( g \) to \( \mathbb{R}^2 \) by \( G(y) := g(|y|) \) for any \( y \in \mathbb{R}^2 \). Then the LHS above gives

\[
b^2 \int_{\varepsilon^2}^{\varepsilon} \frac{|g(r)|^2}{r^2(a + b \log r)^2} r dr = b^2 \int_{\varepsilon^2 \leq |y| \leq \varepsilon} \frac{|G(y)|^2}{|y|^2(a + b \log |y|)^2} dy \leq C \int_{\varepsilon^2 \leq |y| \leq \varepsilon} \frac{|G(y)|^2}{|y|^2|\log |y||^2} dy.
\]

Define the sequence \( \varepsilon_k := 2^{-2^k} \), clearly \( \varepsilon_k \) is \( \varepsilon_{k+1} \). Setting

\[
I_k := \int_{\varepsilon_{k+1} \leq |y| \leq \varepsilon_k} \frac{|G(y)|^2}{|y|^2|\log |y||^2} dy,
\]

we clearly have

\[
\sum_{k=1}^{\infty} I_k = \int_{|y| \leq 1/2} \frac{|G(y)|^2}{|y|^2|\log |y||^2} dy \leq C \int_{\mathbb{R}^2} |\nabla G|^2 = C' \int_0^\infty |g'(r)|^2 r dr,
\]

90
where the last step we used the critical Hardy inequality with logarithmic correction (see, e.g. Theorem 2.8 in [48]). Since in our case $g \in H^1(\mathbb{R}^+, r\,dr)$, from the last displayed inequality we have $I_k \to 0$ as $k \to \infty$. Therefore, we can use the cutoff function $\phi_{\epsilon_k}$ to prove the claim.

The Meyers-Serrin theorem we just proved for $(R^+, r\,dr)$ can be extended to $(\mathbb{R}^N, d\mu)$ using local changes of coordinates if $\beta \geq 1$. Thus one may prove that the form domain of the operator $L^{(1)}$ is $H^1(\Sigma, d\mu_{\Sigma})$ and the dynamics (12.25) is well defined on this space.
13 Entropy and the logarithmic Sobolev inequality (LSI)

13.1 Basic properties of the entropy

In this section, we prove a few well-known inequalities for the entropy which will be used in next section for the logarithmic Sobolev inequality (LSI). In this section we work on an arbitrary probability space \((\Omega, \mathcal{B})\) with an appropriate \(\sigma\)-algebra. We will use the probabilistic and the analysis notations and concepts in parallel. In particular, we interchangeably use the concept of random variables and measurable functions on \(\Omega\). Similarly, both the expectation \(E^\mu[f]\) and the integral notation \(\int f d\mu\) will be used. We will drop the \(\Omega\) in the notation.

Definition 13.1. For any two probability measures \(\nu, \mu\) on the same probability space, we define the relative entropy of \(\nu\) w.r.t. \(\mu\) by

\[
S(\nu | \mu) := \int \log \frac{d\nu}{d\mu} d\mu = \int \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} d\mu
\]

if \(\nu\) is absolutely continuous with respect to \(\mu\) and we set \(S(\nu | \mu) := \infty\) otherwise.

Applying the definition (13.1) to \(\nu = f \mu\) for any probability density \(f\) with \(\int f d\mu = 1\), we may also write

\[
S_\mu(f) := S(f \mu | \mu) = \int f(\log f) d\mu.
\]

In most cases, the reference measure \(\mu\) will be canonical (e.g. the natural equilibrium measure), so often we will drop the subscript \(\mu\) if there is no confusion. We may then call \(S_\mu(f) = S(f)\) the entropy of \(f\). Using the convexity of the function \(x \mapsto x \log x\) on \(\mathbb{R}_+\), a simple Jensen’s inequality,

\[
0 = \left( \int f d\mu \right) \log \left( \int f d\mu \right) \leq \int f \log f d\mu,
\]

shows that the relative entropy is always non-negative, \(S(\nu | \mu) \geq 0\).

Proposition 13.2 (Entropy inequality or Gibbs inequality). Let \(\mu, \nu\) be two probability measures on a common probability space and let \(X\) be a random variable. Then, for any positive number \(\alpha > 0\), the following inequality holds as long as the right hand side is finite.

\[
E^\nu[X] \leq \alpha^{-1} S(\nu | \mu) + \alpha^{-1} \log E^\mu e^{\alpha X}. \tag{13.2}
\]

Proof. First note that we only have to prove the case \(\alpha = 1\) since we can redefine \(\alpha X \to X\). From the concavity of the logarithm and Jensen’s inequality, we have

\[
\int X d\nu - S(\nu | \mu) = \int \log \left[ e^{X} \frac{d\mu}{d\nu} \right] d\nu \leq \log E^\nu e^X,
\]

and this proves (13.2).

Although (13.2) is just an inequality, there is a variational characterization of the relative entropy behind it. Namely,

\[
S(\nu | \mu) = \sup_X \left[ E^\nu[X] - \log E^\mu e^X \right], \tag{13.3}
\]

where the supremum is over all bounded random variables. As we will not need this relation in this book, we leave it to the interesting readers to prove it.

As a corollary of (13.2) we mention that for any set \(A\) we have the bound

\[
\mathbb{P}^\nu(A) \leq \frac{\log 2 + S(\nu | \mu)}{\log \frac{1}{\mathbb{P}^{\nu}(A)}}. \tag{13.4}
\]
The proof is left as an exercise. This inequality will be used in the following context. Suppose that the relative entropy of \( \nu \) with respect to \( \mu \) is finite. Then in order to show that a set has a small probability w.r.t. \( \nu \), we need to verify that this set is exponentially small w.r.t. \( \mu \). In this sense, entropy provides only a relatively weak link between two measures. However, it is still stronger than the total variational norm which we will show now.

For any two probability measures \( f\mu \) and \( \mu \), the \( L^p \) distance between \( f\mu \) and \( \mu \) is defined by

\[
\left[ \int |f - 1|^p \, d\mu \right]^{1/p}.
\]

(13.5)

When \( p = 1 \), it is called the total variational norm between \( f\mu \) and \( \mu \). Entropy is a weaker measure of distance between two probability measures than the \( L^p \) distance for any \( p > 1 \) but stronger than the total variational norm. The former statement can be expressed, for example, by the elementary inequality

\[
\int f \log f \, d\mu \leq 2 \left[ \int |f - 1|^p \, d\mu \right]^{1/p} + \frac{2}{p - 1} \int |f - 1|^p \, d\mu, \quad p > 1,
\]

(13.6)

whose proof is left as an exercise. (There is a more natural way to compare \( L^p \) distance and the relative entropy in terms of the \( L^{\log L} \) Orlicz norm but we will not use this norm in this book.) The latter statement will be made precise in the following proposition. Furthermore, we also remark the following easy relation among \( L^p \) norms and the entropy:

\[
\left. \frac{d}{dp} \right|_{p=1} \left[ \int f^p \, d\mu \right]^{1/p} = \int f \log f \, d\mu
\]

(13.7)

holds for any probability density \( f \) w.r.t. \( \mu \).

**Proposition 13.3** (Entropy and total variation norm, Pinsker inequality). Suppose that \( \int f \, d\mu = 1 \) and \( f \geq 0 \). Then we have

\[
2 \left[ \int |f - 1| \, d\mu \right]^2 \leq \int f \log f \, d\mu.
\]

(13.8)

*Proof.* By variational principle, we first rewrite

\[
\int |f - 1| \, d\mu = \sup_{|g| \leq 1} \int fg \, d\mu - \int g \, d\mu.
\]

(13.9)

For any such function \( g \), we have by the entropy inequality (13.2) that for any \( t > 0 \)

\[
\int fg \, d\mu - \int g \, d\mu \leq t^{-1} \log \int e^{tg} \, d\mu - \int e^{tg} \, d\mu + t^{-1} \int f \log f \, d\mu.
\]

(13.10)

We now define the function

\[
h(t) := \log \int e^{tg} \, d\mu
\]

for any \( t \geq 0 \). A simple calculation shows that the second derivative of \( h \) is given by

\[
h''(t) = \langle g; g \rangle_\omega, \quad \omega_t := \frac{e^{tg} \mu}{\int e^{tg} \, d\mu},
\]

where \( \omega_t \) is a probability measure and

\[
\langle f; g \rangle_\omega := \int fg \, d\omega - \left( \int f \, d\omega \right) \left( \int g \, d\omega \right)
\]

denotes the covariance. Recall that the covariance is positive definite, i.e., it satisfies the usual Schwarz inequality,

\[
|\langle f; g \rangle_\omega|^2 \leq \langle f; f \rangle_\omega \langle g; g \rangle_\omega.
\]

93
Since $|g| \leq 1$, we have $0 \leq (g; g)_{\omega_1} \leq 1$, so $h''(t) \leq 1$. Thus by Taylor’s theorem, $h(t) \leq h(0) + h'(0) + t^2/2$, i.e.,

$$t^{-1} \log \int e^{tg} d\mu \leq \int g d\mu + t/2.$$ 

Together with (13.10), we have

$$\int f g d\mu - \int g d\mu \leq t/2 + t^{-1} \int f \log f d\mu \leq \sqrt{\frac{1}{2} \int f \log f d\mu},$$

where we optimized $t$ in the last step. Since this bound holds for any $g$ with $|g| \leq 1$, using (13.9) we have proved (13.8).

### 13.2 Entropy on product spaces and conditioning

Now we consider a product measure, i.e., we assume that the probability space has a product structure, $\Omega = \Omega_1 \times \Omega_2$ and $\mu = \mu_1 \otimes \mu_2$, $\nu = \nu_1 \otimes \nu_2$, where $\mu_j$ and $\nu_j$ are probability measures on $\Omega_j$, $j = 1, 2$. For simplicity, we denote the elements of $\Omega$ by $(x, y)$ with $x \in \Omega_1$, $y \in \Omega_2$. Writing $\nu_j = f_j \mu_j$, clearly we have

$$S(\nu_1 \otimes \nu_2|\mu_1 \otimes \mu_2) = S_{\mu_1 \otimes \mu_2}(f_1 \otimes f_2) = \int_{\Omega_1 \times \Omega_2} f_1(x) f_2(y) [\log f_1(x) + \log f_2(y)] \mu_1(dx) \mu_2(dy)$$

$$= S_{\mu_1}(f_1) + S_{\mu_2}(f_2) = S(\nu_1|\mu_1) + S(\nu_2|\mu_1), \quad (13.11)$$

i.e., the entropy is additive for product measures.

This property of the entropy makes it an especially suitable tool to measure closeness in very high dimensional analysis. For example, if the probability space is a large product space $\Omega = \Omega_1^N$ equipped with two product measures $\mu = \mu_1 \otimes N$ and $\nu = f^\otimes N \mu$, then their relative entropy

$$S_{\mu}(f^\otimes N) = NS_{\mu_1}(f) \quad (13.12)$$

grows only linearly in $N$, the number of degrees of freedom. It is easy to check that any $L^p$ distance (13.5) for $p > 1$ grows exponentially in $N$, which usually renders it useless. Thus the entropy is still stronger than the $L^1$ norm, but its growth in $N$ is much more manageable. An additional advantage is that the entropy is often easier to compute than any other $L^p$-norm (with the exception of $L^2$).

Next we discuss how the entropy decomposes under conditioning. For simplicity, we assume the product structure, $\Omega = \Omega_1 \times \Omega_2$ as before. Let $\omega$ be a probability measure on the product space $\Omega$. For any integrable function (random variable) $u(x, y)$ on $\Omega$ we denote its **conditional expectation** by

$$\hat{u} = \mathbb{E}^\omega[u|\mathcal{F}_1], \quad (13.13)$$

where $\mathcal{F}_1$ is the $\sigma$-algebra of $\Omega_1$ canonically lifted to $\Omega$. Note that $\hat{u} = \hat{u}(x)$ depends only on the $x$ variable.

The conditional expectation $\hat{u}$ may also be characterized by the following relation: it is the unique measurable function on $\Omega_1$ such that for any bounded measurable (w.r.t. $\mathcal{F}_1$) function $O(x)$ the following identity holds:

$$\mathbb{E}^\omega[\hat{u} O] = \mathbb{E}^\omega[u O].$$

Written out in coordinates, it means that

$$\int \int \hat{u}(x) O(x) \omega(dx\,dy) = \int \int u(x, y) O(x) \omega(dx\,dy). \quad (13.14)$$

In particular, if $\omega(dx\,dy) = \omega(x, y) dx\,dy$ is absolutely continuous with respect to some reference product measure $dx\,dy$ on $\Omega$, then we have, somewhat formally,

$$\hat{u}(x) = \frac{\int_{\Omega_2} u(x, y) \omega(x, y) dy}{\int_{\Omega_2} \omega(x, y) dy}. \quad (13.15)$$

94
The notation \( dx \, dy \) for the reference measure already indicates that in our applications \( \Omega_j \) will be Euclidean spaces \( \mathbb{R}^{n_j} \) of some dimension \( n_j \), and the reference measure will be the Lebesgue measure.

We remark that the concept of conditioning can be defined in full generality with respect to any sub-\( \sigma \)-algebra; the product structure of \( \Omega \) is not essential. However, we will not need the general definition in this book and the above definition is conceptually simpler.

The conditional expectation gives rise to a trivial martingale decomposition:

\[
u = \hat{u} + (u - \hat{u}) ,
\]

where \( \hat{u} \) is \( \mathcal{F}_1 \)-measurable, while \( u - \hat{u} \) has zero expectation on any \( \mathcal{F}_1 \)-measurable set. Subtracting the expectation \( \bar{u} := \mathbb{E}^\omega u = \mathbb{E}^\omega \hat{u} \) and squaring this formula, we have the **martingale decomposition of the variance** of \( u \):

\[
\text{Var}^\omega(u) := \mathbb{E}^\omega[(u - \hat{u})^2] = \mathbb{E}^\omega(u - \hat{u})^2 + \mathbb{E}^\omega(\hat{u} - \bar{u})^2 = \mathbb{E}^\omega \text{Var}(u(x, \cdot)) + \mathbb{E}^\omega (\hat{u}) ,
\]

where we defined the conditional variance

\[
\text{Var}(u(x, \cdot)) := \mathbb{E}[u(x, \cdot)^2|\mathcal{F}_1] = \mathbb{E}[u^2|\mathcal{F}_1] - [\hat{u}]^2 .
\]

The identity (13.17) is a triviality, but its interpretation is important. It means that the variance is additive w.r.t. the martingale decomposition (13.16). The first term \( \mathbb{E}^\omega \text{Var}(u(x, \cdot)) \) is the expectation of the variance w.r.t. \( y \) conditioned on \( x \); the second term \( \mathbb{E}^\omega (\hat{u} - \bar{u})^2 \) is the variance of the marginal w.r.t. \( x \). In other words, we can compute the variance one by one.

The martingale decomposition has an analogue for the entropy. For simplicity, we assume that \( \omega \) has a density \( \omega(x,y) \) w.r.t. a reference measure \( dx \, dy \). Denote by \( \tilde{\omega} \) the marginal probability density on \( \Omega_1 \),

\[
\tilde{\omega}(x) = \int_{\Omega_2} \omega(x,y)dy ,
\]

and by \( \hat{f} \) the marginal density of \( f\omega \) w.r.t. \( \tilde{\omega} \), i.e.,

\[
\hat{f}(x) = \frac{\int_{\Omega_2} f(x,y)\omega(x,y)dy}{\tilde{\omega}(x)} .
\]

In particular, for any test function \( O(x) \) we have

\[
\int_{\Omega_1} \int_{\Omega_2} O(x)f(x,y)\omega(x,y)dxdy = \int_{\Omega_1} O(x)\hat{f}(x)\tilde{\omega}(x)dx .
\]

(13.18)

Let

\[
\omega_x(y) := \frac{\omega(x,y)}{\omega(x)}
\]

be the probability density on \( \Omega_2 \) conditioned on a fixed \( x \in \Omega_1 \). Define

\[
f_x(y) := \frac{f(x,y)}{\hat{f}(x)}
\]

(13.19)

to be the corresponding density of the measure \( f\omega \) conditioned on a fixed \( x \in \Omega_1 \). Note that with these definitions we have

\[
(f\omega)_x(y) = \frac{f(x,y)\omega(x,y)}{\int_{\Omega_2} f(x,y)\omega(x,y)dy} = f_x(y)\omega_x(y) .
\]

Now we are ready to state the following proposition on the additivity of entropy w.r.t. martingale decomposition:
Proposition 13.4. With the notations above, we have

$$S_\omega(f) = \mathbb{E} [\omega \log \omega] S_\omega (f_x) + S_\omega (\hat{f}) \quad (13.20)$$

In particular, the marginal entropy is bounded by the total entropy, i.e.,

$$S_\omega (\hat{f}) \leq S_\omega (f) \quad (13.21)$$

In other words, the entropy is additive w.r.t. the martingale decomposition, i.e., the entropy is the sum of the expectation of the entropy in $y$ conditioned on $x$ and the entropy of the marginal in $x$. The additivity of entropy in this sense is an important tool in the application of the LSI as we will demonstrate shortly. It also indicates that entropy is an extensive quantity, i.e., the entropy of two probabilities on a product space $\Omega^N$ is often of order $N$, generalizing the formulas (13.11)-(13.12) to non-product measures.

Proof. We can decompose the entropy as follows:

$$S_\omega(f) = \int f \log f \, d\omega = \int f \log(f/\hat{f}) \, d\omega + \int f \log \omega \, d\omega = \int f \log(f/\hat{f}) \, d\omega + \int \hat{f} \log \omega \, d\omega, \quad (13.22)$$

where in the last step we used (13.18). The first term can be rewritten as

$$\int f \log(f/\hat{f}) \, dx = \int dx \omega (x) \hat{f}(x) = \mathbb{E} [\omega \log \omega] S_\omega (f_x). \quad (13.23)$$

We have thus proved (13.20).

13.3 Logarithmic Sobolev inequality

The logarithmic Sobolev inequality requires that the underlying probability space admits a concept of differentiation. While the theory can be developed for more general spaces, we restrict our attention to the probability space $\Omega = \mathbb{R}^N$ in this subsection. We will work with probability measures $\mu$ on $\mathbb{R}^N$ that are defined by a Hamiltonian $\mathcal{H}$:

$$d\mu(x) = \frac{e^{-\mathcal{H}(x)}}{Z} \, dx, \quad (13.24)$$

where $Z$ is a normalization factor. In Section 13.7 we will comment on how to extend the results of this section to certain subsets of $\mathbb{R}^N$, most importantly, to the simplex $\Sigma_N = \{ x \in \mathbb{R}^N : x_1 < x_2 < \ldots < x_N \}$.

Note that for simplicity in this presentation we neglect the $\beta N$ prefactor compared with (12.13). Let $\mathcal{L}$ be the generator of the dynamics associated with the Dirichlet form

$$D(f) = D_{\mu} (f) := -\int f (\mathcal{L} g) \, d\mu = \int |\nabla f|^2 \, d\mu = \sum_{j=1}^N \int (\partial_j f)^2 \, d\mu, \quad \partial_j = \partial x_j. \quad (13.25)$$

Formally we have $\mathcal{L} = \Delta - (\nabla \mathcal{H}) \cdot \nabla$. The operator $\mathcal{L}$ is symmetric with respect to the measure $d\mu$, i.e.,

$$\int f (\mathcal{L} g) \, d\mu = \int (\mathcal{L} f) g \, d\mu = -\int \nabla f \cdot \nabla g \, d\mu. \quad (13.26)$$

We remark that in many books on probability, e.g. [43], the Dirichlet form (13.25) is defined with a factor $1/2$, but this convention is not compatible with the $1/(\beta N)$ prefactor in (12.14). The lack of this $1/2$ factor in (13.25) causes slight deviations from their customary form in the following theorems.

Definition 13.5. The probability measure $\mu$ on $\mathbb{R}^N$ satisfies the logarithmic Sobolev inequality if there exists a constant $\gamma$ such that

$$S(f) = \int f \log f \, d\mu \leq \gamma \int |\nabla \sqrt{f}|^2 \, d\mu = \gamma D(\sqrt{f}) \quad (13.27)$$

holds for any smooth density function $f \geq 0$ with $\int f \, d\mu = 1$. The smallest such $\gamma$ is called the logarithmic Sobolev inequality constant of the measure $\mu$. 

96


A simple density argument shows that (13.27) extends from smooth functions to any nonnegative function \( f \in C_0^\infty(\mathbb{R}^N) \), the space of smooth functions with compact support. In fact, it is easy to extend it to all \( \sqrt{f} \in H^1(\text{d}\mu) \). Since we will not use the space \( H^1(\text{d}\mu) \) in this book, we will just use \( f \in C_0^\infty(\mathbb{R}^N) \) in the LSI.

**Theorem 13.6** (Bakry-Émery). [13] Consider a probability measure \( \mu \) on \( \mathbb{R}^N \) of the form (13.24), i.e., 
\[
\mu = e^{-\mathcal{H}}/Z.
\]
Suppose that a convexity bound holds for the Hamiltonian, i.e., with some positive constant \( K \) we have
\[
\nabla^2\mathcal{H}(x) \geq K \tag{13.28}
\]
for any \( x \) (in the sense of quadratic forms). Then the logarithmic Sobolev inequality holds (13.27) with an LSI constant \( \gamma \leq 2/K \), i.e.,
\[
S(f) \leq \frac{2}{K} D(\sqrt{f}), \quad \text{for any density } f \text{ with } \int f \text{d}\mu = 1. \tag{13.29}
\]
Furthermore, the dynamics
\[
\partial_t f_t = \mathcal{L} f_t, \quad t > 0, \tag{13.30}
\]
relaxes to equilibrium on the time scale \( t \asymp 1/K \) both in the sense of entropy and Dirichlet form:
\[
S(f_t) \leq e^{-2Kt} S(f_0), \quad D(\sqrt{f_t}) \leq \frac{2}{t} e^{-Kt} S(f_0). \tag{13.31}
\]

**Proof.** Let \( f_t \) be the solution to the evolution equation (13.30) with a given smooth initial condition \( f_0 \). Simple calculation shows that the derivative of the entropy \( S(f_t) \) is given by
\[
\partial_t S(f_t) = \int (\mathcal{L} f_t) \log f_t \text{d}\mu + \int f_t \frac{\mathcal{L} f_t}{f_t} \text{d}\mu = - \int \frac{(\nabla f_t)^2}{f_t} \text{d}\mu = -4D(\sqrt{f_t}), \tag{13.32}
\]
where we used that \( \int \mathcal{L} f_t \text{d}\mu = 0 \) by (13.26). Similarly, we can compute the evolution of the Dirichlet form. Let \( h_t := \sqrt{f_t} \) for simplicity, then
\[
\partial_t h_t = \frac{1}{2h_t} \partial_t h_t^2 = \frac{1}{2h_t} \mathcal{L} h_t^2 = \mathcal{L} h_t + \frac{1}{h_t} (\nabla h_t)^2.
\]
We compute (dropping the \( t \) subscript for brevity)
\[
\partial_t D(\sqrt{f_t}) = \partial_t \int (\nabla h)^2 \text{d}\mu = 2 \int (\nabla h) \cdot \nabla \partial_t h \\
= 2 \int (\nabla h) \cdot (\nabla \mathcal{L} h) \text{d}\mu + 2 \int (\nabla h) \cdot \nabla (\nabla h)^2 \text{d}\mu \\
= 2 \int (\nabla h) \cdot |\nabla, \mathcal{L}| h \text{d}\mu + 2 \int (\nabla h) \cdot \mathcal{L}(\nabla h) \text{d}\mu + 2 \int \sum_{ij} \partial_i h [\frac{2(\partial_j h) \partial_i h_j h}{h} - \frac{(\partial_j h)^2 \partial_i h}{h^2}] \text{d}\mu \\
= -2 \int (\nabla h) \cdot (\nabla^2 \mathcal{H}) \nabla h \text{d}\mu - 2 \int \sum_{ij} (\partial_i \partial_j h)^2 \text{d}\mu + 2 \int \sum_{ij} [\frac{2(\partial_j h) \partial_i h_j h}{h} - \frac{(\partial_j h)^2 \partial_i h}{h^2}] \text{d}\mu \\
= -2 \int (\nabla h) \cdot (\nabla^2 \mathcal{H}) \nabla h \text{d}\mu - 2 \int \sum_{ij} \left( \partial_i h - \frac{(\partial h) (\partial_j h)}{h} \right)^2 \text{d}\mu, \tag{13.33}
\]
where we used the commutator
\[
[\nabla, \mathcal{L}] = -(\nabla^2 \mathcal{H}) \nabla.
\]
Therefore, under the convexity condition (13.28), we have
\[
\partial_t D(\sqrt{f_t}) \leq -2KD(\sqrt{f_t}). \tag{13.34}
\]
Integrating (13.34), we have
\[ D(\sqrt{f_t}) \leq e^{-2tK}D(\sqrt{f_0}). \] (13.35)
This proves that the equilibrium is achieved at \( t = \infty \) with \( f_\infty = 1 \), and both the entropy and the Dirichlet form are zero. Integrating (13.32) from \( t = 0 \) to \( t = \infty \), and using the monotonicity of \( D(\sqrt{f}) \) from (13.34), we obtain
\[ -S(f_0) = -4 \int_0^\infty D(\sqrt{f_t})dt \geq -4D(\sqrt{f_0}) \int_0^\infty e^{-2tK}dt = -\frac{2}{K}D(\sqrt{f_0}). \] (13.36)
This proves the LSI (13.29) for any smooth function \( f = f_0 \). By a standard density argument, it can be extended to any function \( f \) with finite Dirichlet form, \( D(\sqrt{f}) < \infty \). For functions with unbounded Dirichlet form we interpret the LSI (13.29) as a tautology. In particular, (13.29) holds for \( f_t \) as well, \( S(f_t) \leq (2/K)D(\sqrt{f_t}) \). Inserting this back to (13.32), we have
\[ \partial_t S(f_t) \leq -2KS(f_t). \]
Integrating this inequality from time zero, we obtain the exponential relaxation of the entropy on time scale \( t \simeq 1/K \)
\[ S(f_t) \leq e^{-2tK}S(f_0). \] (13.37)
Finally, we can integrate (13.32) from time \( t/2 \) to \( t \) to get
\[ S(f_t) - S(f_{t/2}) = -4 \int_{t/2}^t D(\sqrt{f_\tau})d\tau. \]
Using the positivity of the entropy \( S(f_t) \geq 0 \) on the left side and the monotonicity of the Dirichlet form (from (13.34)) on the right side, we get
\[ D(\sqrt{f_t}) \leq 2\frac{t}{t}S(f_{t/2}), \] (13.38)
thus, using (13.37), we obtain exponential relaxation of the Dirichlet form on time scale \( t \simeq 1/K \)
\[ D(\sqrt{f_t}) \leq 2e^{-tK}S(f_0). \]

**Standard example.** Let \( \mu \) be the centered Gaussian measure with variance \( a^2 \) on \( \mathbb{R}^n \), i.e.,
\[ d\mu(x) = \frac{1}{(2\pi a^2)^{n/2}}e^{-x^2/2a^2}dx. \] (13.39)
Written in the form \( e^{-\mathcal{H}/Z} \) with the quadratic Hamiltonian \( \mathcal{H}(x) = x^2/2a^2 \), we see that (13.28) holds with \( K = a^{-2} \). Then (13.29) with the replacement \( f \rightarrow f^2 \) yields that
\[ \int f^2 \log f^2 d\mu \leq 2a^2 \int (\nabla f)^2 d\mu \] (13.40)
for any normalized \( \int f^2 d\mu = 1 \). The constant of this inequality is optimal.

**Alternative formulation of LSI.** We remark that there is another formulation of the LSI (see [99]). To make this connection, let
\[ g(x) := f(x) \frac{1}{(2\pi a^2)^{n/4}}e^{-x^2/4a^2}, \] (13.41)
and notice that \( \int g^2(x)dx = \int f^2d\mu \), where \( dx \) is the Lebesgue measure and \( d\mu \) is of the form (13.39). Rewriting (13.40) to an inequality for \( g \) and with the replacement \( 2\pi a^2 \rightarrow a^2 \), we get the following version of the LSI
\[ \int_{\mathbb{R}^n} g^2 \log(g^2/\|g\|^2)dx + n[1 + \log a] \int_{\mathbb{R}^n} g^2 dx \leq (a^2/\pi) \int_{\mathbb{R}^n} |\nabla g|^2dx \] (13.42)
that holds for any $a > 0$ and any function $g$, where $\|g\| = (\int g^2 dx)^{1/2}$ is the $L^2$ norm with respect to the Lebesgue measure.

**Proposition 13.7** (LSI implies spectral gap). Let $\mu$ satisfy the LSI (13.27) with a LSI constant $\gamma$. Then for any $v \in L^2(\mu)$ with $\int v \, d\mu = 0$ we have

$$
\int v^2 \, d\mu \leq \frac{\gamma}{2} \int |\nabla v|^2 \, d\mu = \frac{\gamma}{2} D(v), \tag{13.43}
$$

i.e., $\mu$ has a spectral gap of size at least $\gamma/2$.

**Proof.** By definition of the LSI constant, we have

$$
\int u \log u \, d\mu \leq \gamma D(\sqrt{u})
$$

for any $u$ with $\int u \, d\mu = 1$. For any bounded smooth function $v$ with $\int v \, d\mu = 0$, define $u = 1 + \varepsilon v$. Then we have

$$
\varepsilon^{-2} \int (1 + \varepsilon v) \log(1 + \varepsilon v) \, d\mu \leq \frac{\gamma}{4} \int \frac{|\nabla v|^2}{1 + \varepsilon v} \, d\mu. \tag{13.44}
$$

Taking the limit $\varepsilon \to 0$, we get that the right hand side converges to $\frac{1}{2} \int v^2 \, d\mu$ by dominated convergence. This proves the proposition. \qed

**Proposition 13.8** (Concentration inequality (Herbst bound)). Suppose that the measure $\mu$ satisfies the LSI with a constant $\gamma$. Let $F$ be a function with $E^\mu F = 0$. Then we have

$$
E^\mu e^F \leq \exp \left( \frac{\gamma}{4} \|\nabla F\|_\infty^2 \right). \tag{13.45}
$$

where

$$
\|\nabla F\|_\infty := \sup_x \sqrt{\sum_i |\partial_i F(x)|^2}.
$$

In particular, we have

$$
P^\mu(|F| \geq \alpha) \leq \exp \left( - \frac{\alpha^2}{\gamma \|\nabla F\|_\infty^2} \right) \tag{13.46}
$$

for any $\alpha > 0$.

Notice that we get an exponential tail estimate from the LSI. If we only have the spectral gap estimate, (13.43), we can only bound the variance of $F$ which yields a quadratic tail estimate

$$
P^\mu(|F| \geq \alpha) \leq \frac{\gamma \|\nabla F\|^2_\infty}{2\alpha^2}.
$$

In our typical applications, we have $\gamma \|\nabla F\|_\infty^2 \asymp N^{-1}$. We often need to control the concentration of many ($\asymp N^C$) different functions $F$ in parallel. Thus the simple union bound is applicable with the LSI but not with the spectral gap estimate.

**Proof.** Denote by

$$
u = u(t) := \frac{\exp(e^t F)}{E^\mu \exp(e^t F)}.
$$

By differentiation and the LSI, we have

$$
\frac{d}{dt} \left[ e^{-t} \log E^\mu \exp(e^t F) \right] = e^{-t} E^\mu u \log u \leq e^{-t} \gamma E^\mu |\nabla \sqrt{u}|^2. \tag{13.47}
$$

99
Clearly,
\[ \mathbb{E}^\mu |\nabla \sqrt{u}|^2 \leq \frac{e^{2t}}{4} \| \nabla F \|_\infty^2. \]  
(13.48)

Integrating from any \( t < 0 \) to 0 yields that
\[ \left[ \log \mathbb{E}^\mu \exp (F) \right] - \left[ e^{-t} \log \mathbb{E}^\mu \exp (te^t F) \right] = \frac{\gamma}{4} \| \nabla F \|_\infty^2 \int_t^0 ds e^s. \]  
(13.49)

From the condition \( \mathbb{E}^\mu F = 0 \) we have
\[ \lim_{t \to -\infty} \left[ e^{-t} \log \mathbb{E}^\mu \exp (te^t F) \right] = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \log \mathbb{E}^\mu [1 + \varepsilon F + O(\varepsilon^2 e^F)] = 0 \]
by dominated convergence. This proves the inequality (13.45). The concentration bound (13.46) follows from an exponential Markov inequality:
\[ \mathbb{P}^\mu (F \geq \alpha) \leq \mathbb{E}^\mu e^{(F - \alpha)} \leq \exp \left( -\alpha t + t^2 \frac{\gamma}{4} \| \nabla F \|_\infty^2 \right), \]
where we chose the optimal \( t = 2\alpha / \| \nabla F \|_\infty \) in the last step. Changing \( F \to -F \) we obtain the opposite bound and thus we proved (13.46).

**Proposition 13.9 (Stability of the LSI constant).** Consider two measures \( \mu, \nu \) on \( \mathbb{R}^N \) that are related by \( \nu = g\mu \) with some bounded function \( g \). Let \( \gamma_\mu \) denote the LSI constant for \( \mu \). Then
\[ \gamma_\nu \leq \| g \|_\infty \| g^{-1} \|_\infty \gamma_\mu. \]

**Proof.** Take an arbitrary function \( f \geq 0 \) with \( \int f d\nu = 1 \). Let \( \alpha := \int f d\mu \leq \| g^{-1} \|_\infty \) and by definition of the entropy we have
\[ S_\mu (f/\alpha) = \int (f/\alpha) \log (f/\alpha) d\mu. \]

The following inequality for any two nonnegative numbers \( a, b \) can be checked by elementary calculus:
\[ a \log a - b \log b - (1 + \log b)(a - b) \geq 0. \]

Hence,
\[ \int \left[ f \log f - \alpha \log \alpha - (1 + \log \alpha)(f - \alpha) \right] d\nu \leq \| g \|_\infty \int \left[ f \log f - \alpha \log \alpha - (1 + \log \alpha)(f - \alpha) \right] d\mu = \| g \|_\infty \alpha S_\mu (f/\alpha). \]

The left hand side of the last inequality is equal to
\[ S_\nu (f) - [\log \alpha - (\alpha - 1)] \geq S_\nu (f), \]
where we have used the concavity of log. This proves that
\[ S_\nu (f) \leq \| g \|_\infty \alpha S_\nu (f/\alpha). \]

Suppose that the LSI holds for \( \mu \) with constant \( \gamma_\mu \). Then we have
\[ S_\nu (f) \leq \| g \|_\infty \gamma_\mu \alpha \int (\nabla \sqrt{f/\alpha})^2 d\mu = \| g \|_\infty \gamma_\mu \int (\nabla \sqrt{f})^2 g^{-1} d\nu \leq \| g \|_\infty \| g^{-1} \|_\infty \gamma_\mu \int (\nabla \sqrt{f})^2 d\nu, \]
and this proves the lemma. \( \square \)
Proposition 13.10 (Tensorial property of the LSI). Consider two probability measures $\mu, \nu$ on $\mathbb{R}^n, \mathbb{R}^m$ respectively. Suppose that the LSI holds for them with LSI constants $\gamma_\mu$ and $\gamma_\nu$, respectively. Let $\omega = \mu \otimes \nu$ be the product measure on $\mathbb{R}^{m+n}$. Then the LSI holds for $\omega$ with LSI constant $\gamma_\omega \leq \max\{\gamma_\mu, \gamma_\nu\}$.

Proof. We will use the notations introduced in Section 13.2 with $\Omega_1 = \mathbb{R}^n$, $\Omega_2 = \mathbb{R}^m$. Recall the additivity of entropy (13.20) w.r.t. martingale decomposition. In the current situation, $\omega = \mu \otimes \nu$ and thus $\hat{\omega} = \mu$ and $\omega_x = \nu$ for any $x$. Furthermore,

$$\hat{f}(x) = \hat{\omega}(x)^{-1} \int f(x,y)\omega(x,y)dy = \int f(x,y)\nu(y)dy. \quad (13.50)$$

By the additivity of entropy and the LSI w.r.t. $\mu$ and $\nu$, we have

$$S_\omega(f) = E^{\hat{\mu}} S_\nu(f_x) + S_\mu(\hat{f}) \leq \gamma_\nu \int \hat{f}(x)\mu(x)dx \int \frac{\left|\nabla_y \sqrt{f(x,y)}\right|^2}{f(x)} \nu(y)dy + \gamma_\mu \int \left|\nabla_x \sqrt{f(x,y)}\right| \mu(x)dx \int \nu(y)dy \mid \mu(x)dx. \quad (13.51)$$

The integral in the first term on the right hand side is equal to

$$\iint |\nabla_y \sqrt{f(x,y)}|^2 \mu(x)\nu(y)dx dy.$$

The integral of the second term on the right hand side is bounded by

$$\int \frac{1}{4} \int \frac{\left|\nabla_x f(x,y)\right|^2}{f(x,y)} \mu(x)dx \leq \int \left|\nabla_x \sqrt{f(x,y)}\right|^2 \mu(x)\nu(y)dx dy,$$

where we have written $\nabla_x f = 2(\nabla_x \sqrt{f}) \sqrt{f}$ and used the Schwarz inequality. Summarizing, we have proved that

$$S_\omega(f) \leq \gamma_\nu \int |\nabla_y \sqrt{f(x,y)}|^2 \omega(x,y)dx dy + \gamma_\mu \int |\nabla_x \sqrt{f(x,y)}|^2 \omega(x,y)dx dy,$$

and this proves the proposition.

The following lemma is a useful tool to control the entropy flow w.r.t non-equilibrium measure.

Lemma 13.11. \cite{139} Suppose we have evolution equation $\partial_t f_t = \mathcal{L} f_t$ with $\mathcal{L}$ defined via the Dirichlet form $\langle f, (-\mathcal{L}) f \rangle_{\mu} = \sum_j \langle \partial_j f \rangle^2 d\mu$. Then for any time dependent probability density $\psi_t$ w.r.t. $\mu$, we have the entropy flow identity

$$\partial_t S_\mu(f_t|\psi_t) = -4 \sum_j \int (\partial_j \sqrt{g_t})^2 \psi_t d\mu + \int g_t (\mathcal{L} - \partial_t) \psi_t d\mu,$$

where $g_t := f_t/\psi_t$ and

$$S_\mu(f_t|\psi_t) := \int f_t \log \frac{f_t}{\psi_t} d\mu = S(f_t \mu|\psi_t \mu)$$

is the relative entropy of $f_t \mu$ w.r.t. $\psi_t \mu$.

Proof. A simple computation then yields that

$$\frac{d}{dt} S_\mu(f_t|\psi_t) = \int (\mathcal{L} f_t)(\log g_t) d\mu + \int f_t \frac{\mathcal{L} f_t}{f_t} d\mu - \int (\partial_t \psi_t) g_t d\mu$$

$$= \int (\mathcal{L} f_t)(\log g_t) d\mu - \int (\partial_t \psi_t) g_t d\mu$$

$$= \int g_t \partial_t \psi_t \mathcal{L} (\log g_t) d\mu - \int g_t \partial_t \psi_t d\mu$$

$$= \int \psi_t \left[ g_t \mathcal{L} (\log g_t) - g_t \frac{\mathcal{L} g_t}{g_t} \right] d\mu + \int g_t (\mathcal{L} - \partial_t) \psi_t d\mu.$$
By definition of $L$, we have
\[
L(\log g) - \frac{Lg}{g} = -\sum_j \left( \frac{\partial_j g}{g} \right)^2 = -4 \sum_j (\partial_j \sqrt{g})^2,
\]
(13.52)
and we have proved Lemma 13.11.

We remark that stochastic processes and their generators can be defined in more general setups, e.g. the underlying probability spaces may be different from $\mathbb{R}^N$ and the generators may involve discrete jumps. In this case, the stochastic generator $L$ is defined directly, without an a priori Dirichlet form. It can, however, be proved that $-g[L(\log g) - (Lg)/g] \geq 0$ which can then be viewed as a generalization of the “Dirichlet form” associated with a generator.

### 13.4 Hypercontractivity

We now present an interesting connection between the LSI of a probability measure $\mu$ and the hypercontractivity properties of the semigroup generated by $L = L_\mu$. Since this result will not be used later in this book, this section can be skipped.

To state the result, we define the semigroup $\{P_t\}_{t \geq 0}$ by $P_t f := f_t$, where $f_t$ solves the equation $\partial_t f_t = L f_t$ with initial condition $f_0 = f$.

**Theorem 13.12** (L. Gross [78]). For a measure $\mu$ on $\mathbb{R}^n$ and for any fixed constants $\beta \geq 0$ and $\gamma > 0$ the following two statements are equivalent:

(i) The generalized LSI

\[
\int f \log f \, d\mu \leq \gamma \int |\nabla \sqrt{f}|^2 \, d\mu + \beta, \quad \text{for any } f \geq 0 \text{ with } \int f \, d\mu = 1,
\]
(13.53)

holds;

(ii) The hypercontractivity estimate

\[
\|P_t f\|_{L^q(\mu)} \leq \exp \left\{ \beta \left[ \frac{1}{p} - \frac{1}{q} \right] \right\} \|f\|_{L^p(\mu)}
\]
(13.54)
holds for all exponents satisfying

\[
\frac{q - 1}{p - 1} \leq e^{4t/\gamma}, \quad 1 < p \leq q < \infty.
\]

**Proof.** We will only prove (i)$\Rightarrow$(ii), i.e., that the generalized LSI implies the decay estimate, the proof of the converse statement can be found in [43]. First we assume that $f \geq 0$, hence $f_t \geq 0$. Direct differentiation yields the following identity

\[
\frac{d}{dt} \log \|f_t\|_{p(t)} = \frac{\dot{p}(t)}{p(t)^2} \left[ -\frac{4(p(t) - 1)}{p(t)} D(u(t)) + \int u(t)^2 \log(u(t)^2) \, d\mu \right],
\]
(13.55)
with $\dot{p}(t) = \frac{d}{dt} p(t)$ and where we defined

\[
u(t) := f_t^{p(t)/2} \|f_t\|_{p(t)}^{-p(t)/2}, \quad D(u) = \int (\nabla u)^2 \, d\mu.
\]

Now choose $p(t)$ to solve the differential equation

\[
g = \frac{4(p(t) - 1)}{\dot{p}(t)}, \quad \text{with } p(0) = p, \quad \text{i.e., } \quad p(t) = 1 + (p - 1) e^{4t/\gamma},
\]

102
where $\gamma$ is the constant given in the theorem. Using (13.53) for the $L^1(\mu)$-normalized function $u(t)^2$, we have
\[
\frac{d}{dt} \log \| f_t \|_{\mu(t)} \leq \frac{\beta \dot{p}(t)}{p(t)^2}.
\]
Integrating both sides, from $t = 0$ to some $T$ we have
\[
\log \| f_T \|_{\mu(T)} - \log \| f \|_{\mu} \leq \beta \left[ \frac{1}{p} - \frac{1}{p(T)} \right].
\]
Choosing $T$ such that $p(T) = q$, we have proved (13.54) for $f \geq 0$. The general case follows from separating the positive and negative parts of $f$.

Exercise. In this exercise, we show that the idea of the LSI can be useful even if the invariant measure is not a probability measure. The sketch below follows the paper by Carlen-Loss [33] and it works for any parabolic equation of the type
\[
\partial_t f_t = (\nabla \cdot (D(x,t) \nabla) + b(x,t) \cdot \nabla) f_t,
\]
for any divergence free $b$ and $D(x,t) \geq c > 0$. For simplicity of notation, we consider only the heat equation on $\mathbb{R}^n$
\[
\partial_t f_t = \Delta f_t. \tag{13.56}
\]
The invariant measure of this flow is the standard Lebesgue measure on $\mathbb{R}^n$.

(i) Check the formula (13.55) in this setup, i.e., show that the following identity holds:
\[
\frac{d}{dt} \log \| f_t \|_{\mu(t)} = \frac{1}{p(t)} \| f_t \|_{\mu(t)} - \frac{\dot{p}(t)}{p(t)^2} \int f_t^{p(t)} \log f_t^p dx - \frac{\dot{p}(t)}{p(t)^2} \int f_t^{p(t)} \log f_t^p dx \int (\nabla u(t))^2 dx + \int u(t)^2 \log(u(t)^2) dx, \tag{13.57}
\]
where all norms are w.r.t. the Lebesgue measure and
\[
u(t) = f_t^{p/2} \| f_t \|_{\mu}^{-p/2}, \quad p = p(t).
\]
(ii) For any $t$ fixed, choose the constant in (13.42) by setting
\[
\alpha^2 = \frac{4(p(t) - 1)}{\dot{p}(t)}.
\]
Thus we have
\[
\frac{d}{dt} \log \| f_t \|_{\mu(t)} \leq -\frac{n \dot{p}(t)}{p(t)^2} \left[ 1 + \frac{1}{2} \log \left( \frac{4\pi(p(t) - 1)}{\dot{p}(t)} \right) \right].
\]
For a suitable choice of the function $p(t)$ with $p(0) = 1$, $p(T) = \infty$, show that
\[
\| f_T \|_{\infty} \leq (4\pi T)^{-n/2} \| f_0 \|_1. \tag{13.58}
\]
We remark that in the case of $D \equiv 1$ and $b \equiv 0$, this heat kernel estimate is also a trivial consequence of the explicit formula for the heat kernel $e^{t\Delta}(x,y)$. The point is, as remarked earlier, that this proof works for a general class of second order parabolic equations.
13.5 Brascamp-Lieb inequality

The following inequality will not be needed for the main result of this book, so the reader may skip it. Nevertheless, we included it here, since it is an important tool in the analysis of probability measures in very high dimensions and it was used in universality proofs for the invariant ensembles [25]. We will formulate it on \( \mathbb{R}^N \) and in Section 13.7 we comment on how to extend it to certain probability measures on the simplex \( \Sigma_N \) defined in (12.3).

**Theorem 13.13** (Brascamp-Lieb inequality [29]). Consider a probability measure \( \mu \) on \( \mathbb{R}^N \) of the form (13.24), i.e., \( \mu = e^{-\mathcal{H}}/Z \). Suppose that the Hamiltonian is strictly convex, i.e.,

\[
\nabla^2 \mathcal{H}(x) \geq K > 0
\]

(13.59)
as a matrix inequality for some positive constant \( K \). Then for any smooth function \( f \in L^2(\mu) \) function we have

\[
(f; f)_\mu \leq \langle \nabla f, \left[ \nabla^2 \mathcal{H} \right]^{-1} \nabla f \rangle_\mu.
\]

(13.60)

Recall that \( \langle f, g \rangle_\mu = \int f g d\mu \) denotes the scalar product and \( \langle f; g \rangle_\mu = \langle f, g \rangle_\mu - \langle 1, f \rangle_\mu \langle 1, g \rangle_\mu \) is the covariance. With a slight abuse of notation, we also use the notation \( \langle F, G \rangle_\mu = \sum_{i=1}^N \langle F_i, G_i \rangle_\mu \) for the scalar product of any two vector-valued functions \( F, G : \mathbb{R}^N \to \mathbb{R}^N \); this extended scalar product is used in the right hand side of (13.60).

**Remark:** Notice that the Brascamp-Lieb inequality is optimal in the sense that, if \( \mu \) is a Gaussian measure, then we can find \( f \) so that the inequality becomes equality. Moreover, if we use the matrix inequality \( \nabla^2 \mathcal{H} \geq K \), then (13.60) is reduced to the spectral gap inequality (13.43), but of course (13.60) is stronger. The following proof is due to Helffer-Sjöstrand [80] and Naddaf-Spencer [108].

**Proof.** Recall that \( \mathcal{L} \) is the generator for a reversible dynamics with reversible measure \( \mu \) defined in (13.26). For the dynamics (13.30), we have

\[
\frac{d}{dt} \langle f, e^{t\mathcal{L}} g \rangle = \langle f, \mathcal{L} e^{t\mathcal{L}} g \rangle = - \sum_j \langle \partial_{x_j} f, \partial_{x_j} e^{t\mathcal{L}} g \rangle.
\]

Define

\[
G(t, x) := (G_1(t, x), \ldots, G_N(t, x)), \quad G_j(t, x) := \partial_{x_j} \left[ e^{t\mathcal{L}} g(x) \right].
\]

Recall the explicit form of \( \mathcal{L} \):

\[
\mathcal{L} = \Delta - (\nabla \mathcal{H}) \cdot \nabla = \Delta + \sum_k b_k \partial_{x_k}, \quad b_k := -\partial_{x_k} \mathcal{H}.
\]

Define a new generator \( \mathcal{L} \) on any vector-valued function \( G : \mathbb{R}^N \to \mathbb{R}^N \) by

\[
(\mathcal{L} G)_j(x) := \mathcal{L} G_j(x) + \sum_k (\partial_{x_j} b_k) G_k(x).
\]

Clearly, by explicit computation, we have

\[
\partial_t G(t, x) = \mathcal{L} G(t, x).
\]

From the decay estimate (13.31), the dynamics are mixing, i.e., \( \lim_{t \to \infty} \langle f, e^{t\mathcal{L}} g \rangle = 0 \) for any smooth function \( f \) with \( \int f d\mu = 0 \). Thus for any such function \( f \) we have

\[
\langle f, g \rangle_\mu = - \int_0^\infty \frac{d}{dt} \langle f, e^{t\mathcal{L}} g \rangle_\mu dt = \sum_j \int_0^\infty \langle \partial_{x_j} f, G_j(t, x) \rangle_\mu dt = \int_0^\infty \langle \nabla f, e^{t\mathcal{L}} \nabla g \rangle_\mu dt
\]

\[
= \langle \nabla f, (-\mathcal{L})^{-1} \nabla g \rangle_\mu.
\]

104
Taking $g = f$ and from the definition of $L$, we have

$$-(L)_{i,j} = - \mathcal{L}_{i,j} - (\partial_{x_i} b_j) \geq (\partial_{x_i} \partial_{x_j} \mathcal{H})$$

as an operator inequality and we have proved that

$$\langle f, f \rangle_{\mu} = \langle \nabla f, (-L)^{-1} \nabla f \rangle_{\mu} \leq \langle \nabla f, [\nabla^2 \mathcal{H}]^{-1} \nabla f \rangle_{\mu}.$$  

\[\square\]

### 13.6 Remarks on the applications of the LSI to random matrices

The Herbst bound provides strong concentration results for probability measures that satisfy the LSI. In this section we exploit this connection for random matrices. For definiteness, we will consider the Gaussian orthogonal ensemble (GOE), although the arguments below can be extended to more general Wigner as well as invariant ensembles. There are two ways to view Gaussian matrix ensembles in the context of the LSI: we can either work on the probability space of the matrix $H$ with Gaussian matrix elements, or we can work directly on the space of the eigenvalues $\lambda$ by using the explicit formula (4.12) for invariant ensembles. Both measures satisfy the LSI (in the next Section 13.7 we will comment on how to generalize the LSI from $\mathbb{R}^N$ to the simplex $\Sigma_N$, (12.3), a version of the LSI that we will actually use). We will explore both possibilities, and we start with the matrix point of view.

#### 13.6.1 LSI from the Wigner matrix point of view

Let $\mu = \mu_G$ be the standard Gaussian measure $\mu \sim \exp(-N \text{Tr} H^2)$ on symmetric matrices. Notice that for this measure the family \( \{x_{ij} = N^{1/2} h_{ij}, i \leq j\} \) is a collection of independent standard Gaussian random variables (up to a factor 2). Hence the LSI holds for every matrix element and by its tensorial property, i.e., Proposition 13.10, the LSI holds for any function of the full matrix considered as a function of this collection \( \{x_{ij} = N^{1/2} h_{ij}, i \leq j\} \). In particular, using the spectral gap estimate (13.43) for the Gaussian variables $x_{ij}$, we have for any function $F = F(H)$,

$$\langle F(H); F(H) \rangle_{\mu} \leq C \sum_{i \leq j} \int |\partial_{x_{ij}} F(H)|^2 d\mu = \frac{C}{N} \sum_{i \leq j} \int |\partial_{h_{ij}} F(H)|^2 d\mu,$$

(13.61)

where the additional $N^{-1}$ factor comes from the scaling $h_{ij} = N^{-1/2} x_{ij}$.

Suppose that $F(H) = R(\lambda(H))$ is a real function of the eigenvalues $\lambda = (\lambda_1, \ldots, \lambda_N)$, then by the chain rule we have

$$\frac{C}{N} \sum_{i \leq j} \int |\partial_{h_{ij}} R(\lambda(H))|^2 d\mu = \frac{C}{N} \sum_{i \leq j} \int \left| \sum_{\alpha} \partial_{\lambda_{\alpha}} R(\lambda) \frac{\partial \lambda_{\alpha}}{\partial h_{ij}} \right|^2 d\mu.$$

Expanding the square and using the perturbation formula (12.9) with real eigenvectors, we can compute

$$\frac{1}{N} \sum_{i \leq j} \int \left| \sum_{\alpha} \partial_{\lambda_{\alpha}} R(\lambda) \frac{\partial \lambda_{\alpha}}{\partial h_{ij}} \right|^2 d\mu \leq \frac{1}{N} \sum_{i \leq j} \frac{2}{2 - \delta_{ij}} \int \left| \sum_{\alpha} \partial_{\lambda_{\alpha}} R(\lambda) \frac{\partial \lambda_{\alpha}}{\partial h_{ij}} \right|^2 d\mu$$

$$= \frac{1}{N} \sum_{i \leq j} \frac{2}{2 - \delta_{ij}} \int \partial_{\lambda_{\alpha}} R(\lambda) \frac{\partial \lambda_{\alpha}}{\partial h_{ij}} \frac{\partial \lambda_{\beta}}{\partial h_{ij}} d\mu$$

$$= \frac{2}{N} \sum_{i \leq j} [2 - \delta_{ij}] \sum_{\alpha, \beta} \int \partial_{\lambda_{\alpha}} R(\lambda) \partial_{\lambda_{\beta}} R(\lambda) u_{\alpha}(i) u_{\alpha}(j) u_{\beta}(i) u_{\beta}(j) d\mu$$

$$= \frac{2}{N} \sum_{i \leq j} \sum_{\alpha, \beta} \int \partial_{\lambda_{\alpha}} R(\lambda) \partial_{\lambda_{\beta}} R(\lambda) u_{\alpha}(i) u_{\alpha}(j) u_{\beta}(i) u_{\beta}(j) d\mu$$

$$= \frac{2}{N} \sum_{\alpha} \int |\partial_{\lambda_{\alpha}} R(\lambda)|^2 d\mu,$$

(13.62)
where we have used the orthogonality property and the normalization convention of the eigenvectors in the last step.

We remark that the annoying factor $[2 - \delta_{ij}]$ can be avoided if we first consider $\lambda_\alpha$ as function of all \{${x_{ij}} : 1 \leq i, j \leq N$\} as independent variables. Then the perturbation formula (12.9) becomes

$$\frac{\partial \lambda_\alpha}{\partial h_{ij}} |_{h_{ij} = h_{ij}} = u_\alpha(i) u_\alpha(j),$$

(13.63)
i.e., the derivative evaluated on the submanifold of Hermitian matrices. In this way we can keep the summation in (13.61) unrestricted and up to a constant factor, we will get the same final result as in (13.62).

In summary, we proved

$$\langle F; F \rangle_\mu = \langle R(\lambda(H)); R(\lambda(H)) \rangle_\mu \leq \frac{C}{N} \sum_\alpha \int |\partial_\lambda R(\lambda)|^2 d\mu. \quad (13.64)$$

Notice that this argument holds for any generalized Wigner matrix as long as a spectral gap estimate (13.43) holds for the distribution of every rescaled matrix element $N^{1/2} h_{ij}$. Furthermore, it can be generalized for Wigner matrices with Bernoulli random variables for which there is a spectral gap (and LSI) in discrete form. Similar remarks apply to the following LSI estimates.

To see how good this estimate is, consider a local linear statistics of eigenvalues, i.e., the local average of $K$ consecutive eigenvalues with label near $M$, i.e., set

$$R(\lambda) := \frac{1}{K} \sum_\alpha A\left(\frac{\alpha - M}{K}\right) \lambda_\alpha,$$

where $A$ a smooth function of compact support and $1 \leq M, K \leq N$. Computing the right hand side of (13.64) and combining it with (13.61), we get

$$\langle F; F \rangle_\mu \leq \frac{C}{NK^2} \sum_\alpha A^2\left(\frac{\alpha - M}{K}\right) \leq \frac{C_A}{NK}, \quad (13.65)$$

where the last constant $C_A$ depends on the function $A$. For $K = 1$, this inequality estimates the square of the fluctuation of a single eigenvalue (choosing $A$ appropriately). The bound (13.65) is off by a factor $N$ since the true fluctuation is of order almost $1/N$ by rigidity, Theorem 11.5, at least in the bulk, i.e., if $\delta N \leq M \leq (1 - \delta)N$. On the other hand, for $K = N$ the bound (13.65) is much more precise; it shows that the variance of a macroscopic average of the eigenvalues is at most of order $N^{-2}$. This is the correct order of magnitude, in fact, it is known that $\sum_\alpha A\left(\frac{\alpha}{N}\right) \lambda_\alpha$ converges to a Gaussian random variable, see, e.g., [102, 117] and references therein. Hence the spectral gap argument yields the optimal (up to constant) result for any macroscopic average of eigenvalues.

Another common quantity of interest is the Stieltjes transform of the empirical eigenvalue distribution, i.e.,

$$F(H) = G(\lambda(H)) \quad \text{with} \quad G(\lambda) = m_N(z) = \frac{1}{N} \sum_\alpha \frac{1}{\lambda_\alpha - z}, \quad z = E + i\eta. \quad (13.66)$$

In this case, using (13.64), we can estimate

$$\langle m_N(z); m_N(z) \rangle_\mu \leq \frac{C}{N} \sum_\alpha \int |\partial_\lambda G(\lambda)|^2 d\mu = \frac{C}{N^3} \sum_\alpha \int \frac{1}{|\lambda_\alpha - z|^4} d\mu \leq \frac{C}{N^2 \eta^4}, \quad (13.67)$$

where we used $|\lambda_\alpha - z| \geq \text{Im} z = \eta$.

This shows that the variance of $m_N(z)$ vanishes if $\eta \gg N^{-1/2}$. The last estimate can be improved if we know that the density of the empirical measure is bounded. Roughly speaking, if we know that in the summation over $\alpha$, the number of eigenvalues in an interval of size $\eta$ near $E$ is at most of order $N\eta$, then the last estimate can be improved to

$$\frac{1}{N^3} \sum_\alpha \int \frac{1}{|\lambda_\alpha - z|^4} d\mu \leq \frac{1}{N^3} \sum_{{|\lambda_\alpha - E|} \leq \eta} \int \frac{1}{|\lambda_\alpha - z|^4} d\mu \leq \frac{C}{N^3} (N\eta)^{1/4} = \frac{C}{N^{2/3} \eta^2}. \quad (13.68)$$
This shows that the variance of \( m_N(z) \) vanishes if \( \eta \gg N^{-2/3} \). Since vanishing fluctuations can be used to estimate the density, this argument can actually be made rigorous by a bootstrapping argument [61]. Notice that the scale \( \eta \gg N^{-2/3} \) is still far from the resolution demonstrated in the local semicircle law, Theorem 6.7.

Whenever the LSI is available, the variance bounds can be easily lifted to a concentration estimate with exponential tail. We now demonstrate this mechanism for the fluctuation of a single eigenvalue. In other words, we will apply (13.46) with \( F(H) = \lambda_\alpha(H) - \mathbb{E}^\mu \lambda_\alpha(H) \) for a fixed \( \alpha \). From (12.9), we have

\[
\sum_{i \leq j} |\nabla_{x_{ij}} F|^2 \leq \frac{2}{N},
\]

thus (13.46) implies

\[
\mathbb{P}^\mu(|\lambda_\alpha - \mathbb{E}^\mu \lambda_\alpha| \geq t) \leq \exp \left( -cNt^2 \right)
\]

for any \( \alpha \) fixed. This inequality has two deficiencies compared with the rigidity bounds in Theorem 11.5: First, the control of \( |\lambda_\alpha - \mathbb{E}^\mu \lambda_\alpha| \) is only up to \( N^{-1/2} \) which is still far from the optimal \( N^{-1} \) (in the bulk). Second, it does not give information on the difference between the expectation \( \mathbb{E}^\mu \lambda_\alpha \) and the corresponding classical location \( \gamma_\alpha \) defined as the \( \alpha \)-th \( N \)-quantile of the limiting density, see (11.31).

### 13.6.2 LSI from invariant ensemble point of view

Now we pass to the second point of view, where the basic measure \( \mu_G \) is the invariant ensemble on the eigenvalues. One might hope that the situation can be improved since we look directly at the Gaussian eigenvalue ensemble defined in (12.13) with the Gaussian choice for \( V(\lambda) = \frac{1}{2} \lambda^2 \). Notice that the role of \( \mathcal{H} \) in Theorem 13.6 will be played by \( N \mathcal{H}_N \) defined in (12.13) (with \( \beta = 1 \)). The Hessian of \( \mathcal{H}_N \) is given by (all inner products and norms in the following equation are w.r.t the standard inner product in \( \mathbb{R}^N \))

\[
(\mathbf{v}, \nabla^2 \mathcal{H}_N(\mathbf{x}) \mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|^2 + \frac{1}{N} \sum_{i < j} \frac{(v_i - v_j)^2}{(x_i - x_j)^2} \geq \frac{1}{2} \|\mathbf{v}\|^2, \quad \mathbf{v} \in \mathbb{R}^N,
\]

thus the convexity bound (13.28) holds with a constant \( K = N/2 \) for \( N \mathcal{H}_N \). Hence the spectral gap from (13.43) implies that for any function \( R(\lambda) \) we have

\[
\langle R(\lambda); R(\lambda) \rangle_{\mu_G} \leq \frac{C}{N} \sum_\alpha \int |\partial_{\lambda_\alpha} R(\lambda)|^2 d\mu_G.
\]

Notice that this bound is in the same form as in (13.64). A similar statement holds for the LSI, i.e., we have

\[
\int R \log R d\mu_G \leq \frac{C}{N} \sum_\alpha \int |\partial_{\lambda_\alpha} \sqrt{R(\lambda)}|^2 d\mu_G.
\]

We can apply the concentration inequality (13.46) to the function \( R(\lambda) = \lambda_\alpha \) to get (13.69). Once again, we get exactly the same result as considered earlier, so there is no advantage of using the explicit invariant measure \( \mu_G \).

We have seen that in both the matrix and the invariant ensemble setup, even for Gaussian ensembles, concentration estimates based on spectral gap or LSI do not yield optimal results for the eigenvalues on short scales. In the matrix setup, the proof of the optimal rigidity bounds for Wigner matrices, Theorem 11.5, uses a completely different approach independent of the LSI. In the invariant ensemble setup, while the convexity bound (13.70) cannot be improved for a general vector \( \mathbf{v} \), it becomes much stronger if we consider it only for vectors \( \mathbf{v} \) satisfying \( \sum_j v_j = 0 \) due to the structure of the Hessian (13.70). In the next section, we will explore this idea to improve estimates on certain functions of the eigenvalues.
13.7 Extensions to the simplex, regularization of the DBM

In the above application to either spectral gap or LSI, the measure \( \mu_G \) was restricted to the subset \( \Sigma = \Sigma_N = \{ x \in \mathbb{R}^N : x_1 < x_2 < \ldots < x_N \} \). It is absolutely continuous with respect to the Lebesgue measure on \( \Sigma_N \), but if we write it in the form (13.24) then the Hamiltonian \( \mathcal{H} \) has to be infinite outside of \( \Sigma_N \); in particular it is not differentiable, a property that is implicitly assumed in Section 13.3. This issue is closely related to the boundary terms in the integration by parts, especially in (13.33), that arise if one formally extends the proofs to \( \Sigma \). Therefore, the results of Theorem 13.6 do not apply directly. The proper procedure involves an approximation and extension of the measure from \( \Sigma_N \) to \( \mathbb{R}^N \), using the results of Section 13.3 for the regularized measure on \( \mathbb{R}^N \) and then removing the regularization. Whether the regularization can be removed for any \( \beta > 0 \) or only for \( \beta \geq 1 \) depends on the statement, as we explain now.

For simplicity, we work in the setup of the Gaussian \( \beta \)-ensemble, i.e., we set

\[
\mathcal{H}(x) = \sum_{i=1}^N \frac{1}{4} x_i^2 - \frac{1}{N} \sum_{i<j} \log(x_j - x_i),
\]

and define \( \mu_\beta(x) = Z_\mu^{-1} e^{-N\beta \mathcal{H}(x)} \) on the simplex \( \Sigma_N \), exactly as in (12.13) with the Gaussian potential \( V(x) = \frac{1}{2} x^2 \) and with any parameter \( \beta > 0 \). As usual, \( Z_\mu \) is a normalization constant.

Recall from (12.4) that the DBM on the simplex \( \Sigma_N \) is defined via the stochastic differential equation

\[
dx_i = \frac{\sqrt{2}}{\sqrt{\beta N}} dB_i + \left( -\frac{1}{2} x_i + \frac{1}{N} \sum_{j \neq i} \frac{1}{x_i - x_j} \right) dt \quad \text{for } i = 1, \ldots, N, \tag{13.73}
\]

where \( B_1, \ldots, B_N \) is a family of independent standard Brownian motions. The unique strong solution exists only if \( \beta \geq 1 \) (Theorem 12.2) with equilibrium measure \( \mu_\beta \). Since DBM does not exists on \( \Sigma_N \) unless \( \beta \geq 1 \), we cannot use DBM either in the SDE form (12.4) or in the PDE form (12.17) when \( \beta < 1 \), see a remark below (12.17). On the other hand, certain results may be extended to any \( \beta > 0 \) if their final formulations do not involve DBM.

In this section, we present a regularization procedure to show that substantial parts of the main results of Theorem 13.6, i.e., the LSI for any \( \beta > 0 \) and exponential relaxation decay of the entropy for \( \beta \geq 1 \), remain valid on the simplex \( \Sigma_N \). A similar generalization holds for the Brascamp-Lieb inequality. In the next section, the same regularization will be used to show that the key Dirichlet form inequality (Theorem 14.3) also hold for \( \beta > 0 \).

For later applications, we work with a slightly bigger class of measures than just \( \mu_\beta \). We consider measures on \( \Sigma = \Sigma_N \) of the form

\[
\omega = Z_\omega^{-1} e^{-\beta N \tilde{\mathcal{H}}} \mu_\beta,
\]

where \( \tilde{\mathcal{H}}(x) = \sum_j U_j(x_j) \) for some convex real valued functions \( U_j \) on \( \mathbb{R} \). The total Hamiltonian of \( \omega \) is \( \mathcal{H}_\omega := \mathcal{H} + \tilde{\mathcal{H}} \). Note that \( U_j \) are defined and convex on the entire \( \mathbb{R}^N \). The entropy and the Dirichlet form are defined as before:

\[
S_\omega(f) = \int_\Sigma f \log f \, d\omega, \quad D_\omega(f) = \frac{1}{\beta N} \int_\Sigma |\nabla f|^2 \, d\omega.
\]

The corresponding DBM is given by

\[
dx_i = \frac{\sqrt{2}}{\sqrt{\beta N}} dB_i + \left( -\frac{1}{2} x_i - U'_j(x_i) + \frac{1}{N} \sum_{j \neq i} \frac{1}{x_i - x_j} \right) dt. \tag{13.74}
\]

We assume a lower bound on the Hessian of the total Hamiltonian,

\[
\mathcal{H}'' = \mathcal{H}'' + \tilde{\mathcal{H}}'' \geq \frac{1}{2} \min_{j} \min_{x \in \mathbb{R}} U''_j(x) \geq K \tag{13.75}
\]

for some positive constant \( K \) on the entire \( \mathbb{R}^N \). This bound (13.75) plays the role of (13.28). Let \( D_\omega, S_\omega \) and \( L_\omega \) denote the Dirichlet form, entropy and generator corresponding to the measure \( \omega \). Now we claim that Theorem 13.6 holds for the measure \( \omega \) on \( \Sigma_N \) in the following form:
Theorem 13.14. Assume (13.75). Then for \( \beta > 0 \), the LSI holds, i.e.,

\[
S_\omega(f) \leq \frac{2}{R} D_\omega(\sqrt{f})
\]

(13.76)

for any nonnegative normalized density \( f \) on \( \Sigma_N \), \( \int f d\omega = 1 \), that satisfies \( f \in L^\infty \) and \( \nabla \sqrt{f} \in L^\infty \). For \( \beta \geq 1 \) the requirement that \( f \) and \( \nabla \sqrt{f} \) are bounded can be removed.

Moreover, the Brascamp-Lieb inequality also holds for any \( \beta > 0 \), i.e., for any bounded function \( f \in L^2(\Sigma_N, d\omega) \) we have

\[
(f; f)_\omega \leq (\nabla f, [H'_\omega]^{-1} \nabla f)_\omega.
\]

(13.77)

For \( \beta \geq 1 \) the requirement that \( f \) is bounded can be removed.

Furthermore, for any \( \beta \geq 1 \) the dynamics \( \partial_t f_t = \mathcal{L}_\omega f_t \) with initial condition \( f_0 = f \) is well defined on \( \Sigma_N \) and it approaches to equilibrium in the sense that

\[
S_\omega(f_t) \leq e^{-2tK} S_\omega(f).
\]

(13.78)

The inequalities above are understood in the usual sense that they are relevant only when the right hand side is finite. Moreover, by a standard density argument they extend to the closure of the corresponding spaces. For example, (13.76) holds for any \( f \) that can be approximated by a sequence of bounded normalized densities \( f_n \in L^\infty \) with \( \nabla \sqrt{f_n} \in L^\infty \) such that \( D_\omega(\sqrt{f_n} - \sqrt{f}) \to 0 \).

Before starting the formal proof, we explain the key idea. For the proofs of (13.76) and (13.77), we extend the measure \( \omega \) from \( \Sigma \) to the entire \( \mathbb{R}^N \) in a continuous way by relaxing the strict ordering \( x_i < x_{i+1} \) imposed by \( \Sigma \) but heavily penalizing the opposite order. In this way we can use Theorem 13.6 for the regularized measure and we avoid the problematic boundary terms in the integration by parts in (13.33). At the end we remove the regularization using the additional boundedness assumptions on \( f \) and \( \nabla \sqrt{f} \). For \( \beta \geq 1 \) these additional assumptions are not necessary using that \( C_\infty(\Sigma) \) functions are dense in \( H^1(\Sigma, d\omega) \). The entropy decay (13.78) will follow from the LSI and the time-integral version of the entropy dissipation (13.32) that can be proven directly on \( \Sigma \) if \( \beta \geq 1 \).

We remark that there is an alternative regularization method that mimics the proof of Theorem 13.6 directly on \( \Sigma \) for \( \beta \geq 1 \). This is based on introducing carefully selected cutoff functions in order to approximate \( f_t \) by a function compactly supported on \( \Sigma \). The compact support renders the boundary terms in the integration by parts zero, but the cutoff does not commute with the dynamics; the error has to be tracked carefully. The advantage of this alternative method is that it also gives the exponential decay of the Dirichlet form and it is also the closest in spirit to the strategy of the proof of Theorem 13.6 on \( \mathbb{R}^N \). The disadvantage is that it works only for \( \beta \geq 1 \); in particular it does not yield the LSI for \( \beta \in (0, 1) \). We will not discuss this approach in this book, the interested reader may find details in Appendix B of [64].

Proof. For any \( \delta > 0 \) define the extension \( \mu_\beta^\delta = Z_\beta^{-1} e^{-\beta N \mathcal{H}_\delta} \) of the measure \( \mu_\beta \) from \( \Sigma_N \) to \( \mathbb{R}^N \) by replacing the singular logarithm with a \( C^2 \)-function. To that end, we introduce the approximation parameter \( \delta > 0 \) and define for \( x \in \mathbb{R}^N \),

\[
\mathcal{H}_\delta(x) := \sum_i \frac{1}{4} x_i^2 - \frac{1}{N} \sum_{i<j} \log_\delta(x_j - x_i),
\]

(13.79)

where we set

\[
\log_\delta(x) := 1(x \geq \delta) \log x + 1(x < \delta) \left( \log \delta + \frac{x - \delta}{\delta} - \frac{1}{2\delta^2}(x - \delta)^2 \right), \quad x \in \mathbb{R}.
\]

It is easy to check that \( \log_\delta \in C^2(\mathbb{R}) \), is concave, and satisfies

\[
\lim_{\delta \to 0} \log_\delta(x) = \begin{cases} 
\log x & \text{if } x > 0 \\
-\infty & \text{if } x \leq 0.
\end{cases}
\]
The convergence is monotone decreasing, i.e.,
\[ \log_\delta(x) \geq \log_{\delta'}(x), \quad x \in \mathbb{R}, \]  
for any \( \delta \geq \delta' \). Furthermore, we have
\[ \partial_x \log_\delta(x) = \begin{cases} \frac{1}{x} & \text{if } x > \delta \\ \frac{2\delta - x}{\delta^2} & \text{if } x \leq \delta, \end{cases} \]
and the lower bound
\[ \partial_x^2 \log_\delta(x) \geq \begin{cases} -\frac{1}{x^2} & \text{if } x > \delta \\ -\frac{1}{\delta^2} & \text{if } x \leq \delta, \end{cases} \]
in particular \( H_\delta \) is convex with \( H_\delta'' \geq \frac{1}{2} \). Similarly, we can extend the measure \( \omega \) to \( \mathbb{R}^N \) by setting
\[ \omega_\delta := Z_{\omega,\delta}^{-1} e^{-\beta N H_\delta} \mu_\beta, \]
and its Hamiltonian still satisfies (13.75). By the monotonicity (13.80), we know that \( e^{-\beta N H_\delta(\mathbf{x})} \) is pointwise monotonically decreasing in \( \delta \) and clearly \( Z_{\mu,\delta} \) and \( Z_{\omega,\delta} \) converge to 1 as \( \delta \to 0 \). Clearly, \( \omega_\delta \to \omega \) as \( \delta \to 0 \) in the sense that for any set \( A \subset \mathbb{R}^N \) we have \( \omega_\delta(A) \to \omega(A \cap \Sigma_N) \). We remark that the regularized DBM corresponding to the measure \( \omega_\delta \) is given by
\[ dx_i = \frac{\sqrt{2}}{\sqrt{\beta N}} dB_i + \left( -\frac{1}{2} x_i - U_i'(x_i) + \frac{1}{N} \sum_{j \neq i} \log_\delta(x_i - x_j) - \frac{1}{N} \sum_{j > i} \log_\delta(x_j - x_i) \right) dt. \]

Given a function \( f \) on \( \Sigma_N \) such that \( \int_{\Sigma_N} f d\omega = 1 \) and \( D_\omega(\sqrt{f}) = \int_{\Sigma_N} |\nabla \sqrt{f}|^2 d\omega < \infty \), we extend it by symmetry to \( \mathbb{R}^N \). To do so, we first define the symmetrized version of \( \Sigma_N \), i.e.,
\[ \tilde{\Sigma}_N := \mathbb{R}^N \setminus \{ \mathbf{x} : \exists i \neq j, \ x_i = x_j \}, \]
which has the disjoint union structure
\[ \Sigma_N = \bigcup_{\pi \in \pi_N} \pi(\Sigma_N), \]
where \( \pi \) runs through all \( N \)-element permutations and it acts by permuting the coordinates of any point \( \mathbf{x} \in \mathbb{R}^N \). For any \( \mathbf{x} \in \tilde{\Sigma}_N \) there is a unique \( \pi \) so that \( \mathbf{x} \in \pi(\Sigma_N) \) and we then define the extension \( \tilde{f} \) by \( \tilde{f}(\mathbf{x}) := f(\pi^{-1}(\mathbf{x})) \). Clearly, \( \pi^{-1}(\mathbf{x}) \) is simply the coordinates of \( \mathbf{x} = (x_1, x_2, \ldots, x_N) \) permuted in increasing order. Thus \( \tilde{f} \) is defined on \( \mathbb{R}^N \) apart from a zero measure set and it is bounded since \( f \in L^\infty \). Furthermore, \( \nabla [(\tilde{f})^{1/2}] \) is also bounded since \( \nabla \sqrt{\tilde{f}} \in L^\infty \).

Since \( \tilde{f} \) is bounded, we have \( \int_{\mathbb{R}^N} f d\omega_\delta \to \int_{\Sigma_N} f d\omega = 1 \), so there is a constant \( C_\delta \) such that \( f_\delta := C_\delta \tilde{f} \) is a probability density, i.e., \( \int_{\mathbb{R}^N} f_\delta d\omega_\delta = 1 \) and clearly \( C_\delta \to 1 \) as \( \delta \to 0 \). Applying Theorem 13.6 to the measure \( \omega_\delta \) and the function \( f_\delta \) on \( \mathbb{R}^N \), we see that the LSI holds for \( \omega_\delta \), i.e.,
\[ S_{\omega_\delta}(f_\delta) \leq \frac{2}{K} D_{\omega_\delta}(\sqrt{f_\delta}), \]
or equivalently,
\[ C_\delta \int_{\mathbb{R}^N} \tilde{f} \log f d\omega_\delta + \log C_\delta \leq \frac{2C_\delta}{K} D_{\omega_\delta}(\sqrt{\tilde{f}}). \]

Now we let \( \delta \to 0 \). Using the boundedness of \( \nabla [(\tilde{f})^{1/2}] \), the weak convergence of \( \omega_\delta \) to \( \omega \) and that \( Z_{\omega,\delta} \) and \( C_\delta \) converge to 1, the Dirichlet form on the right side of the last inequality converges to \( D_\omega(\sqrt{f}) \). The first term on the left side of the inequality converges to \( \int f \log f d\omega \) by dominated convergence, and the second term converges to zero. Thus we arrive at (13.76).
In the above argument, the boundedness of \( f \) and \( \nabla \sqrt{f} \) were only used to ensure that \( f \) or rather its extension \( \tilde{f} \) has finite integral and the Dirichlet form w.r.t. the regularized measure \( \omega_\delta \), \( D_{\omega_\delta}(\sqrt{\tilde{f}}) \), converges to \( D_\omega(\sqrt{f}) \). For \( \beta \geq 1 \) we can remove these conditions by using a different extension of \( f \) to \( \mathbb{R}^N \) if \( f \in H^1(\Sigma, d\omega) \). We may assume that \( D_\omega(\sqrt{f}) < \infty \); otherwise (13.76) is a tautology. We first still assume that \( f \in L^\infty(\Sigma) \). We smoothly cut off \( f \) to be zero at the boundary of \( \Sigma_N \), i.e., we find a nonnegative sequence \( f_\varepsilon \in C_0^\infty(\Sigma_N) \) such that \( \sqrt{f_\varepsilon} \to \sqrt{f} \) in \( H^1(\Sigma, d\omega) \) and \( \int f_\varepsilon d\omega = 1 \). For \( \beta \geq 1 \) the existence of a similar sequence but \( f_\varepsilon \to f \) in \( H^1 \) was shown in Section 12.4. In fact, the same construction also shows that we can also guarantee \( \sqrt{f_\varepsilon} \to \sqrt{f} \) in \( H^1(\Sigma, d\omega) \). Now we use the LSI for the smooth functions \( f_\varepsilon \), i.e.,

\[
S_\omega(f_\varepsilon) \leq \frac{2}{K} D_\omega(\sqrt{f_\varepsilon}),
\]

and we let \( \varepsilon \to 0 \). The right hand side converges to \( D_\omega(\sqrt{f}) \) by the above choice of \( f_\varepsilon \). For the left hand side, recall that apart from a smoothing which can be dealt with via standard approximation arguments, the cutoff function was constructed in the form \( f_\varepsilon(x) = C_\varepsilon \phi_\varepsilon(x) f(x) \), where \( \phi_\varepsilon(x) \in (0, 1) \) with \( \phi_\varepsilon \rightharpoonup 1 \) monotonically pointwise and \( C_\varepsilon \) is a normalization such that \( C_\varepsilon \to 1 \) as \( \varepsilon \to 0 \). Clearly,

\[
S_\omega(f_\varepsilon) = C_\varepsilon \log C_\varepsilon \int_\Sigma \phi_\varepsilon f d\omega + C_\varepsilon \int_\Sigma f \phi_\varepsilon \log \phi_\varepsilon d\omega + C_\varepsilon \int_\Sigma \phi_\varepsilon f \log f d\omega.
\]

The first term converges to zero since \( \int \phi_\varepsilon f d\omega \leq 1 \) and \( C_\varepsilon \log C_\varepsilon \to 0 \). The second term also converges to zero by dominated convergence since \( \phi_\varepsilon \log \phi_\varepsilon \) is bounded by one and goes to zero pointwise. Finally, the last term converges to \( S_\omega(f) \) by monotone convergence. This proves (13.76) for \( \beta \geq 1 \) for any bounded \( f \). Finally we remove this last condition. Given any \( f \) with \( D_\omega(\sqrt{f}) < \infty \), we define \( f_M := \min\{f, M\} \) and \( \tilde{f}_M := C_M f_M \), where \( C_M \) is the normalization. Clearly, \( C_M \to 1 \) and \( f_M \rightharpoonup f \) pointwise. Since (13.76) holds for \( \tilde{f}_M \), we have

\[
C_M \log C_M \int f_M d\omega + C_M \int f_M \log f_M d\omega \leq \frac{2}{K} C_M \int |\nabla \sqrt{f_M}|^2 d\omega.
\] (13.84)

Now we let \( M \to \infty \). The first term on the left is just \( \log C_M \to 0 \). The second term on the left converges to \( S_\omega(f) \) by monotone convergence and \( C_M \to 1 \) and similarly the right hand side converges to \( (2/K)D_\omega(\sqrt{f}) \) by monotone convergence. This proves (13.76) for \( \beta \geq 1 \) without any additional condition on \( f \).

The Brascamp-Lieb inequality, (13.77), is proved similarly, starting from its regularized version on \( \mathbb{R}^N \)

\[
(f; f)_{\omega_\delta} \leq (\nabla f, [\mathcal{H}_\delta'' + \tilde{\mathcal{H}}'']^{-1} \nabla f)_{\omega_\delta}
\]

that follows directly from Theorem 13.13. We can then take the limit \( \delta \to 0 \) using monotone convergence on the left and the dominated convergence on the right, using that \( \mathcal{H}_\delta'' \to \mathcal{H} \) and the inverse \([\mathcal{H}_\delta'' + \tilde{\mathcal{H}}'']^{-1}\) is uniformly bounded.

For the third part of the theorem, for the proof of (13.78), we first note that the remark after (12.17) applies to the generator \( \mathcal{L}_\omega \) as well, i.e., \( \beta \geq 1 \) is necessary for the well-posedness of the equation \( \partial_t f_t = \mathcal{L}_\omega f_t \) on \( \Sigma_N \) with initial condition \( f_0 \) supported on \( \Sigma_N \). The construction of the dynamics in Section 12.4 also implies that \( f_t \in H^1(d\omega) \) for any \( t > 0 \) if \( f \in L^2(d\omega) \).

We now mimic the proof of the entropy dissipation (13.32) in our setup. Since we do not know that \( D_\omega(\sqrt{f}) < \infty \), we have to introduce a regularization \( c > 0 \) to keep \( f_t \) away from zero. We compute

\[
\frac{d}{dt} \int f_t \log(f_t + c) d\omega = \int (\mathcal{L} f_t) \log(f_t + c) d\omega + \int f_t \frac{\mathcal{L} f_t}{f_t + c} d\omega
\]

\[
= -\int \frac{|\nabla f_t|^2}{f_t + c} d\omega - \int \frac{c \mathcal{L} f_t}{f_t + c} d\omega
\]

\[
= -\int \frac{|\nabla f_t|^2}{f_t + c} d\omega - \int \frac{c |\nabla f_t|^2}{(f_t + c)^2} d\omega.
\] (13.85)
Owing to the regularization \( c > 0 \), we used integration by parts for \( H^1(d\omega) \) functions only. Dropping the last term that is negative and integrating from 0 to \( t \), we obtain
\[
\int f_t \log(f_t + c) d\omega + \int_0^t ds \int \frac{|\nabla f_s|^2}{f_s + c} d\omega \leq \int f_0 \log(f_0 + c) d\omega. \tag{13.86}
\]
Since \( f_t \log (f_t + c) \geq 0 \), we have
\[
\int f_t \log f_t d\omega + \int_0^t \int ds \frac{|\nabla f_s|^2}{f_s + c} d\omega \leq \int f_0 \log(f_0 + c) d\omega. \tag{13.87}
\]
Note that both terms on the left hand side are nonnegative. Now we let \( c \to 0 \). By monotone convergence and \( S_\omega(f_0) < \infty \), we get
\[
\int f_t \log f_t d\mu + \int_0^t \int ds \frac{|\nabla f_s|^2}{f_s} d\omega \leq \int f_0 \log f_0 d\omega \tag{13.88}
\]
or
\[
S_\omega(f_t) + 4 \int_0^t D_\omega(\sqrt{f_s}) ds \leq S_\omega(f_0). \tag{13.89}
\]
This is the entropy dissipation inequality in a time integral form. Notice that neither equality nor the differential version like (13.32) is claimed. Note that by integrating (13.85) between \( t \) and \( \tau \), a similar argument yields
\[
S_\omega(f_t) + 4 \int_\tau^t D_\omega(\sqrt{f_s}) ds \leq S_\omega(f_\tau), \quad t \geq \tau \geq 0. \tag{13.90}
\]
In particular the entropy decays:
\[
0 \leq S_\omega(f_t) \leq S_\omega(f_\tau), \quad t \geq \tau \geq 0. \tag{13.91}
\]

Now we use the LSI (13.76) to estimate \( D_\omega(\sqrt{f_s}) \) in (13.90) and recall that the LSI holds for any \( f_s \) since \( \beta \geq 1 \). We get
\[
S_\omega(f_t) + 2K \int_\tau^t S_\omega(f_s) ds \leq S_\omega(f_\tau), \quad t \geq \tau \geq 0. \tag{13.92}
\]
A standard calculus exercise shows that \( S_\omega(f_t) \leq e^{-2Kt}S_\omega(f_0) \) for all \( t \geq 0 \). One possible argument is to fix any \( \delta > 0 \) and choose \( \tau = (n - 1)\delta \), \( t = n\delta \) with \( n = 1, 2, \ldots \) in (13.92). By monotonicity of the entropy we have
\[
(1 + 2K\delta)S_\omega(f_{n\delta}) \leq S_\omega(f_{(n-1)\delta}),
\]
for any \( n \), and by iteration we obtain
\[
S_\omega(f_{n\delta}) \leq (1 + 2K\delta)^{-n}S_\omega(f_0).
\]
Setting \( \delta = t/n \) and letting \( n \to \infty \) we get (13.78). This completes the proof of Theorem 13.14. \( \square \)
14 Universality of the Dyson Brownian Motion

In this section we assume $\beta \geq 1$ since Dyson Brownian Motion in a strong sense is well defined only for these values of $\beta$. However, some results will hold for any $\beta > 0$ that we will comment on separately. We consider the Gaussian equilibrium measure $\mu = \mu_G \sim \exp(-\beta NH)$ on $\Sigma_N$, defined in (12.13) in Section 12.3 with the Gaussian potential $V(\lambda) = \frac{1}{2}\lambda^2$. From (12.14) we recall the corresponding Dirichlet form

$$D_\mu(f) := \frac{1}{\beta N} \sum_{i=1}^N (\partial_i f)^2 d\mu = \frac{1}{\beta N} \|\nabla f\|_{L^2(\mu)}^2, \quad (\partial_i \equiv \partial_{\lambda_i}),$$

and the generator $\mathcal{L} = \mathcal{L}_G$ defined via

$$D_\mu(f) = -\int f \mathcal{L}_G f d\mu,$$

and explicitly given by $\mathcal{L}_G = \frac{1}{\beta N} \Delta - (\nabla \mathcal{H}) \cdot \nabla$. The corresponding dynamics is given by (12.17), i.e.,

$$\partial_t f_t = \mathcal{L}_G f_t, \quad t \geq 0.$$  

In this section, we will drop all subscripts $G$.

As remarked in the previous section, the Hamiltonian $\mathcal{H}$ is convex since the Hessian of the Hamiltonian of $\mu$ satisfies $\nabla^2(\beta NH) \geq \beta N/2$ by (13.70). Taking the different normalization of the Dirichlet form in (13.25) and (14.1) into account, Theorem 13.6 (actually, its extension to $\Sigma_N$ in Theorem 13.14 with $U_j \equiv 0$) guarantees that $\mu$ satisfies the LSI in the form

$$S_\mu(f) \leq 4D_\mu(\sqrt{f}),$$

and the relaxation time to equilibrium is of order one. Furthermore, we have the exponential convergence to the equilibrium $\mu_G$ in the sense of total variation norm (13.78) at exponential speed on time scales beyond order one for initial data with finite entropy.

The following theorem is the main result of this section. It shows that under a certain condition on the rigidity of the eigenvalues the relaxation times of the DBM for observables depending only on the eigenvalue differences are in fact much shorter than order one. The main assumption is the a priori estimate (12.18) which we restate here.

**A priori Estimate:** There exists $\xi > 0$ such that

$$Q := \sup_{0 \leq t \leq N} \frac{1}{N} \int \sum_{j=1}^N (\lambda_j - \gamma_j)^2 f_t(\lambda) \mu_G(d\lambda) \leq CN^{-2+2\xi}$$

with a constant $C$ uniformly in $N$. We also assume that after time $1/N$ the solution of the equation (14.3) satisfies the bound

$$S_\mu(f_{1/N}) \leq CN^m$$

for some fixed $m$. Later in Lemma 14.6 we will show that for $\beta = 1, 2$ this bound automatically holds.

**Theorem 14.1** (Gap universality of the Dyson Brownian motion for short time). [64, Theorem 4.1]

Let $\beta \geq 1$ and assume (14.5). Fix $n \geq 1$ and an array of positive integers, $m = (m_1, m_2, \ldots, m_n) \in \mathbb{N}_+^n$. Let $G : \mathbb{R}^n \to \mathbb{R}$ be a bounded smooth function with compact support and define

$$G_{i,m}(x) := G\left(N(x_i - x_{i+m_1}), N(x_{i+m_1} - x_{i+m_2}), \ldots, N(x_{i+m_{n-1}} - x_{i+m_n})\right).$$

Then for any $\xi \in (0, \frac{1}{2})$ and any sufficiently small $\varepsilon > 0$, independent of $N$, there exist constants $C, c > 0$, depending only on $\varepsilon$ and $G$, such that for any $J \subset \{1, 2, \ldots, N - m_n\}$ we have

$$\sup_{t \geq N^{-1+2\xi+\varepsilon}} \left| \frac{1}{|J|} \sum_{i \in J} G_{i,m}(x)(f_i d\mu - d\mu) \right| \leq CN^{\xi} \sqrt{\frac{N^2 Q}{|J|}} + Ce^{-cN^{\xi}}.$$  

113
In particular, if (14.4) holds, then for any \( t \geq N^{-1+2\xi+2\varepsilon+\delta} \) with some \( \delta \geq 0 \), we have
\[
\left| \int \frac{1}{|J|} \sum_{i \in J} G_{i,m}(x)(f_t d\mu - d\mu) \right| \leq C \frac{1}{\sqrt{|J|N^{\delta-1}}} + Ce^{-cN^z}.
\]
(14.8)

Thus the gap distribution, averaged over \( J \) indices, coincides for \( f_t d\mu \) and \( d\mu \) provided that \( |J|N^{\delta-1} \to \infty \).

We stated this result with a very general test function (14.6) because we will need this form later. Obviously, controlling test functions of the form (14.6) for all \( n \) is equivalent to controlling test functions of neighboring gaps only (maybe with a larger \( n \)), so the reader may think only of the case \( m_1 = 1 \), \( m_2 = 2, \ldots, m_n = n \).

In our applications, \( J \) is chosen to be the indices of the eigenvalues in the interval \([E-b, E+b]\) and thus \( |J| \approx Nb \). This identifies the averaged gap distributions of eigenvalues. However, Theorem 12.4 concerns correlation functions and not gap distributions directly. It is intuitively clear that the information carried in gap distribution or correlation functions are equivalent if both statistics are averaged on a scale larger than the typical fluctuation of a single eigenvalue (which is smaller than \( N^{-1+\varepsilon} \) in the bulk by (11.32)). This standard fact will be proved in Section 14.5.

As pointed out after Theorem 12.4, the input of this theorem, the apriori estimate (14.4), identifies the location of the eigenvalues only on a scale \( N^{-1+\varepsilon} \) which is much weaker than the \( 1/N \) precision encoded in the rescaled eigenvalue differences in (14.7). Moreover, by the rigidity estimate (11.32), the a priori estimate (14.4) holds for any \( \xi > 0 \) if the initial data of the DBM is a generalized Wigner ensemble. Therefore, Theorem 14.1 holds for any \( t \geq N^{-1+\varepsilon} \) for any \( \varepsilon > 0 \) and this establishes Dyson’s conjecture (described in Section 12.3) in the sense of averaged gap distributions for any generalized Wigner matrices.

### 14.1 Main ideas behind the proof of Theorem 14.1

The key method is to analyze the relaxation to equilibrium of the dynamics (13.30). This approach was first introduced in Section 5.1 of [63]; the presentation here follows [64].

Recall the convexity inequality (13.70) for the Hessian of the Hamiltonian
\[
\left\langle v, \nabla^2 \mathcal{H}(x)v \right\rangle \geq \frac{1}{2} \|v\|^2 + \frac{1}{N} \sum_{i<j} \frac{(v_i - v_j)^2}{(x_i - x_j)^2} \geq \frac{1}{2} \|v\|^2, \quad v \in \mathbb{R}^N,
\]
(14.9)

which guarantees a relaxation to equilibrium on a time scale of order one. The key idea to prove Theorem 14.1 is that the relaxation time is in fact much shorter than order one for local observables that depend only on the eigenvalue differences. The convexity bound (14.9) shows that the relaxation in the direction \( v_i - v_j \) is much faster than order one provided that \( x_i - x_j \) are close. However, this effect is hard to exploit directly due to that all modes of different wavelengths are coupled. Our idea is to add an auxiliary strongly convex potential \( \widehat{\mathcal{H}}(x) \) to the original Hamiltonian \( \mathcal{H} \) to “speed up” the convergence to local equilibrium. On the other hand, we will also show that the cost of this speeding up can be effectively controlled if the a priori estimate (12.18) holds.

The auxiliary potential \( \widehat{\mathcal{H}}(x) \) is defined by
\[
\widehat{\mathcal{H}}(x) := \sum_{j=1}^{N} U_j(x_j), \quad U_j(x) := \frac{1}{2\tau} (x_j - \gamma_j)^2,
\]
(14.10)
i.e., it is a quadratic confinement on the scale \( \sqrt{\tau} \) for each eigenvalue near its classical location, where the parameter \( 0 < \tau < 1 \) will be chosen later on. The total Hamiltonian is given by
\[
\widehat{\mathcal{H}} := \mathcal{H} + \widehat{\mathcal{H}},
\]
(14.11)
where \( \mathcal{H} \) is the Gaussian Hamiltonian given by (4.12) with \( V(x) = \frac{1}{2}x^2 \). The measure with Hamiltonian \( \widehat{\mathcal{H}} \),
\[
d\omega := \omega(x)dx, \quad \omega := e^{-\beta N \widehat{\mathcal{H}}}/Z,
\]
(14.12)
will be called the \textit{local relaxation measure}.

The \textit{local relaxation flow} is defined to be the flow with the generator characterized by the natural Dirichlet form w.r.t. $\omega$, explicitly, $\tilde{L}$:

\[
\tilde{L} = \mathcal{L} - \sum_j b_j \partial_j, \quad b_j = U_j'(x_j) = \frac{x_j - \gamma_j}{\tau}.
\]

(14.13)

We will choose $\tau \ll 1$ so that the additional term $\hat{H}$ substantially increases the lower bound (13.70) on the Hessian, hence speeding up the dynamics so that the relaxation time is at most $\tau$.

The idea of adding an artificial potential $\hat{H}$ to speed up the convergence appears to be unnatural here. The current formulation is a streamlined version of a much more complicated approach that appeared in [63] which took ideas from the earlier work [59]. Roughly speaking, in the hydrodynamical limit, the short wavelength modes always have shorter relaxation times than the long wavelength modes. A direct implementation of this idea is extremely complicated due to the logarithmic interaction that couples short and long wavelength modes. Adding a strongly convex auxiliary potential $\hat{H}$ shortens the relaxation time of the long wavelength modes, but it does not affect the short modes, i.e., the local statistics, which are our main interest. The analysis of the new system is much simpler since now the relaxation is faster, uniform for all modes. Finally, we need to compare the local statistics of the original system with those of the modified one. It turns out that the difference is governed by $(\nabla \hat{H})^2$ which can be directly controlled by the a priori estimate (14.4).

Our method for enhancing the convexity of $\mathcal{H}$ is reminiscent of a standard convexification idea concerning metastable states. To explain the similarity, consider a particle near one of the local minima of a double well potential separated by a local maximum, or energy barrier. Although the potential is not convex globally, one may still study a reference problem defined by convexifying the potential along with the well in which the particle initially resides. Before the particle reaches the energy barrier, there is no difference between these two problems. Thus questions concerning time scales shorter than the typical escape time can be conveniently answered by considering the convexified problem; in particular the escape time in the metastability problem itself can be estimated by using convex analysis. Our DBM problem is already convex, but not sufficiently convex. The modification by adding $\hat{H}$ enhances convexity without altering the local statistics. This is similar to the convexification in the metastability problem which does not alter events before the escape time.

\subsection*{14.2 Proof of Theorem 14.1}

We will work with the measures $\mu$ and $\omega$ defined on the simplex $\Sigma_N$ and we present the proof as if Theorem 13.6 and its proof were valid not only for measures on $\mathbb{R}^N$ but also on $\Sigma_N$. Theorem 13.14 demonstrated that this is indeed true except that the exponential decay to equilibrium was proven only in entropy sense and not Dirichlet form sense as in (13.31). For pedagogical simplicity, we first neglect this technicality and in Section 14.3 we will comment on how to remedy this problem with the help of the regularization introduced in Section 13.7.

The core of the proof is divided into three theorems. For the flow with generator $\tilde{L}$, we have the following estimates on the entropy and Dirichlet form.

\textbf{Theorem 14.2.} Let $\beta \geq 1$ arbitrary. Consider the forward equation

\[
\partial_t q_t = \tilde{L} q_t, \quad t \geq 0,
\]

(14.14)

with the reversible measure $\omega$ defined in (14.12). Let the initial condition $q_0$ satisfy $\int q_0 d\omega = 1$. Then we have the following estimates

\[
\partial_t D_\omega(\sqrt{q_t}) \leq -\frac{2}{\tau} D_\omega(\sqrt{q_t}) - \frac{1}{\beta N^2} \int \sum_{i,j=1}^N \frac{(\partial_i \sqrt{q_t} - \partial_j \sqrt{q_t})^2}{(x_i - x_j)^2} d\omega,
\]

(14.15)
\[
\frac{1}{\beta N^2} \int_0^\infty ds \int \sum_{i,j=1}^N \frac{(\partial_i \sqrt{q} - \partial_j \sqrt{q})^2}{(x_i - x_j)^2} \, d\omega \leq D_\omega(\sqrt{q_0}),
\]  
(14.16)

and the logarithmic Sobolev inequality
\[
S_\omega(q_0) \leq C \tau D_\omega(\sqrt{q_0})
\]  
(14.17)

with a universal constant \(C\). Thus the relaxation time to equilibrium is of order \(\tau\):
\[
S_\omega(q_t) \leq e^{-Ct/\tau} S_\omega(q_0), \quad D_\omega(q_t) \leq e^{-Ct/\tau} D_\omega(q_0).
\]  
(14.18)

**Proof.** Denote by \(h = \sqrt{q}\) and by a calculation similar to (13.33) we have the equation
\[
\partial_t D_\omega(h_t) = \partial_t \frac{1}{\beta N} \int (\nabla h)^2 e^{-\beta N \tilde{H}} \, dx \leq -\frac{2}{\beta N} \int \nabla h(\nabla^2 \tilde{H}) \nabla h e^{-\beta N \tilde{H}} \, dx.
\]  
(14.19)

In our case, (13.70) and (14.10) imply that the Hessian of \(\tilde{H}\) is bounded from below as
\[
\nabla h(\nabla^2 \tilde{H}) \nabla h \geq \frac{1}{\tau} \sum_j (\partial_j h)^2 + \frac{1}{2 \beta N} \sum_{i,j} \frac{1}{(x_i - x_j)^2} (\partial_i h - \partial_j h)^2.
\]  
(14.20)

This proves (14.15) and (14.16). The rest can be proved by straightforward arguments analogously to (13.32)–(13.37). \(\square\)

Notice that the estimate (14.16) is an additional information that we extracted from the Bakry-Émery argument by using the second term in the Hessian estimate (13.70). It plays a key role in the next theorem.

**Theorem 14.3** (Dirichlet form inequality). Let \(\beta \geq 1\) be arbitrary. Let \(q\) be a probability density with respect to the relaxation measure \(\omega\) from (14.12), i.e., \(\int q \, d\omega = 1\). Fix \(n \geq 1\), an array of positive integers \(m \in \mathbb{N}^n\) and a smooth function \(G : \mathbb{R}^n \to \mathbb{R}\) with compact support as in Theorem 14.1. Then for any \(J \subset \{1, 2, \ldots, N - m_n\}\) and any \(t > 0\) we have
\[
\left| \int \frac{1}{|J|} \sum_{i \in J} G_{i,m}(x) (q \, d\omega - d\omega) \right| \leq C \left( t D_\omega(\sqrt{q}) \right)^{1/2} + C \sqrt{S_\omega(q)} e^{-ct/\tau}.
\]  
(14.21)

**Proof.** We give the proof for the \(\beta \geq 1\) case here since this is relevant for Theorem 14.1. The general case \(\beta > 0\) with additional assumptions will be discussed in Section 14.4. For simplicity of the notation, we consider only the case \(n = 1, m_1 = 1\), \(G_{i,m}(x) = G(N(x_i - x_{i+1}))\). Let \(q_t\) satisfy
\[
\partial_t q_t = \tilde{L} q_t, \quad t \geq 0,
\]
with an initial condition \(q_0 = q\). We write
\[
\int \left[ \frac{1}{|J|} \sum_{i \in J} G(N(x_i - x_{i+1})) \right] (q - 1) \, d\omega
\]  
(14.22)

The second term in (14.22) can be estimated by (13.8), the decay of the entropy (14.18) and the boundedness of \(G\); this gives the second term in (14.21).

To estimate the first term in (14.22), by the evolution equation \(\partial_t q_t = \tilde{L} q_t\) and the definition of \(\tilde{L}\) we have
\[
\int \frac{1}{|J|} \sum_{i \in J} G(N(x_i - x_{i+1})) q_t \, d\omega - \int \frac{1}{|J|} \sum_{i \in J} G(N(x_i - x_{i+1})) q_0 \, d\omega
\]  

\[
= \int_0^t ds \int \frac{1}{|J|} \sum_{i \in J} G'(N(x_i - x_{i+1})) (\partial_i q_s - \partial_{i+1} q_s) \, d\omega.
\]  

116
From the Schwarz inequality and $\partial q = 2\sqrt{q} \partial \sqrt{q}$, the last term is bounded by

$$
2 \left[ \int_0^t ds \int_{\mathbb{R}^N} \frac{N^2}{|J|^2} \sum_{i<j} \left[ G'(N(x_i - x_{i+1})) \right]^2 (x_i - x_{i+1})^2 q_i \omega \right]^{1/2} 
$$

$$
\times \left[ \int_0^t ds \int_{\mathbb{R}^N} \frac{1}{N^2} \sum_i \frac{1}{(x_i - x_{i+1})^2} \left[ \partial_t \sqrt{q_i} \right] - \partial_{i+1} \sqrt{q_i} \right]^2 \right]^{1/2} \leq C \left( \frac{D_\omega(\sqrt{q})}{|J|} \right)^{1/2},
$$

(14.23)

where we have used (14.16) and that \( \left[ G'(N(x_i - x_{i+1})) \right]^2 (x_i - x_{i+1})^2 \leq CN^{-2} \) due to \( G \) being smooth and compactly supported.

Alternatively, we could have directly estimated the left hand side of (14.21) by using the total variation norm between \( q \omega \) and \( \omega \), which in turn could be estimated by the entropy and the Dirichlet form using the logarithmic Sobolev inequality, i.e., by

$$
C \int |q - 1| d\omega \leq C \sqrt{S_\omega(q)} \leq C \sqrt{\tau D_\omega(\sqrt{q})}.
$$

(14.24)

However, compared with this simple bound, the estimate (14.21) gains an extra factor \( |J| \approx N \) in the denominator, i.e., it is in terms of Dirichlet form per particle. The improvement is due to the fact that the observable in (14.21) depends only on the gap, i.e., difference of points. This allows us to exploit the additional term (14.16) gained in the Bakry-Émery argument. This is a manifestation of the general observation that gap-observables behave much better than point-observables.

The final ingredient in proving Theorem 14.1 is the following entropy and Dirichlet form estimates.

**Theorem 14.4.** Let \( \beta \geq 1 \) be arbitrary. Suppose that (13.70) holds and recall the definition of \( Q \) from (12.18). Fix some (possibly \( N \)-dependent) \( \tau > 0 \) and consider the local relaxation measure \( \omega \) with this \( \tau \). Set \( \psi := \omega / \mu \) and let \( g_t := f_t / \psi \) with \( f_t \) solving the evolution equation (14.3). Suppose there is a constant \( m \) such that

$$
S(f, \mu | \omega) \leq CN^m.
$$

(14.25)

Fix any small \( \varepsilon > 0 \). Then for any \( t \in [\tau N^2, N] \) the entropy and the Dirichlet form satisfy the estimates:

$$
S(g_t \omega | \omega) \leq CN^2 Q \tau^{-1}, \quad D_\omega(\sqrt{g_t}) \leq CN^2 Q \tau^{-2},
$$

(14.26)

where the constants depend on \( \varepsilon \) and \( m \).

**Remark 14.5.** It will not be difficult to check that if the initial data is given by a Wigner ensemble, then (14.25) holds without any assumption, see Lemma 14.6.

**Proof of Theorem 14.4.** The evolution of the entropy \( S(f, \mu | \omega) = S(f, \mu | \psi \mu) \) can be computed by Lemma 13.11 as

$$
\partial_t S(f, \mu | \omega) = -\frac{4}{\beta N} \sum_j \left( \partial_j \sqrt{g_t} \right)^2 \psi d\mu + \int (L g_t) \psi d\mu,
$$

where \( L \) is defined in (14.3) and we used that \( \psi = \omega / \mu \) is time independent. Hence we have, by using (14.13),

$$
\partial_t S(f, \mu | \omega) = -\frac{4}{\beta N} \sum_j \left( \partial_j \sqrt{g_t} \right)^2 d\omega + \int \bar{L} g_t d\omega + \sum_j b_j \partial_j g_t d\omega.
$$

Since \( \omega \) is \( \bar{L} \)-invariant and time independent, the middle term on the right hand side vanishes. From the Schwarz inequality and \( \partial g = 2\sqrt{g} \partial \sqrt{g} \), we have

$$
\partial_t S(f, \mu | \omega) \leq -2D_\omega(\sqrt{g_t}) + CN \sum_j b_j^2 g_t d\omega \leq -2D_\omega(\sqrt{g_t}) + CN^2 Q \tau^{-2}.
$$

(14.27)
Notice that (14.27) is reminiscent to (13.32) for the derivative of the entropy of the measure \( g_t \omega = f_t \mu \) with respect to \( \omega \). The difference is, however, that \( g_t \) does not satisfy the evolution equation with the generator \( \mathcal{L} \). The last term in (14.27) expresses the error.

Together with the logarithmic Sobolev inequality (14.17), we have

\[
\partial_t S(f_t \mu | \omega) \leq -2D_\omega(\sqrt{\rho}) + CN^2Qr^{-2} \leq -C\tau^{-1}S(f_t \mu | \omega) + CN^2Qr^{-2}.
\] (14.28)

Integrating the last inequality from \( \tau \) to \( t \) and using the assumption (14.25) and \( t \geq \tau N^\varepsilon \), we have proved the first inequality of (14.26). Using this result and integrating (14.27), we have

\[
\int_\tau^t D_\omega(\sqrt{\rho_s})ds \leq CN^2Qr^{-1}.
\] (14.29)

Notice that

\[
D_\mu(\sqrt{f_s}) = \frac{1}{\beta N} \int \frac{|\nabla (g_s \psi)|^2}{g_s \psi} \, d\omega \leq \frac{C}{\beta N} \int \left[ |\nabla g_s|^2 + |\nabla \log \psi|^2 g_s \right] \, d\omega = CD_\omega(\sqrt{g_s}) + CN^2Qr^{-2}
\] (14.30)

by a Schwarz inequality. Thus from (14.29), after restricting the integration and using \( t \geq 2\tau \), we get

\[
CN^2Qr^{-1} \geq \int_{-\tau}^t D_\omega(\sqrt{g_s})ds \geq \int_{-\tau}^t \left[ \frac{1}{C}D_\mu(\sqrt{f_s}) - CN^2Qr^{-2} \right]ds \geq \frac{T}{C}D_\mu(\sqrt{f_t}) - CN^2Qr^{-1},
\] (14.31)

where, in the last step, we used that \( D_\mu(\sqrt{f_t}) \) is decreasing in \( t \) which follows from the convexity of the Hamiltonian of \( \mu \), see e.g. (13.33). Using the opposite inequality \( D_\omega(\sqrt{g_t}) \leq CD_\mu(\sqrt{f_t}) + CN^2Qr^{-2} \), that can be proven similarly to (14.30), we obtain

\[
D_\omega(\sqrt{g_t}) \leq CN^2Qr^{-2},
\]

i.e., the second inequality of (14.26).

Finally, we complete the proof of Theorem 14.1. For any given \( t > 0 \) we now choose \( \tau := tN^{-\varepsilon} \) and we construct the local relaxation measure \( \omega \) with this \( \tau \) as in (14.14). Set \( \psi = \omega/\mu \) and let \( q := g_t = f_t/\psi \) be the density \( q \) in Theorem 14.3. We would like to apply Theorem 14.4 and for this purpose we need to verify the assumption (14.25). By the definitions of \( \omega \), \( \mu \) and \( \psi = \omega/\mu \), we have

\[
S(f_T \mu | \omega) = \int f_T \log f_T d\mu - \int f_T \log \psi d\mu = S_\mu(f_T) - \int f_T \log \psi d\mu.
\] (14.32)

We can bound the last term by

\[
- \int f_T \log \psi d\mu \leq CN \int f_T W d\mu + \log \frac{Z}{\tilde{Z}} \leq CN^2Qr^{-1} \leq CN^m
\] (14.33)

for some \( m \), where we have used that \( \tilde{Z} \leq Z \) since \( \tilde{\mathcal{H}} \geq \mathcal{H} \). To verify the assumption (14.25), it remains to prove that \( S_\mu(f_T) \leq CN^m \) for some \( m \). This inequality follows from the following lemma.

**Lemma 14.6.** Let \( \beta = 1, 2 \). Suppose the initial data \( f_0 \) of the DBM is given by the eigenvalue distribution of a Wigner matrix. Then for any \( \tau > 0 \) we have

\[
S_\mu(f_\tau) \leq CN^2(1 - \log (1 - e^{-\tau})),
\] (14.34)

**Proof.** For simplicity, we consider only the case \( \beta = 1 \), i.e., the real Wigner matrices. Recall that the probability measure \( f_\tau \mu \) is the same as the eigenvalue distribution of the Gaussian divisible matrix (12.21):

\[
H_\tau = e^{-\tau/2}H_0 + (1 - e^{-\tau})^{1/2} H^G,
\] (14.35)
where $H_0$ is the initial Wigner matrix and $H^G$ is an independent standard GOE matrix. Since the entropy is monotonic w.r.t. taking a marginal, (13.21), we have

$$S_\mu(f) = S(f_\tau|\mu) \leq S(\mu,H_\tau|\mu,H^G),$$

where $\mu$ and $\mu,H_\tau$ are the laws of the matrix $H_\tau$ and $H^G$ respectively. Since the laws of both $\mu$ and $\mu,H_\tau$ and $\mu,H^G$ are given by the product of the laws of the matrix elements, from the additivity of entropy (13.11), $S(\mu,H_\tau|\mu,H^G)$ is equal to the sum of the relative entropies of the matrix elements. Recall that the variances of off-diagonal and diagonal entries for GOE differ by a factor 2. For simplicity of notations, we consider only the off-diagonal terms. Let $\gamma = 1 - e^{-\tau}$ and denote by $g_\alpha$ the standard Gaussian distribution with variance $\alpha$, i.e.,

$$g_\alpha(x) := \frac{1}{\sqrt{2\pi}\alpha} \exp\left(-\frac{x^2}{2\alpha}\right).$$

Let $\varrho$ be the probability density of $(1 - \gamma)^{1/2} \zeta = e^{-\gamma/2} \zeta$ where $\zeta$ is the random variable for an off-diagonal matrix element. By definition, the probability density of the matrix element of $H_\tau$ is given by $\zeta = \varrho \ast g_2/N$. Therefore Jensen’s inequality yields

$$S(\zeta|g_2/N) = S\left(\int dy \varrho(y) g_2/N(y - y) \bigg| g_2/N\right) \leq \int dy \varrho(y) S(g_2/N(y - y)|g_2/N).$$

(14.36)

By explicit computation one finds

$$S(g_\sigma^2(y - y)|g_\sigma^2) = \log \frac{s}{\sigma} + \frac{y^2}{2s^2} + \frac{\sigma^2}{2s^2} - \frac{1}{2}.$$  

In our case, we have

$$S(g_2/N(y - y)|g_2/N) = \frac{1}{2} \left(\frac{N}{2} y^2 - \log \gamma + \gamma - 1\right).$$

We can now continue the estimate (14.36). Using $\int y^2 \varrho, dy = 2/N$, we obtain

$$S(\zeta|g_2/N) \leq C - \log \gamma,$$

and the claim follows. \qed

We now return to the proof of Theorem 14.1. We can apply Lemma 14.6 with the choice $\tau = N^{-1+2\xi}$, where $\xi \in (0,\frac{1}{2})$ is from the assumption (14.4). Together with (14.32)–(14.33), this implies that (14.25) holds. Thus Theorem 14.4 and Theorem 14.3 imply for any $t \in [\tau N^\xi, N]$ that

$$\left|\int \frac{1}{N} \sum_{i \in J} G_{m,i} (x)(f_4 d\mu - d\omega)\right| \leq C \left(\frac{D_\omega \langle \sqrt{q}\rangle}{|J|}\right)^{1/2} + C \sqrt{|S_\omega(q)| e^{-cN^\xi}}$$

(14.37)

$$\leq C \left(\frac{N^2 Q}{|J| \tau^2}\right)^{1/2} + C e^{-cN^\xi} \leq CN^\xi \sqrt{\frac{N^2 Q}{|J| t}} + C e^{-cN^\xi},$$

i.e., the local statistics of $f_4 \mu$ and $\omega$ are the same for any initial data $f_\tau$.

Applying the same argument to the Gaussian initial data, $f_\tau = f_\tau = 1$, we can also compare $\mu$ and $\omega$. We have thus proved the estimate (14.7). Finally, if $t \geq N^{-1+2\xi + s + 2\xi}$, then the assumption (14.4) guarantees that,

$$\sqrt{\frac{N^2 Q}{|J| t}} \leq \frac{1}{\sqrt{|J| N^{s-1}}},$$

which proves Theorem 14.1. \qed
14.3 Restriction to the simplex via regularization

In the previous section we tacitly assumed that the Bakry-Émery analysis from Section 13.3 extends to $\Sigma$. Strictly speaking, this is not fully rigorous. There are two options to handle this technical point.

As we remarked after Theorem 13.14, there is a direct method via cutoff functions to make the Bakry-Émery argument rigorous on $\Sigma$ if $\beta \geq 1$ (see Appendix B of [64]). The same technique can also make every step in Section 14.2 rigorous on $\Sigma$.

We present here a more transparent argument that relies on the regularized measure $\omega_\delta$ on $\mathbb{R}^N$ already used in the proof of the LSI in Theorem 13.14. This path is technically simpler since it performs the core of the analysis on $\mathbb{R}^N$, and it uses an extra input, the strong solution to the DBM, Theorem 12.2, to remove the regularization at the end. We now explain the correct procedure of the proof of Theorem 14.1 using this regularization. The goal is to prove (14.37).

For any small $\delta > 0$ we introduce the regularized versions $\mu_\delta$ and $\omega_\delta$ of the measures $\mu$ and $\omega$ as in Section 13.7. Let $\mathcal{L}_{\omega_\delta}$ be the generator of the regularized measure $\omega_\delta$ on $\mathbb{R}^N$ and let $f_{t,\delta}$ be the solution to $\partial_t f_{t,\delta} = \mathcal{L}_{\omega_\delta} f_{t,\delta}$ on $\mathbb{R}^N$ with the same initial condition as for $\delta = 0$, i.e., $f_{0,\delta}$ is the extension of $f_0$ to the entire $\mathbb{R}^N$ with $f_{0,\delta}(x) = 0$ for $x \not\in \Sigma_N$.

The three key theorems, Theorems 14.2–14.4, of Section 14.2 were formulated and proven for the regularized setup on $\mathbb{R}^N$, so they can be applied to the regularized dynamics. With the notations in these theorems, the conclusion is that if

$$ S(f_{t,\delta} \mu_\delta | \omega_\delta) \leq C N^m $$

(see (14.25)), then

$$ \left| \frac{1}{N} \sum_{i \in J} g_{m,i}(x)(f_{t,\delta} d\mu_\delta - d\omega_\delta) \right| \leq C N^\varepsilon \sqrt{N^2 Q_\delta} / J |t| + C e^{-cN^\varepsilon}, \quad t \in [\tau N^\varepsilon, N], $$

(14.39)

where $Q_\delta$ is defined exactly as in (14.4) with the regularized measures, i.e.,

$$ Q_\delta := \sup_{0 \leq t \leq N} \frac{1}{N} \int \sum_{j=1}^{N} (x_j - \gamma_j)^2 f_{t,\delta}(x) \mu_\delta(\mathrm{d}x). $$

Now we let $\delta \to 0$. Since $\omega_\delta$ converges weakly to $\omega$ in the sense that $\omega_\delta(A) \to \omega(A \cap \Sigma_N)$ for any subset $A \subset \mathbb{R}^N$, we have $\int \omega(x) d\omega_\delta \to \int \omega(x) d\omega$ for any bounded observable $O$. Setting

$$ O(x) = \frac{1}{N} \sum_{i \in J} g_{m,i}(x), $$

(14.41)

we see that the second term within the integral in the left hand side of (14.39) converges to the corresponding term in (14.37). On the right hand side a similar argument shows that $Q_\delta \to Q$ as $\delta \to 0$. Strictly speaking, the observable $(x_j - \gamma_j)^2$ is unbounded, but a very simple cutoff argument shows that

$$ \mathbb{P} f_{t,\delta} \mu_\delta \left( \max_j |x_j| \geq N^2 \right) \leq e^{-cN} $$

uniformly in $\delta$ and $t \in [0, N]$.

For the first term in the left side of (14.39), we use the stochastic representation of the solution to $\partial_t f_{t,\delta} = \mathcal{L}_{\omega_\delta} f_{t,\delta}$ to note that

$$ \int_{\mathbb{R}^N} O(x) f_{t,\delta}(x) d\omega_\delta = \mathbb{E}^{f_{0,\delta}}_{\omega_\delta} \mathbb{E}_{\delta}^x O(x(t)), $$

(14.42)

where $\mathbb{E}_{\delta}^x$ denotes the expectation with respect to the law of the regularized DBM $(x(t))_t$, see (13.82), starting from $x_0$, and $\mathbb{E}^{f_{0,\delta}}_{\omega_\delta}$ denotes the expectation of $x_0$ with respect to the measure $f_{0,\delta} \omega$. Now we use the existence of the strong solution to the DBM (13.74) for any $\beta \geq 1$, that can be obtained exactly as in Theorem 12.2 for the $U_j \equiv 0$ case. Since $x(t)$ is continuous and it remains in the open set $\Sigma_N$, the probability
that up to a fixed time \( t \) it remains away from a \( \delta \)-neighborhood of the boundary of \( \Sigma_N \) goes to 1 as \( \delta \to 0 \), i.e.,

\[
\lim_{\delta \to 0} \mathbb{P}^\omega_\delta \left( \text{dist}(x(s), \partial \Sigma_N) \geq \delta : s \in [0, t] \right) = 1.
\]

Notice that (13.74) and (13.82) are exactly the same for paths that stay away from the boundary of \( \Sigma_N \). This means that the right hand side of (14.42) converges to \( \mathbb{E}^{f_0} \mathbb{E}^{x_0} \mathbb{O}(x(t)) \), where \( \mathbb{E}^{x_0} \) denotes expectation with respect to the law of (13.74). This proves that

\[
\lim_{\delta \to 0} \int_{\mathbb{R}^N} O(x) f_{t, \delta}(x) d\omega_\delta = \int_{\Sigma_N} O(x) f_t(x) d\omega,
\]

which completes the proof of (14.37).

Finally, we need to verify the condition (14.38). Following (14.32)--(14.33), we only need to show that \( S_{\mu_\delta}(f_{\tau, \delta}) \leq CN^m \) for sufficiently small \( \delta > 0 \). We first run the DBM for a very short time and replace the original matrix ensemble \( H_0 \) by \( H_\sigma \), its tiny Gaussian convolution of variance \( \sigma = N^{-10} \), see (14.35). This is not a restriction, since Theorem 14.1 anyway concerns \( f_t \) for much larger times, \( t \gg N^{-1} \). The corresponding eigenvalue density \( f_\sigma \) supported on \( \Sigma_N \) satisfies \( S_{\mu}(f_\sigma) \leq CN^{12} \) by (14.34). As a second cutoff, for some large \( K \) we set \( \hat{f}_{\sigma, K} := C_K \max\{f_\sigma, K\} \) with a normalization constant \( C_K \) such that \( \int \hat{f}_{\sigma, K} d\mu = 1 \) and \( C_K \to 1 \) as \( K \to \infty \). We will actually apply Theorem 14.1 to \( \hat{f}_{\sigma, K} \) as an initial condition. The regularized flow starts with this initial condition, i.e., we define \( f_{t, \delta} \) to be the solution of \( \partial_t f_{t, \delta} = \mathcal{L}_{\mu_\delta} f_{t, \delta} \) with \( f_0, \delta := \hat{f}_{\sigma, K} \).

Since the entropy decays along the regularized flow, we immediately have \( S_{\mu_\delta}(f_{t, \delta}) \leq S_{\mu}(f_{t, \delta}) = S_{\mu}(\hat{f}_{\sigma, K}) \). Since \( \hat{f}_{\sigma, K} \) is supported on \( \Sigma_N \) and it is bounded, clearly \( S_{\mu_\delta}(\hat{f}_{\sigma, K}) \) converges to \( S_{\mu}(\hat{f}_{\sigma, K}) \) as \( \delta \to 0 \). We then have

\[
S_{\mu}(\hat{f}_{\sigma, K}) = C_K \int \max\{f_\sigma, K\} \left[ \log \max\{f_\sigma, K\} + \log C_K \right] d\mu \to S_{\mu}(f_\sigma)
\]

as \( K \to \infty \) by monotone convergence. Thus first choosing \( K \) sufficiently large so that \( S_{\mu}(\hat{f}_{\sigma, K}) \leq 2S_{\mu}(f_\sigma) \leq CN^{12} \), then choosing \( \delta \) sufficiently small, we achieved \( S_{\mu_\delta}(f_{\tau, \delta}) \leq CN^{12} \).

This verifies the condition (14.38) for some sufficiently small \( \delta \), so that (14.39) holds. Letting \( \delta \to 0 \), we obtain (14.7) which completes the fully rigorous proof of Theorem 14.1.

### 14.4 Dirichlet form inequality for any \( \beta > 0 \)

Here we extend the proof of the key Dirichlet form inequality, Theorem 14.3, from \( \beta \geq 1 \) to any \( \beta > 0 \). This extension will not be needed for this book, so the reader may skip this section, but we remark that the Dirichlet form inequality for \( \beta > 0 \) played a crucial role in the proof of the universality for invariant ensembles for any \( \beta > 0 \), see [25].

**Lemma 14.7.** Let \( \beta > 0 \) and \( q \) be a probability density with respect to \( \omega \) defined in (12.12) with \( \nabla \sqrt{q} \in L^\infty(d\omega) \) and \( q \in L^\infty(d\omega) \). Then for any \( J \subset \{1, 2, \ldots, N - m - 1\} \) and any \( t > 0 \) we have

\[
\left| \int \frac{1}{|J|} \sum_{i \in J} G_{i,m} q \, d\omega - \int \frac{1}{|J|} \sum_{i \in J} G_{i,m} \, d\omega \right| \leq C \sqrt{\frac{D_{\omega}(\sqrt{q}) t}{|J|}} + C \sqrt{S_{\omega}(q)} e^{-ct/\tau}.
\]

For \( \beta \geq 1 \), the conditions \( \nabla \sqrt{q} \in L^\infty \) and \( q \in L^\infty(d\omega) \) can be removed.

We emphasize that this lemma holds for any \( \beta > 0 \), i.e., it does not (directly) rely on the existence of the DBM. The parameter \( t \) is not connected with the time parameter of a dynamics on \( \Sigma_N \) (although it emerges as a time cutoff in a regularized dynamics on \( \mathbb{R}^N \) within the proof).

**Proof of Lemma 14.7.** First we record the result if \( \omega \) on \( \Sigma_N \) is replaced with the regularized measure \( \omega_{\delta} \) on \( \mathbb{R}^N \), as defined in (13.81) with the choice of \( \hat{H}(x) = \sum_j U_j(x_j) \), where \( U_j \) is given in (14.10). In this case the proof of Theorem 14.3 applies to the letter even for \( \beta > 0 \) since now we work on \( \mathbb{R}^N \) and complications
arising from the boundary are absent. We immediately obtain for any probability density \( \tilde{q} \) on \( \mathbb{R}^N \) with \( \int \tilde{q} \, d\omega = 1 \), for any \( J \subset \{1, 2, \ldots, N - m_n - 1\} \) and any \( t > 0 \) that
\[
\left| \frac{1}{|J|} \sum_{i \in J} \mathcal{G}_{i,m} \tilde{q} \, d\omega - \frac{1}{|J|} \sum_{i \in J} \mathcal{G}_{i,m} \, d\omega \right| \leq C \sqrt{D_{\omega_i}(\sqrt{\tilde{q}}) t} + C \sqrt{S_{\omega_i}(\tilde{q})} e^{-ct/\tau}. \quad (14.45)
\]

Suppose now that \( \sqrt{\tilde{q}} \in H^1(\omega) \) and \( q \) is a bounded probability density in \( \Sigma_N \) with respect to \( \omega \). Similarly to the proof of (13.76) for the general \( \beta > 0 \) case, we may extend \( q \) to \( \hat{\Sigma} \) by symmetrization. Let \( \tilde{q} \) denote this extension which is bounded and \( \nabla \hat{f}^{1/2} \) is also bounded since \( q \) has these properties. Then there is a constant \( C_\delta \) such that \( q_\delta := C_\delta \tilde{q} \) is a probability density with respect to \( \omega_\delta \). Since \( \tilde{q} \) is bounded, we have \( \int_{\mathbb{R}^N} \tilde{q} \, d\omega_\delta \to \int_{\Sigma_N} q \, d\omega \) by dominated convergence, and thus \( C_\delta \to 1 \) as \( \delta \to 0 \).

Now we apply (14.45) \( \tilde{q} = q_\delta \). Taking the limit \( \delta \to 0 \), the left hand side converges to that of (14.44) since \( \mathcal{G}_{i,m} \) is a bounded smooth function and \( q_\delta \, d\omega_\delta = C_\delta \tilde{q} \, d\omega_\delta \) converges weakly to \( q(\omega) \mathbf{1}(\omega \in \Sigma_N) \, d\omega \) by dominated convergence. We thus have
\[
\left| \frac{1}{|J|} \sum_{i \in J} \mathcal{G}_{i,m} q \, d\omega - \frac{1}{|J|} \sum_{i \in J} \mathcal{G}_{i,m} \, d\omega \right| \leq C \limsup_{\delta \to 0} \sqrt{C_\delta D_{\omega_i}(\sqrt{\tilde{q}}) t} + C \limsup_{\delta \to 0} \sqrt{S_{\omega_i}(C_\delta \tilde{q})} e^{-ct/\tau}. \quad (14.46)
\]

By dominated convergence, using that \( \nabla \hat{f}^{1/2} \in L^\infty, \) \( D_{\omega}(\sqrt{q}) \in L^\infty \). Finally, these conditions can be removed for \( \beta > 1 \) by a simple approximation. We may assume \( D_{\omega}(\sqrt{q}) < \infty \); otherwise (14.44) is an empty statement. By the LSI (13.76) we also know that \( S_{\omega}(q) < \infty \). First, we still keep the assumption that \( q \in L^\infty \). Since \( C^\infty_0(\Sigma) \) is dense in \( H^1(\omega) \) (see Section 12.4), we can find a sequence of densities \( \tilde{q}_n \in L^\infty(\Sigma) \) such that \( \nabla \sqrt{\tilde{q}_n} \in L^2(\omega) \) and \( \sqrt{\tilde{q}_n} \to \sqrt{\tilde{q}} \) in \( L^2(\omega) \) and \( \nabla \sqrt{\tilde{q}_n} \to \nabla \sqrt{\tilde{q}} \) in \( L^2(\omega) \). In fact, the construction in Section 12.4 guarantees that, apart from an irrelevant smoothing, \( \tilde{q}_n \) may be chosen of the form \( \tilde{q}_n = C_n \phi_n \tilde{q} \), where \( \phi_n \) is a cutoff function with \( 0 \leq \phi_n \leq 1 \), converging to 1 pointwise and \( C_n \to 1 \). We know that (14.44) holds for every \( q_n \). Taking the limit \( n \to \infty \), the left hand side converges since
\[
\left| \int O(q_n - q) \, d\omega \right| \leq \|O\|_\infty \|\sqrt{\tilde{q}_n} - \sqrt{\tilde{q}}\|_{L^2(\omega)} \|\sqrt{\tilde{q}_n} + \sqrt{\tilde{q}}\|_{L^2(\omega)} \to 0,
\]
where we have used that \( \|\sqrt{\tilde{q}_n} + \sqrt{\tilde{q}}\|_2 \leq \|\sqrt{\tilde{q}_n}\|_2 + \|\sqrt{\tilde{q}}\|_2 \) and \( \|\sqrt{\tilde{q}}\|_2 = \int q \, d\omega = 1 \). Here \( O \) is given by (14.41). For the right hand side of (14.44) applied to the approximate function \( q_n \), we note that \( D_{\omega}(\sqrt{\tilde{q}_n}) \to D_{\omega}(\sqrt{q}) \) directly by the choice of \( q_n \). Finally, we need to show that \( S_{\omega}(q_n) \to S_{\omega}(q) \) as \( n \to \infty \). Clearly,
\[
S_{\omega}(q_n) - S_{\omega}(q) = \int_\Sigma [q_n \log q_n - q \log q] \, d\omega = \int_\Sigma [C_n \phi_n q \log(C_n \phi_n) + (C_n \phi_n - 1) q \log q] \, d\omega,
\]
and both integrals go to zero by dominated convergence and by \( S_{\omega}(q) < \infty \). This proves (14.44) for any \( q \in L^\infty \). Finally, this last condition can be removed by approximating any density \( q \) with \( q_M := C M \max(q, M) \), where \( C_M \) is the normalization and \( C_M \to 1 \) as \( M \to \infty \). Similarly to the proof in (13.84), one can show that \( S_{\omega}(q_M) \to S_{\omega}(q) \) and \( D_{\omega}(q_M) \to D_{\omega}(q) \). Thus we can pass to the limit in the inequality (14.44) written up for \( q_M \). This proves Lemma 14.7. Notice that the proof did not use the existence of the DBM; instead, it used the existence of the regularized DBM.

### 14.5 From gap distribution to correlation functions: proof of Theorem 12.4.

Theorem 12.4 follows from Theorem 14.1 if we can show that the correlation function difference in (12.23) can be bounded in terms of the difference of the expectation of gap observables in (14.8). We state it as the following lemma which clearly proves Theorem 12.4 with Theorem 14.1 as an input.

\[\text{122}\]
Lemma 14.8. Let \( f \) be a probability density with respect to the Gaussian equilibrium measure (12.13). Suppose that the estimate (12.22) holds with some exponent \( \xi > 0 \) and that
\[
\left| \int \frac{1}{|J|} \sum_{i \in J} G_{i,m}(x)(f d\mu - d\mu) \right| \leq C \frac{1}{\sqrt{|J|N^{3-\delta}}}
\] (14.47)
also holds for some \( \delta > 0 \) and for any \( J \subset \{1, 2, \ldots, N - m_n\} \), where \( G_{i,m} \) is defined in (14.6). Then for any \( \varepsilon > 0 \) and \( N \) large enough we have for any \( b = b_N \) with \( N^{-1} \ll b \ll 1 \) that
\[
\left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha O(\alpha) \left( P_{f,N}^{(n)} - P_{\mu,N}^{(n)} \right) \left( E' + \frac{\alpha}{N\theta_{ac}(E)} \right) \right| \leq N^{2\xi} \left[ N^{-1+\xi} + \sqrt{N^{-\delta}} \right].
\] (14.48)

This is a fairly standard technical argument, and the details will be given in Section 14.6. Here we just summarize the main points. To understand the difference between (14.47) and (14.48), consider \( n = 1 \) for simplicity and let \( m_1 = 1 \), say. The observable (14.7) answers the question: “What is the empirical distribution of differences of consecutive eigenvalues?” in other words (14.7) directly identifies the gap distribution. Correlation functions answer the question: “What is the probability that there are two eigenvalues at a fixed distance away from each other?” in other words they are not directly sensitive to the possible other eigenvalues in between. Of course these two questions are closely related and it is easy to deduce the answers from each other under some assumptions on the distribution of the eigenvalues. We now prove that the correlation functions can be derived from (generalized) gap distributions which is the content of Lemma 14.8.

### 14.6 Details of the proof of Lemma 14.8

By definition of the correlation function in (12.19), we have
\[
\int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha O(\alpha) P_{f,N}^{(n)} \left( E' + \frac{\alpha}{N} \right) = C_{N,n} \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha O(\alpha) \sum_{i_1 \neq i_2 \neq \ldots \neq i_n} O\left( N(x_{i_1} - E'), N(x_{i_1} - x_{i_2}), \ldots, N(x_{i_{n-1}} - x_{i_n}) \right) d\mu,
\]
\[
= C_{N,n} \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha O(\alpha) \sum_{i_1 \neq i_2 \neq \ldots \neq i_n} O\left( N(x_{i_1} - x_i), N(x_{i_2} - x_{i+m_2}), \ldots, N(x_{i_n} - x_{i+m_n}) \right) d\mu,
\] (14.49)

with \( C_{N,n} := N^n (N-n)!/N! = 1 + O_n(N^{-1}) \), where we let \( S_n \) denote the set of increasing positive integers, \( m = (m_2, m_3, \ldots, m_n) \in \mathbb{N}^{n-1} \), \( m_2 < m_3 < \ldots < m_n \), and we introduced
\[
Y_{i,m}(E', x) := O\left( N(x_i - E'), N(x_i - x_{i+m_2}), \ldots, N(x_i - x_{i+m_n}) \right).
\] (14.50)

We will set \( Y_{i,m} = 0 \) if \( i + m_n > N \). By permutational symmetry of \( P_{f,N}^{(n)} \), we can assume that \( O \) is symmetric and thus we restrict the summation to \( i_1 < i_2 < \ldots < i_n \) upon an overall factor \( n! \). Then we changed indices \( i = i_1, i_2 = i + m_2, i_3 = i + m_3, \ldots \), and performed a resummation over all index differences encoded in \( m \). Apart from the first variable \( N(x_i) \), the function \( Y_{i,m} \) is of the form (14.6), so (14.47) will apply. The dependence on the first variable will be negligible after the \( dE' \) integration on a macroscopic interval.

To control the error terms in this argument, especially to show that the error terms in the potentially infinite sum over \( m \in S_n \) converge, one needs an a priori bound on the local density. But this information is provided very precisely by the bulk rigidity estimate (12.22). The details for the rest of the argument will be presented in the next subsection.

Since (14.49) also holds if we set \( f = 1 \), to prove Lemma 14.8, we only need to estimate
\[
\Theta := \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} \sum_{i = 1}^{N} Y_{i,m}(E', x)(f - 1) d\mu \right|.
\] (15.41)
Let $M$ be an $N$-dependent parameter chosen at the end of the proof. Let 

$$S_u(M) := \{ m \in S_u, \ m_n \leq M \}, \quad S_u^c(M) := S_u \setminus S_u(M),$$

and note that $|S_u(M)| \leq M^{n-1}$. We have the simple bound $\Theta \leq \Theta^{(1)}_M + \Theta^{(2)}_M + \Theta^{(3)}_M$, where

$$\Theta^{(1)}_M := \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{i \in S_u(M)} \sum_{i=1}^{N} Y_{i,m}(E',x)(f-1)d\mu \right|$$

and

$$\Theta^{(2)}_M := \sum_{m \in S_u^c(M)} \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{i \in S_u^c(M)} \sum_{i=1}^{N} Y_{i,m}(E',x)f\,d\mu \right|.$$  

We define $\Theta^{(3)}_M$ to be the same as $\Theta^{(2)}_M$ but with $f$ replaced by the constant 1, i.e., the equilibrium measure $\mu$.

**Step 1: Small $m$ case; estimate of $\Theta^{(1)}_M$.**

After performing the $dE'$ integration, we will eventually apply Theorem 14.1 to the function

$$G(u_1,u_2,\ldots) := \int_{\mathbb{R}} O(y,u_1,u_2,\ldots)\,dy,$$

i.e., to the quantity

$$\int_{\mathbb{R}} dE' Y_{i,m}(E',x) = \frac{1}{N} g\left(N(x_i - x_{i+m_2}), \ldots \right)$$

for each fixed $i$ and $m$.

For any $E$ and $0 < \zeta < b$ define sets of integers $J = J_{E,b,\zeta}$ and $J^\pm = J_{E,b,\zeta}^\pm$ by

$$J := \{ i : \gamma_i \in [E - b, E + b] \}, \quad J^\pm := \{ i : \gamma_i \in [E - (b \pm \zeta), E + b \pm \zeta] \},$$

where $\gamma_i$ was defined in (11.31). Clearly, $J^- \subset J \subset J^+$. With these notations, we have

$$\int_{E-b}^{E+b} dE' \sum_{i=1}^{N} Y_{i,m}(E',x) = \int_{E-b}^{E+b} dE' \sum_{i \in J^+} Y_{i,m}(E',x) + \Omega_{J,m}(x).$$

The error term $\Omega_{J,m}^+$, defined by (14.55) indirectly, comes from those $i \notin J^+$ indices, for which $x_i \in [E - b, E + b] + O(1/N)$ since $Y_{i,m}(E',x) = 0$ unless $|x_i - E| \leq C/N$, the constant depending on the support of $O$. Thus

$$|\Omega_{J,m}^+(x)| \leq CN^{-1} \# \{ i : |x_i - \gamma_i| \geq \zeta/2, |x_i - E| \leq 2b \}$$

for any sufficiently large $N$ assuming $\zeta \gg 1/N$ and using that $O$ is a bounded function. The additional $N^{-1}$ factor comes from the $dE'$ integration. Due to the rigidity estimate (12.22) and choosing $\zeta = N^{-1} + \varepsilon'$ with some $\varepsilon' > 0$, we get

$$\int |\Omega_{J,m}^+(x)| f\,d\mu \leq N^{-D}$$

for any $D$. We can also estimate

$$\int_{E-b}^{E+b} dE' \sum_{i \in J^+} Y_{i,m}(E',x) \leq \int_{E-b}^{E+b} dE' \sum_{i \in J^-} Y_{i,m}(E',x) + CN^{-1} |J^+ \setminus J^-|$$

$$= \int_{\mathbb{R}} dE' \sum_{i \in J^-} Y_{i,m}(E',x) + CN^{-1} |J^+ \setminus J^-| + \Xi_{J,m}(x)$$

$$\leq \int_{\mathbb{R}} dE' \sum_{i \in J} Y_{i,m}(E',x) + CN^{-1} |J^+ \setminus J^-| + CN^{-1} |J \setminus J^-| + \Xi_{J,m}(x),$$

124
where the error term $\Xi_{j,m}^+$, defined by (14.57), comes from indices $i \in J^-$ such that $x_i \notin [E - b, E + b] + O(1/N)$. It satisfies the same bound (14.56) as $\Omega_{j,m}^+$. By the continuity of $g$, the density of $\gamma_i$’s is bounded by $CN$, thus $|J^+ \setminus J^-| \leq CN\zeta$ and $|J \setminus J^-| \leq CN\zeta$. Therefore, summing up the formula (14.54) for $i \in J$, we obtain from (14.55), (14.56) and (14.57)

$$
\int_{E-b}^{E+b} dE' \int_1^N \sum_{i=1}^N Y_{i,m}(E', x) f d\mu \leq \int_1^N \sum_{i \in J} G \left(N(x_i - x_i + m_2), \ldots \right) f d\mu + C\zeta + CN^{-D}
$$

for each $m \in S_n$. A similar lower bound can be obtained analogously, and we get

$$
\left| \int_{E-b}^{E+b} dE' \int_1^N \sum_{i=1}^N Y_{i,m}(E', x) f d\mu - \int_1^N \sum_{i \in J} G \left(N(x_i - x_i + m_2), \ldots \right) f d\mu \right| \leq C\zeta + CN^{-D}
$$

(14.58) for each $m \in S_n$. The error term $N^{-D}$ can be neglected. Adding up (14.58) for all $m \in S_n(M)$, we get

$$
\left| \int_{E-b}^{E+b} dE' \int \sum_{m \in S_n(M)} \sum_{i=1}^N Y_{i,m}(E', x) f d\mu - \int \sum_{m \in S_n(M)} \int \sum_{i \in J} G \left(N(x_i - x_i + m_2), \ldots \right) f d\mu \right| \leq CM^{n-1} \zeta.
$$

(14.59)

Clearly, the same estimate holds for the equilibrium, i.e., if we set $f = 1$ in (14.59). Subtracting these two formulas and applying (14.47) to each summand on the second term in (14.58), we conclude that

$$
\Theta^{(1)}_M = \int_{E-b}^{E+b} dE' \int \sum_{m \in S_n(M)} \sum_{i=1}^N Y_{i,m}(E', x) f d\mu - \int \sum_{m \in S_n(M)} \int \sum_{i \in J} G \left(N(x_i - x_i + m_2), \ldots \right) f d\mu \leq CM^{-1} \left(b^{-1} N^{-1+\xi+\varepsilon'} + b^{-1/2} N^{-\delta/2} \right),
$$

(14.60)

where we have used that $|J| \leq CNb$. This completes the estimate of $\Theta^{(1)}_M$.

**Step 2. Large $m$ case; estimate of $\Theta^{(2)}_M$ and $\Theta^{(3)}_M$.**

For a fixed $y \in \mathbb{R}$, $\ell > 0$, let

$$
\chi(y, \ell) := \sum_{i=1}^N \mathbf{1} \left( y - \frac{\ell}{N}, y + \frac{\ell}{N} \right)
$$

denote the number of points in the interval $[y - \ell/N, y + \ell/N]$. Note that for a fixed $m = (m_2, \ldots, m_n)$, we have

$$
\sum_{i=1}^N |Y_{i,m}(E', x)| \leq C \cdot \chi(E', \ell) \cdot \mathbf{1} \left( \chi(E', \ell) \geq m_n \right) \leq C \sum_{m=m_n}^N m \cdot \mathbf{1} \left( \chi(E', \ell) \geq m \right),
$$

(14.61)

where $\ell$ denotes the maximum of $|u_1| + \ldots + |u_n|$ in the support of $O(u_1, \ldots, u_n)$.

Since the summation over all increasing sequences $m = (m_2, \ldots, m_n) \in \mathbb{N}_+^{n-1}$ with a fixed $m_n$ contains at most $m_{n-2}$ terms, we have

$$
\sum_{m \in S_n(M)} \left| \int_{E-b}^{E+b} dE' \int \sum_{i=1}^N Y_{i,m}(E', x) f d\mu \right| \leq C \sum_{m \in S_n(M)} \int_{E-b}^{E+b} dE' \sum_{m=M}^N m^{n-1} \int \mathbf{1} \left( \chi(E', \ell) \geq m \right) f d\mu.
$$

(14.62)

The rigidity bound (12.22) clearly implies

$$
\int \mathbf{1} \left( \chi(E', \ell) \geq m \right) f d\mu \leq C_D \varepsilon' N^{-D}
$$

125
for any $D > 0$ and $\varepsilon' > 0$, as long as $m \geq \ell N^{\varepsilon'}$. Therefore, choosing $D = 2n + 2$, we see from (14.62) that $\Theta^{(2)}$ is negligible, e.g.

$$\Theta_M^{(2)} \leq CN^{-1}$$

if $M \geq \ell N^{\varepsilon'}$. Since all rigidity bounds hold for the equilibrium measure, the last bound is valid for $\Theta_M^{(3)}$ as well. We will choose $M = \ell N^{\varepsilon'}$ and recall that $\ell$ is independent of $N$. Combining these bounds with (14.60), we obtain

$$\Theta \leq CN^{\varepsilon'(n+1)} \left( b^{-1} N^{-1+\xi+\varepsilon'} + b^{-1/2} N^{-\delta/2} \right) + CN^{-1}.$$ 

Choosing $\varepsilon' = \varepsilon/(n + 1)$, we conclude the proof of Lemma 14.8. \qed
15 Continuity of local correlation functions under matrix OU process

We have completed the first two steps of the three step strategy introduced in Section 5, i.e., the local semicircle law and the universality of Gaussian divisible ensembles (Theorem 12.4). In this section, we will complete this strategy by proving a continuity result for the local correlation functions of the matrix OU process in the following Theorem 15.2 and Lemma 15.3. This is the Step 3a defined in Section 5. From these results, we obtain a weaker version of Theorem 5.1, namely we get averaged energy universality of Wigner matrices but only on scale $b \geq N^{-1/2+\varepsilon}$. In Section 16.1, we will use the idea of “approximation by a Gaussian divisible ensemble” and prove Theorem 5.1 down to any scale $b \geq N^{-1+\varepsilon}$.

Theorem 15.1. [68, Theorem 2.2] Let $H$ be an $N \times N$ real symmetric or complex Hermitian Wigner matrix. In the Hermitian case we assume that the real and imaginary parts are i.i.d. Suppose that the distribution $\nu$ of the rescaled matrix elements $\sqrt{N}h_{ij}$ satisfies the decay condition (5.6). Fix a small $\varepsilon > 0$, an integer $n \geq 1$ and let $O : \mathbb{R}^n \to \mathbb{R}$ be a continuous, compactly supported function. Then for any $|E| < 2$ and $b \in [N^{-1/2+\varepsilon}, N^{-\varepsilon}]$, we have

$$\lim_{N \to \infty} \frac{1}{2b} \int_{E-b}^{E+b} dE' \int_{\mathbb{R}^n} d\alpha \ O(\alpha) \left( \hat{p}_{H,N}^{(n)} - \hat{p}_{G,N}^{(n)} \right) \left( E' + \frac{\alpha}{N} \right) = 0. \quad (15.1)$$

To prove this theorem, we first recall the matrix OU process (12.1) defined by

$$dH_t = \frac{1}{\sqrt{N}} dB_t - \frac{1}{2} H_t dt, \quad (15.2)$$

with the initial data $H_0$. The eigenvalue evolution of this process is the DBM and recall that we denote the eigenvalue distribution at the time $t$ by $f_t dp$ with $f_t$ satisfying (12.17). In this section, we assume that the initial data $H_0$ is a $N \times N$ Wigner matrix and the distribution of matrix element satisfies the uniform polynomial decay condition (5.6). We have the following Green function continuity theorem for the matrix OU process.

Theorem 15.2 (Continuity of Green function). Suppose that the initial data $H_0$ is a $N \times N$ Wigner matrix with the distribution of matrix element satisfying the uniform polynomial decay condition (5.6). Let $\kappa > 0$ be arbitrary and suppose that for some small parameter $\sigma > 0$ and for any $y \geq N^{-1+\sigma}$, we have the following estimate on the diagonal elements of the resolvent for any $0 \leq t \leq 1$:

$$\max_{1 \leq k \leq N} \max_{|E| \leq 2-\kappa} \left| \frac{1}{H_t - E - i\eta} \right|_{kk} \lesssim N^{2\sigma} \quad (15.3)$$

with some constants $C, c$ depending only on $\sigma, \kappa$.

Let $G(t,z) = (H_t - z)^{-1}$ denote the resolvent and $m(t,z) = N^{-1} \text{Tr} G(t,z)$. Suppose that $F(x_1, \ldots, x_n)$ is a function such that for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $1 \leq |\alpha| \leq 5$ and for any $\varepsilon' > 0$ sufficiently small, we have

$$\max \left\{ |\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \leq N^{\varepsilon'} \right\} \lesssim N^{C_{0\varepsilon'}} \quad (15.4)$$

and

$$\max \left\{ |\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \leq N^2 \right\} \lesssim N^{C_0} \quad (15.5)$$

for some constant $C_0$.

Let $\varepsilon > 0$ be arbitrary and choose an $\eta$ with $N^{-1-\varepsilon} \leq \eta \leq N^{-1}$. For any sequence of complex parameters $z_j = E_j \pm i\eta, j = 1, \ldots, n$, with $|E_j| \leq 2 - \kappa$, there is a constant $C$, such that for any choices of the signs in the imaginary part of $z_j$ and for any $t \in [0, 1]$, we have

$$|E F(m(t,z_1), \ldots, m(t,z_n)) - E F(m(0,z_1), \ldots, m(0,z_n))| \lesssim C N^{2\sigma + 6\varepsilon} \sqrt{N}. \quad (15.6)$$
We defer the proof of Theorem 15.2 to the next subsection. The following result shows how the Green function continuity, i.e., estimate of the type (15.6), can be used to compare correlation functions. The proof will be given in Section 15.2.

Next we will compare local statistics of two Wigner matrix ensembles. We will use the labels $v$ and $w$ to distinguish them because later in Section 16 we will denote the matrix elements of the two ensembles by different letters, $v_{ij}$ and $w_{ij}$. For any two (generalized) Wigner matrix ensembles $H^v$ and $H^w$, we denote the probability laws of their eigenvalues $\lambda^v$ and $\lambda^w$ by $\mu_v$ and $\mu_w$, respectively. Denote by $m^v(z)$, $m^w(z)$ the Stieltjes transforms of the eigenvalues, i.e.,

$$m^v(z) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_j^v - z},$$

and similarly for $m^w(z)$. Let $p_{v,N}^{(n)}$ and $p_{w,N}^{(n)}$ be the $n$-point correlation functions of the eigenvalues w.r.t. $\mu_v$ and $\mu_w$. We remind the reader that sometimes we will use the notation $EYF(\lambda)$ to denote the expectation of $EYF(\lambda^v)$. Although the setup is motivated by local correlation functions of eigenvalues, we point out that the concept of matrices or eigenvalues plays no role in the following theorem. It is purely a statement about the different letters, to distinguish them because later in Section 16 we will denote the matrix elements of the two ensembles by different letters.

We defer the proof of Theorem 15.2 to the next subsection. The following result shows how the Green function continuity, i.e., estimate of the type (15.6), can be used to compare correlation functions. The proof will be given in Section 15.2.

Next we will compare local statistics of two Wigner matrix ensembles. We will use the labels $v$ and $w$ to distinguish them because later in Section 16 we will denote the matrix elements of the two ensembles by different letters, $v_{ij}$ and $w_{ij}$. For any two (generalized) Wigner matrix ensembles $H^v$ and $H^w$, we denote the probability laws of their eigenvalues $\lambda^v$ and $\lambda^w$ by $\mu_v$ and $\mu_w$, respectively. Denote by $m^v(z)$, $m^w(z)$ the Stieltjes transforms of the eigenvalues, i.e.,

$$m^v(z) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_j^v - z},$$

and similarly for $m^w(z)$. Let $p_{v,N}^{(n)}$ and $p_{w,N}^{(n)}$ be the $n$-point correlation functions of the eigenvalues w.r.t. $\mu_v$ and $\mu_w$. We remind the reader that sometimes we will use the notation $EYF(\lambda)$ to denote the expectation of $EYF(\lambda^v)$. Although the setup is motivated by local correlation functions of eigenvalues, we point out that the concept of matrices or eigenvalues plays no role in the following theorem. It is purely a statement about the different letters, to distinguish them because later in Section 16 we will denote the matrix elements of the two ensembles by different letters.

Theorem 15.3 (Correlation function comparison). Let $\kappa > 0$ be arbitrary and suppose that for some small parameters $\sigma, \delta > 0$ the following two conditions hold.

(i) For any $\varepsilon > 0$ and any $k$ integer

$$E|\text{Im } m^v(E + iN^{-1+\varepsilon})|^k + E|\text{Im } m^w(E + iN^{-1+\varepsilon})|^k \leq C$$

holds for any $|E| \leq 2 - \kappa$ and $N \geq N_0(\varepsilon, k, \kappa)$;

(ii) For any sequence $z_j = E_j + i\eta_j$, $j = 1, \ldots, n$ with $|E_j| \leq 2 - \kappa$ and $\eta_j = N^{-1-\sigma_j}$ for some $\sigma_j \leq \sigma$, we have

$$|E(\text{Im } m^v(z_1) \cdots \text{Im } m^v(z_n)) - E(\text{Im } m^w(z_1) \cdots \text{Im } m^w(z_n))| \leq N^{-\delta}. \quad (15.8)$$

Then for any integer $n \geq 1$ there are positive constants $c_n = c_n(\sigma, \delta)$ such that for any $|E| \leq 2 - 2\kappa$ and for any $C^1$ function $O : \mathbb{R}^n \to \mathbb{R}$ with compact support,

$$\int_{\mathbb{R}^n} d\alpha O(\alpha) \left( p_{v,N}^{(n)} - p_{w,N}^{(n)} \right)(E + \frac{\alpha}{N}) \leq CN^{-c_n}, \quad (15.9)$$

where $C$ depends on $O$ and $N$ is sufficiently large.

We remark that in some applications we will use slightly different conditions. Instead of (15.8) we may assume

$$\left| E\left( \text{Im } m^v(z_1), \ldots, \text{Im } m^v(z_n) \right) - E\left( \text{Im } m^w(z_1), \ldots, \text{Im } m^w(z_n) \right) \right| \leq N^{-\delta}, \quad (15.10)$$

where $F$ is as in Theorem 15.2. Then (15.8) holds since we can approximate $E\left[ \text{Im } m^w(z_1) \cdots \text{Im } m^w(z_n) \right]$ by the expression in (15.10) where the function $F$ is chosen to be $F(x_1, \ldots, x_n) := x_1x_2\ldots x_n$ if $\max_j |x_j| \leq N^c$ and it is smoothly cutoff to go to zero in the regime $\max_j |x_j| \geq N^{2c}$ for some small $c > 0$.

To see this, recall the elementary inequality,

$$\theta_{\eta_1} \leq \frac{\eta_2}{\eta_1} \theta_{\eta_2}, \quad \text{implying} \quad \text{Im } m(E + i\eta_1) \leq \frac{\eta_2}{\eta_1} \text{Im } m(E + i\eta_2) \quad (15.11)$$
Lemma 15.4. Suppose that (elements $H$ with some $0 \theta$)

The first ingredient to prove Theorem 15.2 is the following continuity estimate of matrix OU process. To prove the theorem, we need to adjust our approach to this case.

This compares the correlation functions of $H_t$ (12.24), we have for any $\eta$, with any sufficiently small $\eta$, where the last supremum runs through all deformations $H_t$ of eigenvalues and eigenvectors (12.9), (12.10), higher derivatives of eigenvalues involve singularities of the form $(\lambda_i - \lambda_j)^{-\alpha}$ for some positive integers $n$. These singularities are very difficult to control precisely.

Proof of Theorem 15.1. Recall the matrix OU (15.2) can be solved by the formula (12.21) so that the probability distribution of the matrix OU is given by a Wigner matrix ensemble if the initial data is a Wigner matrix ensemble. More precisely, if we denote the initial Wigner matrix by $H_0$ then the distribution of $H_t$ is the same as $e^{-t/2}H_0 + (1 - e^{-t})^{1/2}H_G$. Hence the rigidity holds in this case by (11.32) and (15.3) holds with any sufficiently small $\sigma > 0$: we may choose $\sigma = \varepsilon_1$. Recall $p_{t, N}$ denotes the correlation functions of $H_t$. We now apply Theorem 15.3 with $H^x = H_0$ and $H^w = H_t$. The assumption (15.8) can be verified by (15.6) if $t \leq N^{-1/2-\varepsilon}$ for any $\varepsilon > 0$. From (15.9), we have

This compares the correlation functions of $H_0$ and $H_t$ if $t$ is not too large. To compare $H_t$ with $H_\infty = H_G$, by (12.24), we have for $t = N^{-1/2-\varepsilon}$ and $b \geq N^{-1/2+10\varepsilon}$ that (15.13) holds with $p_{t, N}$ (which is the correlation functions of $H_0$) replaced by $p_{t, N}^{(n)}$. We have thus completed the proof of Theorem 15.1.

15.1 Proof of Theorem 15.2

The first ingredient to prove Theorem 15.2 is the following continuity estimate of matrix OU process. To state it, we need to introduce the deformation of a matrix $H$ at certain matrix elements. For any $i < j$, let $\theta^{ij}H$ denote a new matrix with matrix elements $(\theta^{ij}H)_{k\ell} = H_{k\ell}$ if $\{k, \ell\} \neq \{i, j\}$ and $(\theta^{ij}H)_{k\ell} = \theta_{k\ell}^{ij}H_{k\ell}$ with some $0 \leq \theta_{k\ell}^{ij} \leq 1$ if $\{k, \ell\} = \{i, j\}$. In other words, the deformation operation $\theta^{ij}$ multiplies the matrix elements $H_{ij}$ and $H_{ji}$ by a factor between 0 and 1 and leaves all other matrix elements intact.

**Lemma 15.4.** Suppose that $H_0$ is a Wigner ensemble and $t \in [0, 1]$. Let $g$ be a smooth function of the matrix elements $(h_{ij})_{i \leq j}$ and set

$$M_t := \sup_{0 \leq s \leq t} \sup_{i \leq j} \sup_{\theta^{ij}} \mathbb{E}\left((N^{3/2}|h_{ij}(s)|^3 + \sqrt{N}|h_{ij}(s)|)|\partial_{h_{ij}}^3 g(\theta^{ij}H_s)|\right),$$

where the last supremum runs through all deformations $\theta^{ij}$. Then

$$\mathbb{E}g(H_t) - \mathbb{E}g(H_0) = O(t\sqrt{N})M_t.$$

This lemma holds also for generalized Wigner matrices. We refer the reader to [12] for the minor adjustment needed to this case.
Define a dyadic decomposition \( \eta \) obtained by establishing the estimate first for a fine grid of matrix elements and for spectral parameters with imaginary parts slightly below a Taylor expansion of the first derivative \( \partial_{ij}g \) in the direction \( h_{ij} \) yields

\[
E(h_{ij}(t)\partial_{ij}g(H_t)) = Eh_{ij}(t)\partial_{ij}g_{h_{ij}(t)=0} + E(h_{ij}(t)^2\partial_{ij}^2g_{h_{ij}(t)=0}) + O\left(\sup_{\theta^{ij}} E(|h_{ij}(t)^3\partial_{ij}^3g(\theta^{ij}H_t)|)\right)
\]

Integration over time finishes the proof.

The second ingredient to prove Theorem 15.2 is an estimate of the type (15.3) but not only for diagonal matrix elements and for spectral parameters with imaginary parts slightly below \( N^{-1} \).

**Lemma 15.5.** Suppose for a Wigner matrix \( H \) we have the following estimate

\[
\max_{1 \leq k \leq N} \sup_{|E| \leq 2^{-\kappa} \eta} \sup_{\eta \geq N^{-1-\varepsilon}} \left| \operatorname{Im} \left( \frac{1}{H - E \pm i\eta} \right) \right|_{kk} \prec N^{3\sigma+\varepsilon}. \tag{15.15}
\]

Then we have that for any \( \eta \geq N^{-1-\varepsilon} \)

\[
\sup_{1 \leq k, \ell \leq N} \sup_{|E| \leq 2^{-\kappa}} \left| \frac{1}{H - E \pm i\eta} \right|_{k\ell} \prec N^{3\sigma+\varepsilon}. \tag{15.16}
\]

We remark that (15.16) could be strengthened by including the supremum over \( \eta \geq N^{-1-\varepsilon} \). This can be obtained by establishing the estimate first for a fine grid of \( \eta \)'s with spacing \( N^{-10} \) and then extend the bound for all \( \eta \) by using that the Green functions are Lipschitz continuous in \( \eta \) with a Lipschitz constant \( \eta^{-2} \).

**Proof.** Let \( \lambda_m \) and \( u_m \) denote the eigenvalues and eigenvectors of \( H \), then by the definition of the Green function, we have

\[
\left| \left( \frac{1}{H - z} \right)_{jk} \right| \leq \sum_{m=1}^{N} \frac{|u_m(j)||u_m(k)|}{|\lambda_m - z|} \leq \left[ \sum_{m=1}^{N} \frac{|u_m(j)|^2}{|\lambda_m - z|} \right]^{1/2} \left[ \sum_{m=1}^{N} \frac{|u_m(k)|^2}{|\lambda_m - z|} \right]^{1/2}. \tag{15.17}
\]

Define a dyadic decomposition

\[
U_n = \{ m : 2^{n-1}\eta \leq |\lambda_m - E| < 2^n\eta \}, \quad n = 1, 2, \ldots, n_0 := C \log N, \tag{15.18}
\]

\[
U_0 = \{ m : |\lambda_m - E| < \eta \}, \quad U_{\infty} := \{ m : 2^n\eta \leq |\lambda_m - E| \}.
\]
and divide the summation over \( m \) into \( \cup_n U_n \)

\[
\sum_{m=1}^{\infty} \left| \frac{u_m(j)}{\lambda_m - z} \right|^2 = \sum_n \sum_{m \in U_n} \left| \frac{u_m(j)}{\lambda_m - z} \right|^2 \leq C \sum_n \sum_{m \in U_n} \text{Im} \left( \frac{|u_m(j)|^2}{\lambda_m - E - i2^n \eta} \right) \leq C \sum_n \text{Im} \left( \frac{1}{H - E - i2^n \eta} \right)_{jj}.
\]

Now using (15.15) we can control the right hand side of (15.17) and conclude (15.16).

Now we can finish the proof of Theorem 15.2. First note that from the trivial bound

\[
\text{Im} \left( \frac{1}{H - E - i\eta} \right)_{jj} \leq \frac{y}{\eta} \text{Im} \left( \frac{1}{H - E - i\eta} \right)_{jj}, \quad \eta \leq y, \tag{15.19}
\]

and (15.3), the assumption (15.15) in Lemma 15.5 holds. Therefore the bounds (15.16) on the matrix elements are available.

Next, we will first consider the specific function

\[
g(H) = G_{ab}(z), \tag{15.20}
\]

for a fixed index pair \( a, b \), where \( z = E + i\eta \) with \( N^{-1-\varepsilon} \leq \eta \leq 1 \) and \( |E| \leq 2 - \kappa \). While this function is not of the form appearing in (15.6), once the comparison for \( G_{ab}(z) \) is established, we take \( a = b \) and average over \( a \) to get the normalized trace of \( G \), i.e., the Stieltjes transform. This will then prove (15.6) for the function \( F(x) = x \), i.e., compare \( \mathbb{E}m(t, z) \) with \( \mathbb{E}m(0, z) \). The case of a polynomial \( F \) with several arguments is analogous, and similarly any function satisfying (15.4)–(15.5) can be sufficiently well approximated by a Taylor expansion.

Returning to the case (15.20), in order to apply Lemma 15.4 we need to bound the third derivatives of \( g(H) \) to estimate \( M_t \) from (15.14). We have

\[
\partial^3_{ij} G(z)_{ab} = -\sum_{\alpha, \beta} G(z)_{\alpha\alpha_1} G(z)_{\beta_1, \alpha_2} G(z)_{\beta_2, \alpha_3} G(z)_{\beta_3, \beta},
\]

where \( \{\alpha_k, \beta_k\} = \{i, j\} \) or \( \{j, i\} \). By (15.16), the following four expressions

\[
G(z)_{\alpha_\alpha_1}, \ G(z)_{\beta_1, \alpha_2}, \ G(z)_{\beta_2, \alpha_3}, \ G(z)_{\beta_3, \beta}
\]

are bounded by \( N^{4\sigma + \varepsilon} \) with very high probability provided \( N^{-1-\varepsilon} \leq \eta \leq 1 \). Consequently, we proved that uniformly in \( E \in (-2 + \kappa, 2 - \kappa) \), \( N^{-1-\varepsilon} \leq \eta \leq 1 \),

\[
\partial^3_{ij} G(z)_{ab} = O(N^{2\sigma + 5\varepsilon})
\]

with very high probability. The same argument holds if some matrix elements are reduced by deformation, i.e., we have

\[
|\partial^3_{ij} g(\theta^t H_s)| = O(N^{2\sigma + 5\varepsilon})
\]

and thus (15.14) holds with \( M_t = N^{2\sigma + 6\varepsilon} \) using the bound \( h_{ij} \prec N^{-1/2} \). Hence by Lemma 15.4, we have proved that for any \( t \leq 1 \) we have

\[
|\mathbb{E}g(H_t) - \mathbb{E}g(H_0)| \leq CN^{2\sigma + 6\varepsilon}(t\sqrt{N}).
\]

This completes the proof of Theorem 15.2.

### 15.2 Proof of the correlation function comparison Theorem 15.3

Define an approximate delta function at the scale \( \eta \) by

\[
\theta_\eta(x) := \frac{1}{\pi} \text{Im} \frac{1}{x - i\eta}.
\]
We will choose $\eta \sim N^{-1-a}$, with some small $a > 0$, i.e., slightly smaller than the typical eigenvalue spacing. This means that an observable of the form $\theta_\eta$ have sufficient resolution to detect individual eigenvalues. Moreover, polynomials of such observables detect correlation functions. On the other hand,

$$\frac{1}{N} \sum_i \theta_\eta(\lambda_i - E) = \pi \text{Im } m(z),$$

therefore expectation values of such observables are covered by the condition (15.8). The rest of the proof consists of making this idea precise. There are two technicalities to resolve. First, correlation functions involve distinct eigenvalues (see (15.25) below), while polynomials of the resolvent include an overcounting of coinciding eigenvalues. Thus an exclusion-inclusion formula will be needed. Second, although $\eta$ is much smaller than the relevant scale $1/N$, it still does not give pointwise information on the correlation functions. However, the correlation functions in (4.38) are identified only as a weak limit, i.e., tested against a continuous function $O$. The continuity of $O$ can be used to show that the difference between the exact correlation functions and the smeared out ones on the scale $\eta \sim N^{-1-a}$ is negligible. This last step requires an a priori upper bound on the density to ensure that not too many eigenvalues fall into an irrelevantly small interval; this bound is given in (15.7), and it will eventually be verified by the local semicircle law.

Proof of Theorem 15.3. For notational simplicity, we give the detailed proof only for the case of $n = 3$-point correlation functions; the proof is analogous for the general case.

Denote by $\eta = N^{-1-a}$ for some small $a > 0$; by following the proof one may check that any $a \leq c \min \{\delta, \sigma\}/n^2$ will do. Let $O$ be a compactly supported test function and let

$$O_\eta(\beta_1, \beta_2, \beta_3) := \frac{1}{N^3} \int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) \theta_\eta \left( \frac{\beta_1 - \alpha_1}{N} \right) \ldots \theta_\eta \left( \frac{\beta_3 - \alpha_3}{N} \right)$$

be its smoothing on scale $N\eta$. We note that for any nonnegative $C^1$ function $O$ with compact support, there is a constant depending on $O$ such that

$$O_\eta \leq CO_{\eta'}, \quad \text{for any } 0 \leq \eta \leq \eta' \leq 1/N. \quad (15.22)$$

We can apply this bound with $\eta' = 1/N$ and combine it with (15.11) with $\eta_1 = 1/N$ and $\eta_2 = N^{-1+\epsilon}$ for any small $\epsilon > 0$ to have

$$O_\eta \leq C N^{3\epsilon} O_{N^{-1+\epsilon}}, \quad \text{for any } 0 \leq \eta \leq 1/N. \quad (15.23)$$

After the change of variables $x_j = E + \beta_j/N$ and with $E_j := E + \frac{\alpha_j}{N}$, we have

$$\int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 \, O_\eta(\beta_1, \beta_2, \beta_3) \delta_{w,N}^{(3)} \left( E + \frac{\beta_1}{N}, \ldots, E + \frac{\beta_3}{N} \right) \quad (15.24)$$

$$= \int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 \, O(\alpha_1, \alpha_2, \alpha_3) \int_{\mathbb{R}^3} dx_1 dx_2 dx_3 \delta_{w,N}^{(3)}(x_1, x_2, x_3) \theta_\eta(x_1 - E_1) \theta_\eta(x_2 - E_2) \theta_\eta(x_3 - E_3).$$

By definition of the correlation function, for any fixed $E$, $\alpha_1, \alpha_2, \alpha_3$,

$$\int dx_1 dx_2 dx_3 \delta_{w,N}^{(3)}(x_1, x_2, x_3) \theta_\eta(x_1 - E_1) \theta_\eta(x_2 - E_2) \theta_\eta(x_3 - E_3) \quad (15.25)$$

$$= \text{E}^w \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k} \theta_\eta(\lambda_i - E - \frac{\alpha_1}{N}) \theta_\eta(\lambda_j - E - \frac{\alpha_2}{N}) \theta_\eta(\lambda_k - E - \frac{\alpha_3}{N}),$$

where $\text{E}^w$ indicates expectation w.r.t. the $w$ variables. By the exclusion-inclusion principle,

$$\text{E}^w \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k} \theta_\eta(x_1 - E_1) \theta_\eta(x_2 - E_2) \theta_\eta(x_3 - E_3) = \text{E}^w A_1 + \text{E}^w A_2 + \text{E}^w A_3, \quad (15.26)$$

132
where

\[
A_1 := \frac{N^3}{N(N-1)(N-2)} \prod_{j=1}^{3} \left[ \frac{1}{N} \sum_{i} \theta_{\eta}(\lambda_i - E_j) \right],
\]

\[
A_3 := \frac{2}{N(N-1)(N-2)} \sum_{i} \theta_{\eta}(\lambda_i - E_1)\theta_{\eta}(\lambda_i - E_2)\theta_{\eta}(\lambda_i - E_3),
\]

and \( A_2 := B_1 + B_2 + B_3 \) with \( B_j \) defined as follows:

\[
B_3 = -\frac{1}{N(N-1)(N-2)} \sum_{i} \theta_{\eta}(\lambda_i - E_1)\theta_{\eta}(\lambda_i - E_2)\sum_{k} \theta_{\eta}(\lambda_k - E_3);
\]

\( B_1 \) consists of terms with \( j = k \), and \( B_2 \) consists of terms with \( i = k \) from the triple sum (15.26). We remark that \( A_3 + A_2 \leq 0 \) and thus

\[
\int dx_1 dx_2 dx_3 p^{(3)}_{w,N}(x_1, x_2, x_3)\theta(x_1 - E_1)\theta(x_2 - E_2)\theta(x_3 - E_3) \leq C\mathbb{E}^{w} \prod_{j=1}^{3} \text{Im}(E_j + i\eta). \tag{15.27}
\]

The same bound holds for \( p^{(3)}_{v,N} \) as well. Therefore, we can combine (15.23) and (15.7) to obtain for any small \( \varepsilon > 0 \) we have

\[
\int d\beta_1 d\beta_2 d\beta_3 O_{\eta}(\beta_1, \beta_2, \beta_3)(p^{(3)}_{w,N} + p^{(3)}_{v,N}) \left( E + \frac{\beta_1}{N}, \ldots, E + \frac{\beta_4}{N} \right) = O(N^{3\varepsilon}). \tag{15.28}
\]

To approximate \( \mathbb{E}^{w} B_3 \), we define \( \phi_{E_1, E_2}(x) = \theta(x - E_1)\theta(x - E_2) \). Recall \( \eta = N^{-1-a} \) and let \( \tilde{\eta} = N^{-1-a} \). Decompose \( \tilde{\theta} = \tilde{\theta}_1 + \tilde{\theta}_2 \) where \( \tilde{\theta}_2(y) = \theta_0(y)1(|y| \geq \tilde{\eta}N^3a) \). Denote

\[
\tilde{\theta} = (1 - \psi(a))^{-1} \tilde{\theta}_1, \quad \psi(a) := \int \tilde{\theta}_2(y)dy \leq N^{-3a}
\]

so that \( \int \tilde{\theta} = 1 \). Using \( |\phi''_{E_1, E_2}(x)| \leq C\eta^{-2}\theta_{\eta}(x - E_1) \) and \( |\phi''_{E_1, E_2}(x)| \leq C\eta^{-3}\theta_{\eta}(x - E_1) \), we have

\[
|\tilde{\theta} \ast \phi_{E_1, E_2}(x) - \phi_{E_1, E_2}(x)| \leq \frac{1}{1 - \psi(a)} \left| \int dy\tilde{\theta}_1(y)[\phi_{E_1, E_2}(x - y) - \phi_{E_1, E_2}(x)] \right| \leq C \left| \int dy\tilde{\theta}_1(y)[\phi''_{E_1, E_2}(x)y + O(\phi''_{E_1, E_2}(x)y^2)] \right| \leq C\tilde{\eta}^2 N^{6a}\theta_{\eta}(x - E_1), \tag{15.30}
\]

where we used the symmetry of \( \tilde{\theta}_1 \), i.e., \( \int dy\tilde{\theta}_1(y)y = 0 \) and that \( \tilde{\theta}_1(y) \) is supported on \( |y| \leq \tilde{\eta}N^3a \). Together with (15.11) with the choice of \( \eta = N^{-1+\varepsilon} \) for some small \( \varepsilon > 0 \), we have

\[
\left| \mathbb{E}^{w} \left[ \sum_{i,k} \tilde{\theta} \ast \phi_{E_1, E_2}(\lambda_i) - \phi_{E_1, E_2}(\lambda_i) \right] \theta_{\eta}(\lambda_k - E_3) \right| \leq \frac{\tilde{\eta}^2 N^{6a}}{N\eta^3} \left| \mathbb{E}^{w} \left[ \sum_{i,k} \theta_{\eta}(\lambda_i - E_1)\theta_{\eta}(\lambda_k - E_3) \right] \right| \leq \frac{\tilde{\eta}^2 N^{6a+2\varepsilon}}{N^2} \leq N^{-6a}, \tag{15.32}
\]

where we used the a priori bound (15.7) and then \( \varepsilon \leq a/2 \) in the two last steps. Hence we can approximate \( \mathbb{E}^{w} B_3 \) by

\[
\mathbb{E}^{w} B_3 = \mathbb{E}^{w} \int dy\phi_{E_1, E_2}(y)N^{-3} \sum_{i,k} \tilde{\theta}(\lambda_i - y)\theta_{\eta}(\lambda_k - E_3) + O(N^{-a}). \tag{15.33}
\]

Recall \( \theta_{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2 \) and \( \tilde{\theta} = (1 - \psi(a))^{-1}\tilde{\theta}_1 \). We wish to replace \( \tilde{\theta} \) in the right hand side of (15.33) by \( (1 - \psi(a))^{-1}\theta_{\eta} \) with an error \( O(N^{-a}) \). For this purpose, we need to prove that the additional contribution from \( \tilde{\theta}_2 \) is bounded by \( O(N^{-a}) \).
Since $\hat{\theta}_2(y) \leq C N^{-3a} \theta_{N^3 \hat{\eta}}(y)$, we have
\[
\int dy \phi_{E_1, E_2}(y) E^w N^{-3} \sum_{i,k} \hat{\theta}_2(\lambda_i - y) \theta_\eta(\lambda_k - E_3) \leq \int dy \phi_{E_1, E_2}(y) E^w N^{-3} \sum_{i,k} N^{-3a} \theta_{N^3 \hat{\eta}}(\lambda_i - y) \theta_\eta(\lambda_k - E_3).
\] (15.34)

Now we show that the contribution of this error term to the r.h.s. of (15.24) is negligible. Recalling $E_1 = E + \alpha_1/N$ and using $\int d\alpha_1 \theta_\eta(y - E_1) \leq CN$, we have
\[
| \int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) \int dy \theta_\eta(y - E_1) \theta_\eta(y - E_2) E^w N^{-3} \sum_{i,k} \theta_\eta(\lambda_i - y) \theta_\eta(\lambda_k - E_3) | \leq C N^{-3a} \int_{|\alpha_2|+|\alpha_3| \leq C} d\alpha_2 d\alpha_3 E^w N^{-2} \sum_i \theta_\eta \ast \theta_{N^3 \hat{\eta}}(\lambda_i - E_2) \sum_k \theta_\eta(\lambda_k - E_3) \leq C N^{-3a} \int_{|\alpha_2|+|\alpha_3| \leq C} d\alpha_2 d\alpha_3 E^w N^{-2} \sum_i \theta_\eta(\lambda_i - E_2) \sum_k \theta_\eta(\lambda_k - E_3),
\] (15.35)

where we have used $\theta_\eta \ast \theta_\eta' \leq C \theta_\eta$ if $\eta > \eta'$ with $C$ independent of $\eta, \eta'$. Notice that we have only used that $O$ is compactly supported and $\|O\|_\infty$ is bounded. Using (15.11) and (15.7), we can bound the last line by $N^{-a + C\varepsilon}$.

As we will choose $\varepsilon \ll a$, neglecting the $C\varepsilon$ exponent, we can thus approximate the contribution of $E^w B_3$ to the r.h.s. of (15.24) by
\[
\int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) E^w B_3 = (1 - \psi(a))^{-1} \int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) \int dy \phi_{E_1, E_2}(y) E^w N^{-3} \sum_{i,k} \theta_\eta(\lambda_i - y) \theta_\eta(\lambda_k - E_3) + O(N^{-a}).
\] (15.36)

The same estimate holds if we replace the expectation $E^w$ by $E^\nu$. Recalling (15.8), we can thus estimate their difference by
\[
| \int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) [E^w - E^\nu] B_3 | \leq C \int dy \phi_{E_1, E_2}(y) \left| E^w - E^\nu \right| N^{-3} \sum_{i,k} \theta_\eta(\lambda_i - y) \theta_\eta(\lambda_k - E_3) + O(N^{-a})
\] (15.37)
\[
\leq C(N\eta)^{-1} N^{-\delta} + O(N^{-a}) \leq O(N^{-a}),
\]
provided that $a \leq \delta/2$. Similar arguments can be used to prove that
\[
\int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) [E^w - E^\nu] [A_1 + A_2 + A_3] = O(N^{-a}).
\] (15.38)

Recalling (15.24), we have thus proved that
\[
\int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 O_\eta(\beta_1, \beta_2, \beta_3) (p_{w,N}^{(3)} - p_{\nu,N}^{(3)}) \left( E + \frac{\beta_1}{N}, \ldots, E + \frac{\beta_3}{N} \right) = O(N^{-a}),
\] (15.39)
for any $O$ compactly supported with $\|O\|_\infty$ bounded.

In order to prove Theorem 15.3, it remains to replace $O_\eta$ with $O$ in (15.39); at this point we will use that $O$ is differentiable. For this purpose, we only need to bound the error
\[
\int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 (O - O_\eta)(\beta_1, \beta_2, \beta_3) p_{w,N}^{(3)} \left( E + \frac{\beta_1}{N}, \ldots, E + \frac{\beta_3}{N} \right) = O(N^{-a+\varepsilon})
\] (15.40)
and use the similar estimate with \( w \) replaced by \( v \). One can check easily that there is a \( C^1 \) nonnegative function \( \tilde{O} \) with compact support and \( \| \tilde{O} \|_{\infty} \leq 1 \) such that

\[
|O - O_\eta| \leq C(N\eta)\tilde{O}_\eta. \tag{15.41}
\]

Hence (15.40) follows from applying (15.28) to \( \tilde{O} \). Choosing \( \varepsilon \) much smaller than \( a \), this completes the proof of Theorem 15.3.
16 Universality of Wigner matrices in small energy window: GFT

We have proved the universality of Wigner matrices in energy windows bigger than $N^{-1/2+\varepsilon}$ in Section 15. In this section, we will use the Green function comparison theorem to improve it to any energy windows of size bigger than $N^{-1+\varepsilon}$. This is the Step 3 in the three step strategy introduced in Section 5. In the following, we will first prove the Green function comparison theorem and then use it to prove our main universality theorem, i.e., Theorem 5.1.

16.1 The Green function comparison theorems

The main ingredient to prove this result is the following Green function comparison theorem stating that the correlation functions of eigenvalues of two matrix ensembles are identical on scale $1/N$ provided that the first four moments of all matrix elements of these two ensembles are almost identical. Here we do not assume that the real and imaginary parts are i.i.d., hence the $k$-th moment of $h_{ij}$ is understood as the collection of numbers $\int h^s h^{k-s} \nu_{ij}(dh)$, $s = 0, 1, 2, \ldots, k$.

Before stating the result we explain a notation. We will distinguish between the two ensembles by using different letters, $v_{ij}$ and $w_{ij}$ for their matrix elements and we often use the notation $H^{(v)}$ and $H^{(w)}$ to indicate the difference. Alternatively, one could denote the matrix elements of $H$ by $h_{ij}$ and make the distinction of the two ensembles in the measure, especially in the expectation, by using notations $E^v$ and $E^w$. Since the matrix elements will be replaced one by one from one distribution to the other, the latter notation would have been cumbersome and so we will follow the first convention in this section.

**Theorem 16.1** (Green function comparison). [69, Theorem 2.3] Suppose that we have two generalized $N \times N$ Wigner matrices, $H^{(v)}$ and $H^{(w)}$, with matrix elements $h_{ij}$ given by the random variables $N^{-1/2} v_{ij}$ and $N^{-1/2} w_{ij}$, respectively, with $v_{ij}$ and $w_{ij}$ satisfying the uniform polynomial decay condition (5.6). Fix a bijective ordering map on the index set of the independent matrix elements,

$$\phi : \{(i,j) : 1 \leq i \leq j \leq N\} \rightarrow \{1, \ldots, \gamma(N)\}, \quad \gamma(N) := \frac{N(N+1)}{2},$$

and denote by $H_\gamma$ the generalized Wigner matrix whose matrix elements $h_{ij}$ follow the $v$-distribution if $\phi(i,j) \leq \gamma$ and they follow the $w$-distribution otherwise; in particular $H^{(v)} = H_0$ and $H^{(w)} = H_{\gamma(N)}$. Let $\kappa > 0$ be arbitrary and suppose that, for any small parameter $\tau > 0$ and for any $y \geq N^{-1+\tau}$, we have the following estimate on the diagonal elements of the resolvent

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k \leq N} \max_{|E| \leq 2-\kappa} \left| \left( \frac{1}{H_\gamma - E - iy} \right)_{kk} \right| \lesssim N^{2\tau}$$

with some constants $C, c$ depending only on $\tau, \kappa$. Moreover, we assume that the first four moments of $v_{ij}$ and $w_{ij}$ satisfy that

$$|E \nu_{ij}^s v_{ij}^{4-s} - E \nu_{ij}^s w_{ij}^{4-s}| \lesssim N^{-\delta+2+2s/2}, \quad s = 0, 1, 2, 3, 4,$$

for some given $\delta > 0$. Let $\varepsilon > 0$ be arbitrary and choose an $\eta$ with $N^{-1-\varepsilon} \leq \eta \leq N^{-1}$. For any sequence of complex parameters $z_j = E_j \pm \eta j$, $j = 1, \ldots, n$, with $|E_j| \leq 2 - 2\kappa$ and with an arbitrary choice of the $\pm$ signs. Let $G^{(v)}(z) = (H^{(v)} - z)^{-1}$ denote the resolvent and let $F(x_1, \ldots, x_n)$ be a function such that for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $1 \leq |\alpha| \leq 5$ and for any $\varepsilon' > 0$ sufficiently small, we have

$$\max \left\{ |\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \leq N^{\varepsilon'} \right\} \leq N^{C_0 \varepsilon'}$$

and

$$\max \left\{ |\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \leq N^2 \right\} \leq N^{C_0}$$

for some constant $C_0$.  

136
Then, there is a constant $C_1$, depending on $\hat{v}$, $\sum_m k_m$ and $C_0$ such that for any $\eta$ with $N^{-1-\varepsilon} \leq \eta \leq N^{-1}$ and for any choices of the signs in the imaginary part of $z^m_j$, we have

$$\left| \mathbb{E} F\left( G_{a_1 b_1}^{(v)} (z_1), \ldots, G_{a_n b_n}^{(v)} (z_n) \right) - \mathbb{E} F\left( G(v) \rightarrow G(w) \right) \right| \leq C_1 N^{-1/2+\varepsilon} + C_1 N^{-\delta+\varepsilon},$$

(16.5)

where the arguments of $F$ in the second term are changed from the Green functions of $H(v)$ to $H(w)$ and all other parameters remain unchanged.

Moreover, for any $|E| \leq 2 - 2\kappa$, any $k \geq 1$ and any compactly supported smooth test function $O : \mathbb{R}^n \rightarrow \mathbb{R}$ we have

$$\lim_{N \rightarrow \infty} \int_{\mathbb{R}^n} d\alpha \ O(\alpha) \left( p^{(n)}_{v,N} - p^{(n)}_{w,N} \right) \left( E + \frac{\alpha}{N} \right) = 0,$$

(16.6)

where $p^{(n)}_{v,N}$ and $p^{(n)}_{w,N}$ denote the $n$-point correlation functions of the $v$ and $w$-ensembles, respectively.

Remark 1: We formulated Theorem 16.1 for functions of traces of monomials of the Green function because this is the form we need in the application. However, the result (and the proof we are going to present) holds directly for matrix elements of monomials of Green functions as well, for the precise statement, see [69]. We also remark that Theorem 16.1 holds for generalized Wigner matrices if $C_{\text{sup}} = \sup_{ij} N s_{ij} < \infty$. The positive lower bound on the variances, $C_{i,n,f} > 0$ in (2.6) is not necessary for this theorem.

Remark 2: Although we state Theorem 16.1 for Hermitian and symmetric ensembles, similar results hold for real and complex sample covariance ensembles; the modification of the proof is obvious.

Proof of the Green function comparison Theorem 16.1: The basic idea is to estimate the effect of changing matrix elements of the resolvent one by one by a resolvent expansion. Since each matrix element has a typical size of $N^{-1/2}$ and the resolvents are essentially bounded thanks to (16.1), a resolvent expansion up to the fourth order will identify the change of each element with a precision $O(N^{-5/2})$, modulo some tiny corrections of order $N^{O(r)}$. The expectation values of the terms up to fourth order involve only the first four moments of the single entry distribution, which can be directly compared. The error terms are negligible even when we sum them up $N^2$ times, the number of comparison steps needed to replace all matrix elements.

Now we turn to the one by one replacement. For notational simplicity, we will consider the case when the test function $F$ has only one variable and $k_1 = 1$, i.e., we consider the trace of a first order monomial; the general case follows analogously. Consider the telescopic sum of differences of expectations

$$\mathbb{E} F\left( \frac{1}{N} \text{Tr} \frac{1}{H(v) - z} \right) = \sum_{\gamma=1}^{N} \left[ \mathbb{E} F\left( \frac{1}{N} \text{Tr} \frac{1}{H_\gamma - z} \right) - \mathbb{E} F\left( \frac{1}{N} \text{Tr} \frac{1}{H_{\gamma-1} - z} \right) \right].$$

(16.7)

Let $E^{(ij)}$ denote the matrix whose matrix elements are zero everywhere except at the $(i, j)$ position, where it is 1, i.e., $E^{(ij)}_{k\ell} = \delta_{ik}\delta_{j\ell}$. Fix an $\gamma \geq 1$ and let $(i, j)$ be determined by $\phi(i, j) = \gamma$. We will compare $H_{\gamma-1}$ with $H_\gamma$. Note that these two matrices differ only in the $(i, j)$ and $(j, i)$ matrix elements and they can be written as

$$H_{\gamma-1} = Q + \frac{1}{\sqrt{N}} V,$$

$$H_{\gamma} = Q + \frac{1}{\sqrt{N}} W,$$

$$V := v_{ij} E^{(ij)} + v_{ji} E^{(ji)},$$

$$W := w_{ij} E^{(ij)} + w_{ji} E^{(ji)},$$

with a matrix $Q$ that has zero matrix element at the $(i, j)$ and $(j, i)$ positions and where we set $v_{ji} := v_{ij}$ for $i < j$ and similarly for $w$. Define the Green functions

$$R := \frac{1}{Q - z}, \quad S := \frac{1}{H_{\gamma} - z}.$$
We first claim that the estimate (15.16) holds for the Green function $R$ as well. To see this, we have, from the resolvent expansion,

$$R = S + N^{-1/2}SVS + \ldots + N^{-9/5}(SV)^9S + N^{-5}(SV)^{10}R.$$  

Since $V$ has only at most two nonzero element, when computing the $(k, \ell)$ matrix element of this matrix identity, each term is a finite sum involving matrix elements of $S$ or $R$ and $v_{ij}$, e.g. $(SVS)_{kl} = S_{kl}v_{ij}S_{jl} + S_{kj}v_{ji}S_{lk}$. Using the bound (15.16) for the $S$ matrix elements, the subexponential decay for $v_{ij}$ and the trivial bound $|R_{ij}| \leq \eta^{-1}$, we obtain that the estimate (15.16) holds for $R$.

We can now start proving the main result by comparing the resolvents of $H^{(\gamma - 1)}$ and $H^{(\gamma)}$ with the resolvent $R$ of the reference matrix $Q$. By the resolvent expansion,

$$S = R - N^{-1/2}RVR + N^{-1}(RV)^2R - N^{-3/2}(RV)^3R + N^{-2}(RV)^4R - N^{-5/2}(RV)^5S,$$  

so we can write

$$\frac{1}{N} \text{Tr} S = \hat{R} + \xi, \quad \xi := \sum_{m=1}^{4} N^{-m/2} \hat{R}_v^{(m)} + N^{-5/2} \Omega_v$$

with

$$\hat{R} := \frac{1}{N} \text{Tr} R, \quad \hat{R}_v^{(m)} := (-1)^m \frac{1}{N} \text{Tr}(RV)^m R, \quad \Omega_v := -\frac{1}{N} \text{Tr}(RV)^5 S.$$  

For each diagonal element in the computation of these traces, the contribution to $\hat{R}, \hat{R}_v^{(m)}$ and $\Omega_v$ is a sum of a few terms. E.g.

$$\hat{R}_v^{(2)} = \frac{1}{N} \sum_k \left[ R_{kj}v_{ij}R_{j\ell}v_{ki} + R_{kj}v_{ij}R_{j\ell}v_{ki} + R_{kj}v_{ij}R_{j\ell}v_{ki} + R_{kj}v_{ij}R_{j\ell}v_{ki} \right]$$

and similar formulas hold for the other terms. Then we have

$$\mathbb{E}F\left( \frac{1}{N} \text{Tr} \frac{1}{H_\gamma - z} \right) = \mathbb{E}F\left( \hat{R} + \xi \right)$$

$$= \mathbb{E}\left[ F(\hat{R}) + F'(\hat{R})\xi + F''(\hat{R})\xi^2 + \ldots + F^{(5)}(\hat{R} + \xi')\xi^5 \right]$$

$$= \sum_{m=0}^{5} N^{-m/2} \mathbb{E}A_v^{(m)},$$

where $\xi'$ is a number between 0 and $\xi$ and it depends on $\hat{R}$ and $\xi$; the $A_v^{(m)}$s are defined as

$$A_v^{(0)} = F(\hat{R}), \quad A_v^{(1)} = F'(\hat{R})\hat{R}_v^{(1)}, \quad A_v^{(2)} = F''(\hat{R})(\hat{R}_v^{(1)})^2 + F'(\hat{R})\hat{R}_v^{(2)},$$

and similarly for $A_v^{(3)}$ and $A_v^{(4)}$. Finally,

$$A_v^{(5)} = F'(\hat{R})\Omega + F^{(5)}(\hat{R} + \xi')(\hat{R}_v^{(1)})^5 + \ldots.$$  

The expectation values of the terms $A_v^{(m)}$, $m \leq 4$, with respect to $v_{ij}$ are determined by the first four moments of $v_{ij}$, for example

$$\mathbb{E}A_v^{(2)} = F'(\hat{R})\left[ \frac{1}{N} \sum_k R_{kj}R_{j\ell}R_{\ell i} + \ldots \right] \mathbb{E}|v_{ij}|^2 + F''(\hat{R})\left[ \frac{1}{N^2} \sum_{k,\ell} R_{kj}R_{j\ell}R_{\ell i}R_{\ell j} + \ldots \right] \mathbb{E}|v_{ij}|^2$$

$$+ F''(\hat{R})\left[ \frac{1}{N} \sum_k R_{kj}R_{i\ell}R_{\ell j} + \ldots \right] \mathbb{E}v_{ij}^2 + F^{(5)}(\hat{R})\left[ \frac{1}{N^2} \sum_{k,\ell} R_{kj}R_{i\ell}R_{\ell i}R_{j\ell} + \ldots \right] \mathbb{E}v_{ij}^2.$$  

138
Note that the coefficients involve up to four derivatives of $F$ and normalized sums of matrix elements of $R$. Using the estimate (15.16) for $R$ and the derivative bounds (16.3) for the typical values of $\hat{R}$, we see that all these coefficients are bounded by $N^{C(\tau+\varepsilon)}$ with a very large probability, where $C$ is an explicit constant.

We use the bound (16.4) for the extreme values of $\hat{R}$ but this event has a very small probability by (15.16). Therefore, the coefficients of the moments $\mathbb{E} \hat{v}_{ij}^s v_{ij}^s$, $u+s \leq 4$, in the quantities $A_0^{(0)}, \ldots, A_4^{(4)}$ are essentially bounded, modulo a factor $N^{C(\tau+\varepsilon)}$. Notice that the $k$-th moment of $v_{ij}$ appears only in the $m = k$ term that already has a prefactor $N^{-k/2}$ in (16.10).

Finally, we have to estimate the error term $A_5^{(5)}$. All terms without $\Omega$ can be dealt with as before; after estimating the derivatives of $F$ by $N^{C(\tau+\varepsilon)}$, one can perform the expectation with respect to $v_{ij}$ that is independent of $\hat{R}^{(m)}$. For the terms involving $\Omega$ one can argue similarly, by appealing to the fact that the matrix elements of $S$ are also essentially bounded by $N^{C(\tau+\varepsilon)}$, see (15.16), and that $v_{ij}$ has subexponential decay. Alternatively, one can use Hölder inequality to decouple $S$ from the rest and use (15.16) directly, for example:

$$\mathbb{E} |F'(\hat{R})\Omega| = \frac{1}{N} |\mathbb{E} F'(\hat{R}) \text{Tr} (RV)^5 S| \leq \frac{1}{N} \left[ \mathbb{E} (F'(\hat{R}))^2 \text{Tr} S^2 \right]^{1/2} \left[ \mathbb{E} \text{Tr} (RV)^5 (VR^*)^5 \right]^{1/2} \leq CN^{-\frac{1}{2}+C(\tau+\varepsilon)}.$$

Note that exactly the same perturbation expansion holds for the resolvent of $H_{\tau-1}$ with $A_{\nu}$ replaced by $A_{\nu}$, i.e., $v_{ij}$ is replaced with $w_{ij}$ everywhere. By the moment matching condition (16.2), the expectation values

$$N^{-m/2} \left| \mathbb{E} A_{\nu}^{(m)} - \mathbb{E} A_{\nu}^{(m)} \right| \leq N^{C(\tau+\varepsilon)-\delta-2}$$

for $m \leq 4$ in (16.10). Choosing $\tau = \varepsilon$, we have

$$\mathbb{E} F \left( \frac{1}{N} \text{Tr} \frac{1}{H_{\tau-1} - z} \right) - \mathbb{E} F \left( \frac{1}{N} \text{Tr} \frac{1}{H_{\tau} - z} \right) \leq CN^{-5/2+C\varepsilon} + CN^{-2-\delta+C\varepsilon}.$$

After summing up in (16.7) we have thus proved that

$$\mathbb{E} F \left( \frac{1}{N} \text{Tr} \frac{1}{H^{(\nu)} - z} \right) - \mathbb{E} F \left( \frac{1}{N} \text{Tr} \frac{1}{H^{(\omega)} - z} \right) \leq CN^{-1/2+C\varepsilon} + CN^{-\delta+C\varepsilon}.$$

The proof can be easily generalized to functions of several variables. This concludes the proof of Theorem 16.1.

### 16.2 Conclusion of the three step strategy

In this section we put the previous results together to prove our main result Theorem 5.1. Recall the main result of the step 2, Theorem 12.4, asserts the universality of Gaussian divisible ensemble. However, in order to reach energy window $N^{-1+\varepsilon}$, we need to substantial large Gauss component. In terms of DBM, we need the time $t \geq N^{-\varepsilon c}$ for some $c > 0$. The eigenvalue continuity result in the previous section is unfortunately restricted to $t \leq N^{-1-\varepsilon}$. To over come the deficiency of the continuity, we will approximate the Wigner ensemble by a suitably chosen Gaussian divisible ensemble. Thus we term this step “Approximation by a Gaussian divisible ensemble”.

We first review the classical moment matching problem for real random variables. For any real random variable $\xi$, denote by $m_k(\xi) = \mathbb{E} \xi^k$ its $k$-th moment. By Schwarz inequality, the sequence of moments, $m_1, m_2, \ldots$ are not arbitrary numbers, for example $|m_1|^k \leq m_k$ and $m_2^2 \leq m_4$, etc, but there are more subtle relations. For example, if $m_1 = 0$, then

$$m_4m_2 - m_2^3 \geq m_2^3,$$

which can be obtained by

$$m_3^2 = [\mathbb{E} \xi^3]^2 = [\mathbb{E} \xi(\xi^2-1)]^2 \leq [\mathbb{E} \xi^2] [\mathbb{E}(\xi^2-1)^2] = m_2(m_4-2m_2^2+1)$$

for $m_1 = 0$. We observe that the moments $m_1, m_3, m_4$ can be obtained directly from the means $m_1, m_2, m_3, m_4$.
and noticing that (16.11) is scale invariant, so it is sufficient to prove it for \( m_2 = 1 \). In fact, it is easy to see that (16.11) saturates if and only of the support of \( \xi \) consists of exactly two points (apart from the trivial case when \( \xi \equiv 0 \)).

This restriction shows that given a sequence of four admissible moments, \( m_1 = 0, m_2 = 1, m_3, m_4 \), there may not exist a Gaussian divisible random variable \( \xi \) with these moments; e.g. the moment sequence \((m_1, m_2, m_3, m_4) = (0, 1, 0, 1)\) uniquely characterizes the standard Bernoulli variable \((\xi = \pm 1 \text{ with } 1/2, 1/2 \text{ probability})\). However, if we allow a bit room in the fourth moment, then one can match any four admissible moments by a small Gaussian divisible variable. This is the content of the next lemma \([69, \text{Lemma 6.5}]\). This matching idea in the context of random matrices appeared earlier in \([130]\). For simplicity, we formulate it in the case of real variables, the extension to complex variables is standard.

**Lemma 16.2. [Moment matching]** Let \( m_3 \) and \( m_4 \) be two real numbers such that

\[
m_4 - m_3^2 - 1 \geq 0, \quad m_4 \leq C_2
\]  

(16.12)

for some positive constant \( C_2 \). Let \( \xi^G \) be a Gaussian random variable with mean 0 and variance 1. Then for any sufficient small \( \gamma > 0 \) (depending on \( C_2 \)), there exists a real random variable \( \xi_\gamma \) with subexponential decay and independent of \( \xi^G \), such that the first four moments of

\[
\xi' = (1 - \gamma)^{1/2} \xi_\gamma + \gamma^{1/2} \xi^G
\]  

(16.13)

are \( m_1(\xi') = 0, m_2(\xi') = 1, m_3(\xi') = m_3 \) and \( m_4(\xi') \), and

\[
|m_4(\xi') - m_4| \leq C\gamma
\]  

(16.14)

for some \( C \) depending on \( C_2 \).

**Proof.** We first construct a random variable \( X \) with the first four moments given by \( 0, 1, m_3, m_4 \) satisfying \( m_4 - m_3^2 - 1 \geq 0 \). We take the law of \( X \) to be of the form

\[
p\delta_a + q\delta_{-b} + (1 - p - q)\delta_0,
\]

where \( a, b, p, q \geq 0 \) are parameters satisfying \( p + q \leq 1 \). The conditions \( m_1(X) = 0 \) and \( m_2(X) = 1 \) imply

\[
p = \frac{1}{a(a + b)}, \quad q = \frac{1}{b(a + b)}.
\]

Furthermore, the condition \( p + q \leq 1 \) reads \( ab \geq 1 \). By an explicit computation we find

\[
m_3(X) = a - b, \quad m_4(X) = m_3(X)^2 + ab.
\]  

(16.15)

Clearly, the condition \( m_4 - m_3^2 - 1 \geq 0 \) implies that (16.15) has a solution with \( ab \geq 1 \). This proves the existence of a random variable supported at three points with given four moments \( 0, 1, m_3, m_4 \) satisfying \( m_4 - m_3^2 - 1 \geq 0 \).

Our main task is to construct a Gaussian divisible one which approximate the four moment matching condition stated in the lemma. For any real random variable \( \zeta \), independent of \( \xi^G \), and with the first four moments being \( 0, 1, m_3(\zeta) \) and \( m_4(\zeta) < \infty \), the first four moments of

\[
\zeta' = (1 - \gamma)^{1/2} \zeta + \gamma^{1/2} \xi^G
\]  

(16.16)

are \( 0, 1, m_3(\zeta') = (1 - \gamma)^{3/2} m_3(\zeta) \)

(16.17)

and

\[
m_4(\zeta') = (1 - \gamma)^2 m_4(\zeta) + 6\gamma - 3\gamma^2.
\]  

(16.18)
Since we can match four moments by a random variable, for any $\gamma > 0$ there exists a real random variable $\xi_\gamma$ such that the first four moments are 0, 1,

$$m_3(\xi_\gamma) = (1 - \gamma)^{-3/2}m_3$$  \hspace{1cm} (16.19)

and

$$m_4(\xi_\gamma) = m_3(\xi_\gamma)^2 + (m_4 - m_3^2).$$

With $m_4 \leq C_2$, we have $m_3^2 \leq C_2^{3/2}$, thus

$$|m_4(\xi_\gamma) - m_4| \leq C\gamma$$ \hspace{1cm} (16.20)

for some $C$ depending on $C_2$. Hence with (16.17) and (16.18), we obtain that $\xi' = (1 - \gamma)^{1/2}\xi_\gamma + \gamma^{1/2}\xi_G$ satisfies $m_3(\xi') = m_3$ and (16.14). This completes the proof of Lemma 16.2.

We now prove Theorem 5.1, i.e, that the limit in (15.1) holds. We restrict ourselves to the real symmetric case, since the moment matching Lemma 16.2 was formulated for real random variables. A similar argument works for the complex case. From the universality of Gaussian divisible ensembles (more precisely, (12.24)) for any $b = N^{-1+10\varepsilon}$ and $t \geq N^{-\varepsilon}$, we have

$$\frac{1}{2b} \int_{E-b}^{E+b} \int_{\mathbb{R}^n} d\alpha O(\alpha)(p_{t,N}^{(n)} - p_{G,N}^{(n)})(E' + \frac{\alpha}{N}) \rightarrow 0,$$ \hspace{1cm} (16.21)

where $p_{t,N}^{(n)}$ is the eigenvalue correlation function of the matrix ensemble $H_t = e^{-t/2}H_0 + \sqrt{1 - e^{-t}}H_G$. The initial matrix ensemble $H_0$ can be any Wigner ensemble. Recall $H$ is the Wigner ensemble for which we wish to prove (15.1). Given the first four moments $m_1 = 0, m_2 = 1, m_3, m_4$ of the rescaled matrix elements $\sqrt{N}h_{ij}$ of $H$, we first use Lemma 16.2 to construct a distribution $\xi_\gamma$ with $\gamma = 1 - e^{-t}$. This distribution has the property that the first four moments of $\xi'$, defined by (16.13), matches the first four moments of the rescaled matrix elements of $H$. We now choose $H_0$ to be the Wigner ensemble with rescaled matrix elements distributed according to $\xi_\gamma$. Then, on one hand, the matrix $H_t$ will have universal local spectral statistics in the sense of (16.21); on the other hand we can apply the Green function comparison Theorem 16.1 to the matrix ensembles $H$ and $H_t$ so that (15.9) holds for these two ensembles. Clearly, (15.1) follows from (16.21) and (15.9). Note that averaging in the energy parameter $E$ was necessary only for the first step in (16.21), the comparison argument works for any fixed energy $E$. \hfill $\Box$
17 Edge Universality

The main result of edge universality for generalized Wigner matrix is given by the following theorem. For simplicity, we formulate it only for the real symmetric case, the complex Hermitian case is analogous.

**Theorem 17.1 (Universality of extreme eigenvalues).** Suppose that we have two real symmetric generalized $N \times N$ Wigner matrices, $H^{(v)}$ and $H^{(w)}$, with matrix elements $h_{ij}$ given by the random variables $N^{-1/2}v_{ij}$ and $N^{-1/2}w_{ij}$, respectively, with $v_{ij}$ and $w_{ij}$ satisfying the uniform subexponential decay condition (2.7). Suppose that

$$Ev_{ij}^2 = Ew_{ij}^2.$$  \hspace{1cm} (17.1)

Let $\lambda_N^{(v)}$ and $\lambda_N^{(w)}$ denote the largest eigenvalues of $H^{(v)}$ and $H^{(w)}$, respectively. Then there is an $\varepsilon > 0$ and $\delta > 0$ depending on $\vartheta$ in (2.7) such that for any real parameter $s$ (may depend on $N$) we have

$$P(N^{2/3}(\lambda_N^{(v)} - 2) \leq s - N^{-\varepsilon}) - N^{-\delta} \leq P(N^{2/3}(\lambda_N^{(w)} - 2) \leq s) \leq P(N^{2/3}(\lambda_N^{(v)} - 2) \leq s + N^{-\varepsilon}) + N^{-\delta}$$ \hspace{1cm} (17.2)

for $N \geq N_0$ sufficiently large, where $N_0$ is independent of $s$. Analogous result holds for the smallest eigenvalue $\lambda_1$.

This theorem shows that the statistics of the extreme eigenvalues depend only on the second moments of the centered matrix entries, but the result does not determine the distribution. For the Gaussian case, the corresponding distribution was identified by Tracy and Widom in [133,134]. Theorem 17.1 immediately gives the same result for Wigner matrices, but not yet for generalized Wigner matrices. In fact, the Tracy-Widom law for generalized Wigner matrices does not follow from Theorem 17.1 and its proof in [24] required a quite different argument, see Theorem 18.7.

We first give an outline of the proof of Theorem 17.1. Denote by $\varrho_N^{(v)}$ the empirical eigenvalue distribution

$$\varrho_N^{(v)}(x) = \frac{1}{N} \sum_{\alpha=1}^{N} \delta(x - \lambda_N^{(v)}).$$

Our goal is to compare the difference of $\varrho_N^{(v)}$ and $\varrho_N^{(w)}$ via their Stieltjes transforms, $m_N^{(v)}(z)$ and $m_N^{(w)}(z)$. We will be able to locate the largest eigenvalue via the distribution of certain functionals of the Stieltjes transforms. This will be done in the preparatory Lemma 17.2 and its Corollary 17.3. Therefore, if $m_N^{(v)}(z)$ and $m_N^{(w)}(z)$ are sufficiently close, we will be able to compare $\lambda_N^{(v)}$ and $\lambda_N^{(w)}$.

The main ingredient of the proof of Theorem 17.1 is the Green function comparison theorem at the edge (Theorem 17.4), which shows that the distributions of $m_N^{(v)}(z)$ and $m_N^{(w)}(z)$ on the critical scale $\Im z \approx N^{-2/3}$ are the same provided the second moments of the matrix elements match. This will be done by a resolvent expansion, similarly to the Green function comparison theorem in the bulk (Theorem 16.1). However, the resolvent expansion is more efficient at the edge for a reason that we explain now.

Recall (6.13) stating that $m_{sc}(z) \approx \eta$ if $\kappa = |E| - 2 |\leq \eta$ and $z = E + i\eta$. Since the extremal eigenvalue gaps are expected to be order $N^{-2/3}$, we need to set $\eta \ll N^{-2/3}$ in order to identify individual eigenvalues. This is a smaller scale than that of the local semicircle law; in fact $m_N(z)$ does not have a deterministic limit; we need to identify its distribution. Hence the reference size of $\Im m_N(z)$ which we should keep in mind is $N^{-1/3}$. The largest eigenvalue can be located via the Stieltjes transform as follows. By definition of $m_N$, if there is an eigenvalue in a neighborhood of $E$ within distance $\eta$ then

$$\Im m_N(E + i\eta) = \frac{1}{N} \sum_{\alpha} \frac{\eta}{(\lambda_{\alpha} - E)^2 + \eta^2} \geq \frac{1}{N\eta} \gg N^{-1/3},$$

and otherwise $\Im m_N(E + i\eta) \lesssim N^{-1/3}$. Based upon this idea, we will construct a certain functional of $\Im m_N$ that expresses the distribution of $\lambda_N$. In other words, we expect that if we can control the imaginary part of the trace of the Green function by $N^{-1/3}$, then we can identify the individual extremal eigenvalues.
Roughly speaking, our basic principle is that whenever we understand \( m_N(z) \) with a precision better than \((N\eta)^{-1}\), i.e., we can improve over the strong local semicircle law (6.32)

\[
|m_N(z) - m_{ac}(z)| < \frac{1}{N\eta}, \quad \text{Im} z = \eta
\]

at some scale \( \eta \), then we can "identify" the eigenvalue distribution at that scale precisely. It is instructive to compare the edge situation with the bulk, where the critical scale \( \eta \approx 1/N \) identifies individual eigenvalues and the typical size of \( \text{Im} m_N \) is of order 1. In other words, at the critical scale \( \text{Im} m_N \) is of order one in the bulk and it is of order \( N^{-1/3} \) at the edge. Owing to (6.31), this fact extends to each resolvent matrix element, i.e., \(|G_{ij}| \lesssim N^{-1/3}\) at the edge. Thus a resolvent expansion is more powerful near the edge, which explains that less moment matching is needed than in the bulk.

Now we start the detailed proof of Theorem 17.1. We first introduce a few notation. For any \( E_1 \leq E_2 \), let

\[
\mathcal{N}(E_1, E_2) := \# \{ j : E_1 \leq \lambda_j \leq E_2 \} = \text{Tr} \mathbf{1}_{[E_1, E_2]}(H)
\]

be the unnormalized counting function of the eigenvalues. Set \( E_+ := 2 + N^{-2/3+\varepsilon} \). From rigidity, Theorem 11.5, we know that \( \lambda_N \lesssim E_+ \), i.e.,

\[
\mathcal{N}(E, \infty) = \mathcal{N}(E, E_+)	ag{17.3}
\]

with very high probability, so it is sufficient to consider the counting function \( \mathcal{N}(E, E_+) = \text{Tr} \chi_E(H) \), with \( \chi_E(x) := \mathbf{1}_{[E, E_+]}(x) \). We will approximate the sharp cutoff function \( \chi_E \) its smoothened version \( \chi_E \ast \theta_\eta \) on scale \( \eta \ll N^{-2/3} \), where we recall the notation \( \theta_\eta(x) \) from (15.21). The following lemma shows that the error caused by this replacement is negligible.

**Lemma 17.2.** For any \( \varepsilon > 0 \) set \( \ell_1 = N^{-2/3-3\varepsilon} \) and \( \eta = N^{-2/3-9\varepsilon} \). Then there exist constants \( C, c \) such that for any \( E \) satisfying \(|E - 2| \leq \frac{1}{4} N^{-2/3+\varepsilon} \) and for any \( D \) we have

\[
\mathbb{P}\left\{ \left| \text{Tr} \chi_E(H) - \text{Tr} \chi_E \ast \theta_\eta(H) \right| \leq C\left(N^{-2\varepsilon} + \mathcal{N}(E - \ell_1, E + \ell_1)\right) \right\} \geq 1 - N^{-D} \tag{17.4}
\]

if \( N \gg N_0(D) \) is large enough.

We will not give the detailed proof of this lemma since it is a straightforward approximation argument. We only point out that \( \theta_\eta \) is an approximate delta function on scale \( \eta \ll N^{-2/3} \), and if it were compactly supported, then the difference between \( \text{Tr} \chi_E(H) \) and \( \text{Tr} \chi_E \ast \theta_\eta(H) \) were clearly bounded by the number of eigenvalues in an \( \eta \)-vicinity of \( E \). Since \( \theta_\eta(x) = \pi^{-1} \eta/(x^2 + \eta^2) \), i.e., it has a quadratically decaying tail, the above argument must be complemented by estimating the density of eigenvalues away from \( E \). This estimate comes from two contributions. Eigenvalues within the \( \ell_1 \)-vicinity of \( E \) are directly estimated by the term \( \mathcal{N}(E - \ell_1, E + \ell_1) \). Eigenvalues farther away come with an additional factor \( \eta/\ell_1 = N^{-6\varepsilon} \) due to the decay of \( \theta_\eta \), thus it is sufficient to estimate their density only up to an \( N^{\varepsilon} \) factor precision, which is provided by the optimal local law at the edge. For precise details, see the proof of Lemma 6.1 in [70] for essentially the same argument.

Using (17.3) and the local law to estimate a slightly averaged version of \( \mathcal{N}(E - \ell_1, E + \ell_1) \), we arrive at the following Corollary:

**Corollary 17.3.** Let \(|E - 2| \leq N^{-2/3+\varepsilon} \), \( \ell = \frac{1}{2} \ell_1 N^{2\varepsilon} = \frac{1}{2} N^{-2/3-\varepsilon} \). Then the inequality

\[
\text{Tr} \chi_{E+\ell} \ast \theta_\eta(H) - N^{-\varepsilon} \leq \mathcal{N}(E, \infty) \leq \text{Tr} \chi_{E-\ell} \ast \theta_\eta(H) + N^{-\varepsilon} \tag{17.5}
\]

holds with probability bigger than \( 1 - N^{-D} \) for any \( D \) if \( N \) is large enough. Furthermore, we have

\[
\mathbb{E} F(\text{Tr} \chi_{E-\ell} \ast \theta_\eta(H)) \leq \mathbb{P}(\mathcal{N}(E, \infty) = 0) \leq \mathbb{E} F(\text{Tr} \chi_{E+\ell} \ast \theta_\eta(H)) + CN^{-D}. \tag{17.6}
\]
Proof. We have $E_+ - E \gg \ell$ thus $|E - 2 - \ell| \leq \frac{3}{2} N^{-2/3+\varepsilon}$, therefore (17.4) holds for $E$ replaced with $y \in [E - \ell, E]$ as well. We thus obtain

$$
\text{Tr} \chi_E(H) \leq \ell^{-1} \int_{E-\ell}^E dy \text{Tr} \chi_y(H)
\leq \ell^{-1} \int_{E-\ell}^E dy \text{Tr} \chi_y * \theta_\eta(H) + C \ell^{-1} \int_{E-\ell}^E dy \left[N^{-2\varepsilon} + N(y - \ell_1, y + \ell_1)\right]
\leq \text{Tr} \chi_{E-\ell} * \theta_\eta(H) + CN^{-2\varepsilon} + C \ell^3 \ell^{-1} N(E - 2\ell, E + \ell)
$$

with a probability larger than $1 - N^{-D}$ for any $D > 0$. From the rigidity estimate in the form (11.25), the conditions $|E| \leq N^{-2/3+\varepsilon}$, $\ell_1/\ell = 2N^{-2\varepsilon}$ and $\ell \leq N^{-2/3}$, we can bound

$$
\frac{\ell_1}{\ell} N(E - 2\ell, E + \ell) \leq N^{1-2\varepsilon} \int_{E-2\ell}^{E+\ell} \theta_{\varepsilon}(x) dx + N^{-2\varepsilon} N^{\varepsilon} \leq CN^{-\varepsilon}
$$

with a very high probability, where we estimated the explicit integral using that the integration domain is in a $CN^{-2/3+\varepsilon}$-vicinity of the edge at 2. We have thus proved

$$
N(E, E_+) = \text{Tr} \chi_E(H) \leq \text{Tr} \chi_{E-\ell} * \theta_\eta(H) + N^{-\varepsilon}.
$$

By (17.3), we can replace $N(E, E_+)$ by $N(E, \infty)$. This proves the upper bound of (17.5) and the lower bound can be proved similarly.

On the event that (17.5) holds, the condition $N(E, \infty) = 0$ implies that $\text{Tr} \chi_{E+\ell} * \theta_\eta(H) \leq 1/9$. Thus we have

$$
P(N(E, \infty) = 0) \leq P(\text{Tr} \chi_{E+\ell} * \theta_\eta(H) \leq 1/9) + CN^{-D}.
$$

Together with the Markov inequality, this proves the upper bound in (17.6). For the lower bound, we use

$$
\mathbb{E} F(\text{Tr} \chi_{E-\ell} * \theta_\eta(H)) \leq P(\text{Tr} \chi_{E-\ell} * \theta_\eta(H) \leq 2/9) \leq P(N(E, \infty) \leq 2/9 + N^{-\varepsilon}) = P(N(E, \infty) = 0),
$$

where we used the upper bound from (17.5) and that $N$ is an integer. This completes the proof of the Corollary.

Since $P(\lambda_N < E) = P(N(E, \infty) = 0)$, we will use (17.6) and the identity

$$
\text{Tr} \chi_E * \theta_\eta(H) = N \int_E^E dy \text{Im} m_N(y + i\eta)
$$

(17.8) to relate the distribution of $\lambda_N$ with that of the Stieltjes transform $m_N$. The following Green function comparison theorem shows that the distribution of the right hand side of (17.8) depends only on the second moments of the matrix elements. Its proof will be given after we have completed the proof of Theorem 17.1.

**Theorem 17.4 (Green function comparison theorem on the edge).** Suppose that the assumptions of Theorem 17.1, including (17.1), hold. Let $F : \mathbb{R} \to \mathbb{R}$ be a function whose derivatives satisfy

$$
\max_x |F^{(\alpha)}(x)| (|x| + 1)^{-C_1} \leq C_1, \quad \alpha = 1, 2, 3,
$$

(17.9) with some constant $C_1 > 0$. Then there exists $\varepsilon_0 > 0$ depending only on $C_1$ such that for any $\varepsilon < \varepsilon_0$ and for any real numbers $E_1$ and $E_2$ satisfying

$$
|E_1 - 2| \leq CN^{-2/3+\varepsilon}, \quad |E_2 - 2| \leq CN^{-2/3+\varepsilon},
$$

and setting $\eta = N^{-2/3-\varepsilon}$, we have

$$
\left| \mathbb{E} F \left( \int_{E_1}^{E_2} dy \text{Im} m^{(\nu)}(y + i\eta) \right) - \mathbb{E} F \left( \int_{E_1}^{E_2} dy \text{Im} m^{(\omega)}(y + i\eta) \right) \right| \leq CN^{-1/6+C\varepsilon}
$$

(17.10) for some constant $C$ and large enough $N$ depending only on $C_1$, $\vartheta$, $\delta_\pm$ and $C_0$ (in (14.26)).
Note that the $N$ times the integration gives a factor $N|E_1 - E_2| \sim N^{1/3+\varepsilon}$ on the left hand side which compensates for the "natural" size of the imaginary part of the Stieltjes transform obtained from the local semicircle law and makes the argument of $F$ to be of order one. The bound (17.10) shows that the distributions of this order one quantity with respect to the $\nu$ and $w$ ensembles are asymptotically the same.

Assuming that Theorem 17.4 holds, we now prove Theorem 17.1.

**Proof of Theorem 17.1.** By the rigidity of eigenvalues in the form of (11.25), we know that

$$|\lambda_N - 2| \leq N^{-2/3+\varepsilon}, \quad N(2 - CN^{-2/3+\varepsilon}, 2 + CN^{-2/3+\varepsilon}) \leq CN^\varepsilon$$

with very high probability. Thus we can assume that $|s| \leq N^\varepsilon$ holds for the parameter $s$ in Theorem 17.1; the statement is trivial for other values of $s$. We define $E := 2 + sN^{-2/3}$. With the left side of (17.6), for any sufficiently small $\varepsilon > 0$, we have

$$\mathbb{E}_\nu F(\text{Tr } \chi_{E-\ell} * \theta_\eta(H)) \leq \mathbb{P}_w(\mathcal{N}(E, \infty) = 0)$$

(17.11)

with the choice

$$\ell := \frac{1}{2}, \quad \eta := N^{-2/3-9\varepsilon}.$$

The bound (17.10) applying to the case $E_1 = E - \ell$ and $E_2 = E_+$ shows that there exist $\delta > 0$, for sufficiently small $\varepsilon > 0$, such that

$$\mathbb{E}_\nu F(\text{Tr } \chi_{E-\ell} * \theta_\eta(H)) \leq \mathbb{E}_w F(\text{Tr } \chi_{E-\ell} * \theta_\eta(H)) + N^{-\delta}$$

(17.12)

(note that $9\varepsilon$ plays the role of the $\varepsilon$ in the Green function comparison theorem). Then applying the right side of (17.6) to the l.h.s of (17.12), we have

$$\mathbb{E}_\nu F(\text{Tr } \chi_{E-2\ell, \infty}) = 0) \leq \mathbb{E}_w F(\text{Tr } \chi_{E-\ell} * \theta_\eta(H)) + CN^{-D}.$$

(17.13)

Combining these inequalities, we have

$$\mathbb{E}_w F(\text{Tr } \chi_{E-2\ell, \infty} = 0) \leq \mathbb{E}_w F(\text{Tr } \chi_{E-\ell} * \theta_\eta(H)) + 2N^{-\delta}$$

(17.14)

for sufficiently small $\varepsilon > 0$ and sufficiently large $N$. Recalling that $E = 2 + sN^{-2/3}$, this proves the first inequality of (17.2) and, by switching the role of $\nu, w$, the second inequality of (17.2) as well. This completes the proof of Theorem 17.1.

**Proof of Theorem 17.4.** We follow the notations and the setup in the proof of Theorem 16.1. First we prove a simpler version of (17.10) with $F(x) = x$ and without integration. Namely, we show that for any $E$ with $|E - 2| \leq N^{-2/3+\varepsilon}$ we have

$$N\eta|\text{Em}_N^{(x)}(z) - \text{Em}_N^{(w)}(z)| = N\eta|\mathbb{E}_N \frac{1}{\text{Tr } H^{(\nu)} - z} - \mathbb{E}_N \frac{1}{\text{Tr } H^{(w)} - z}| = CN^{-1/6+C\varepsilon}, \quad z = E + i\eta.$$  

(17.15)

The prefactor $N\eta$ accounts for the $N$ times the integration from $E_1$ to $E_2$ in (17.10) since $|E_1 - E_2| \leq C\eta$.

We write

$$\mathbb{E}_N \frac{1}{\text{Tr } H^{(\nu)} - z} - \mathbb{E}_N \frac{1}{\text{Tr } H^{(w)} - z} = \sum_{\gamma=1}^{\gamma^{(N)}} \left[ \mathbb{E}_N \frac{1}{\text{Tr } H^{(\gamma)} - z} - \mathbb{E}_N \frac{1}{\text{Tr } H^{(\gamma-1)} - z} \right],$$

(17.16)

with $R := (Q - z)^{-1}$, $S := (H_\gamma - z)^{-1}$ and

$$\frac{1}{N} \text{Tr } S = \hat{R} + \xi_\nu, \quad \xi_\nu := \sum_{m=1}^{4} N^{-m/2} \hat{R}_\nu^{(m)} + N^{-5/2} \Omega_\nu$$

(17.17)
with
\[ \hat{R} := \frac{1}{N} \text{Tr} R, \quad \hat{R}^{(m)} := (-1)^m \frac{1}{N} \text{Tr}(RV)^m R, \quad \Omega_{\gamma} := \frac{1}{N} \text{Tr}(RV)^3 S. \] (17.18)

All these quantities depend on \( \gamma \), equivalently on the index pair \((i, j)\) with \( \phi(i, j) = \gamma \) (see the proof of Theorem 16.1), i.e., \( \hat{R} = \hat{R}(\gamma) \) etc., but we will omit this fact from the notation. We also define the same quantities for \( \nu \) replaced with \( \nu \). By the moment matching condition (17.1),
\[ \mathbb{E} \hat{R}^{(m)}_{\gamma} = \mathbb{E} \hat{R}^{(m)}_{\nu}, \quad m = 0, 1, 2, \]
so we can consider the \( m = 3, 4 \) terms only in the summation in (17.17). Since \( |E - 2| \leq N^{-2/3 + \varepsilon} \) and \( \eta = N^{-2/3 - \varepsilon} \), the strong local semicircle law (6.32) implies that
\[ |R_{ij}(z) - \delta_{ij} m_{sc}(z)| \leq \sqrt{\frac{\text{Im} m_{sc}(z)}{N\eta}} + \frac{1}{N\eta} \leq N^{-1/3 + C\varepsilon}, \] (17.19)
and similar bound holds for \( S \). In particular, the off-diagonal terms of the Green functions are of order \( N^{-1/3 + C\varepsilon} \) and the diagonal terms are bounded with high probability. As a crude bound, we immediately obtain that \( |\hat{R}^{(m)}_{\nu}|, |\Omega_{\nu}| \leq C^{\varepsilon} \) with high probability.

Hence we have
\[ \sum_{\gamma=1}^{\gamma(N)} (N\eta) \left| \frac{1}{N} \text{Tr} \frac{1}{H_{\gamma} - z} - \frac{1}{N} \text{Tr} \frac{1}{H_{\gamma-1} - z} \right| \leq \sum_{m=3}^{4} N^2 (N\eta) N^{-m/2} \max_{\gamma} \mathbb{E} \left| \hat{R}^{(m)}_{\gamma} - \hat{R}^{(m)}_{\nu} \right| + \eta N^{1/2 + C\varepsilon}. \] (17.20)

Given a fixed \( \gamma \) with \( \phi(i, j) = \gamma \), we have the explicit formula
\[ \hat{R}^{(m)}_{\gamma} = \hat{R}^{(m)}_{\nu}(\gamma) = (-1)^m \frac{1}{N} \sum_{k} \sum_{l=1}^{m} \sum_{(\alpha_k, \beta_l) = (i, j)} \left[ R_{k\alpha_1 \beta_1} v_{\alpha_1 \beta_1} R_{\beta_1 \beta_2} \ldots v_{\alpha_m \beta_m} R_{\beta_m \beta_l} \right]. \] (17.21)

First we discuss the generic case, when all indices are different, in particular \( i \neq j \), and later we comment on the case of the coinciding indices. Notice that generically, the first and the last resolvents \( R_{k\alpha_1}, R_{\beta_m \beta_l} \) are off-diagonal terms, thus their size is \( N^{-1/3 + C\varepsilon} \). Every other factor in (17.21) is \( O(N^\varepsilon) \). Hence the generic contributions to the sum (17.20) give
\[ N^2 (N\eta) N^{-m/2} \hat{R}^{(m)}_{\nu} \leq N^{2 - m/2} N^{-1/3 + C\varepsilon}. \] (17.22)
For \( m = 4 \) this is \( N^{-1/3 + C\varepsilon} \), so it is negligible. For \( m = 3 \), we have
\[ N^2 (N\eta) N^{-m/2} \hat{R}^{(3)}_{\nu} \leq N^{1/6 + C\varepsilon}, \] (17.23)
so the straightforward power counting is not sufficient. Our goal is to show that the expectation of this term is in fact much smaller. To see this, we can assume that on the right hand side of (17.21), all the Green functions except the first and the last one are diagonal since otherwise we have an additional (third) off-diagonal term, which yields an extra factor \( N^{-1/3 + C\varepsilon} \), which would be sufficient if combined with the trivial power counting yielding (17.23). If both intermediate Green functions in (17.21) are diagonal, then we can replace each of them with \( m_{sc} \) using the strong local semicircle law (17.19) at the expense of a multiplicative error \( N^{-1/3 + C\varepsilon} \). This improves the previous power counting of order \( N^{1/6 + C\varepsilon} \) to \( N^{-1/6 + C\varepsilon} \) as required in (17.15). Thus we only have to consider the contributions from the terms
\[ \frac{1}{N} \sum_{k} \sum_{(\alpha_k, \beta_l) = (i, j)} m_{sc}^2 \mathbb{E} \left[ R_{k\alpha_1 \beta_1} v_{\alpha_1 \beta_1} v_{\alpha_2 \beta_2} v_{\alpha_3 \beta_3} R_{\beta_3 \beta_l} \right]. \] (17.24)
contributing to \( \hat{R}^{(3)}_{\nu} \), where the summation is restricted to the choices consistent with the fact that the all Green functions in the middle are diagonal, i.e., \( b_1 = a_2, b_2 = a_3 \). Together with the restriction \( \{a_k, b_k\} = \{i, j\} \) and the assumption that all indices are distinct, it yields only two terms:
\[ \frac{1}{N} \sum m_{sc}^2 \mathbb{E} \left[ R_{ki} v_{ij} v_{ij} R_{jk} \right] + \frac{1}{N} \sum m_{sc}^2 \mathbb{E} \left[ R_{kj} v_{ij} v_{ij} R_{ik} \right]. \] (17.25)
Considering the prefactors $N^2(N\eta)N^{-3/2} \leq N^{5/6}$ in (17.23), in order to prove (17.15) we need to show that both terms in (17.25) are bounded by $N^{-1+C\varepsilon}$, i.e., they are at least by a factor $N^{-1/3}$ smaller than the trivial power counting $N^{-2/3+C\varepsilon}$ from the two off-diagonal elements would give.

We will focus on the first term in (17.25). For the last resolvent entry, $R_{jk}$, we use the identity

$$R_{jk} = R_{ik}^{(i)} + \frac{R_{ik} R_{kj}}{R_{ii}}$$

from (8.1). The second factor has already two off-diagonal elements, $R_{ij} R_{ik}$, yielding an additional $N^{-1/3+C\varepsilon}$ factor with high probability (i.e., even without taking the expectation). So we focus on the term

$$\frac{1}{N} \sum_k m_{sc}^2 \mathbb{E} \left[ R_{ki} v_{ij} v_{ji} v_{ij} R_{jk}^{(i)} \right].$$

(17.26)

Recall the identity (8.2)

$$R_{ki} = -R_{kk} \sum_\ell R_{\ell i}^{(i)} (N^{-1/2} v_{\ell i}).$$

Hence we have that

$$1 \cdot (17.26) = -\frac{1}{N} \sum_k m_{sc}^2 N^{-1/2} \sum_\ell (i) \mathbb{E} \left[ R_{kk} R_{\ell i}^{(i)} v_{\ell i} v_{ij} v_{ji} v_{ij} R_{jk}^{(i)} \right].$$

(17.27)

We may again replace the diagonal term $R_{kk}$ with $m_{sc}$ at a negligible error and we are left with

$$-m_{sc}^3 N^{-3/2} \sum_k \sum_\ell (i) \mathbb{E} \left[ R_{\ell i}^{(i)} v_{\ell i} v_{ij} v_{ji} v_{ij} R_{jk}^{(i)} \right].$$

(17.28)

The expectation with respect to the variable $v_{\ell i}$ renders this term zero, unless $\ell = j$, in which case we gain an additional factor $N^{-1/3+C\varepsilon}$ in the power counting compared with (17.26). So the contribution of the last displayed expression to (17.23) is improved from $N^{1/6+C\varepsilon}$ to $N^{-1/3+C\varepsilon}$.

The argument so far assumed that all indices are distinct. In case of each coinciding index we gain a factor $1/N$ from the restriction in the summation and one off-diagonal element, accounted for as a factor $N^{-1/3+C\varepsilon}$, may be "lost" in a sense that it may become diagonal. The total balance of a coinciding index in the power counting is thus a factor $N^{-2/3}$, so we conclude that terms with coinciding indices are negligible. This proves the simplified version (16.5) of Theorem 17.4.

To prove (17.10), for simplicity we ignore the integration and we prove only that

$$\left| \mathbb{E} F \left( N\eta \text{Im } m^{(v)}(z) \right) - \mathbb{E} F \left( N\eta \text{ Im } m^{(w)}(z) \right) \right| \leq CN^{-1/6+C\varepsilon}, \quad z = E + i\eta,$$

(17.29)

for any $E$ with $|E - 2| \leq N^{-2/3+\varepsilon}$. The $\eta$ factor, up to an irrelevant $N^{2\varepsilon}$ accounts for the integration domain of length $|E_2 - E_1| \leq N^{-2/3+\varepsilon}$ in the arguments in (17.10). The general form (17.10) follows in the same way. We point out that (17.29) with $F(N\eta \text{Im } m(z))$ replaced by $F(N\eta m(z))$ also holds with the same argument given below.

Taking into account the telescopic summation over all $\gamma$’s, for the proof of (17.29) it is sufficient to prove that

$$N^2 \left| \mathbb{E} F \left( \frac{N\eta}{N} \text{Im Tr } \frac{1}{H - z} \right) - \mathbb{E} F \left( \frac{N\eta}{N} \text{ Im Tr } \frac{1}{H - z} \right) \right| \leq CN^{-1/6+C\varepsilon},$$

(17.30)

and then sum it up for all $\gamma = \phi(i, j)$. As before, we assume the generic case, i.e., $i \neq j$.

We Taylor expand the function $F$ around the point

$$\Xi := N\eta \text{Im } m^{(R)}(z), \quad z = E + i\eta,$$
where $m^{(R)} = \frac{1}{N} \text{Tr} R$. Notice that $m^{(R)}$ is independent of $v_{ij}$ and $w_{ij}$. Recalling the definition of

$$\xi_v = \sum_{m=1}^{4} N^{-m/2} \hat{R}_v^{(m)} + N^{-5/2} \Omega_v$$  \hspace{1cm} (17.31)$$

from (17.17), and a similar definition for $\xi_w$, we obtain

$$\left| EF\left( N\eta \frac{1}{N} \text{Im} \text{Tr} \frac{1}{H_\gamma - z} \right) - EF\left( N\eta \frac{1}{N} \text{Im} \text{Tr} \frac{1}{H_{\gamma - 1} - z} \right) \right| \leq EF'(\Xi) \left( N\eta [\xi_v - \xi_w] \right) + \frac{1}{2} EF''(\Xi) \left( [N\eta \xi_v]^2 - [N\eta \xi_w]^2 \right) + \ldots .$$

We now substitute (17.31) into this expression. By the naive power counting

$$(N\eta) N^{-m/2} |\hat{R}_v^{(m)}| \leq N^{-m/2 - 1/3 + C\varepsilon}, \quad (N\eta) N^{-5/2} |\Omega_v| \leq N^{-13/6 + C\varepsilon}$$

from the previous proof. The contributions of all the $\Omega$ error terms as well as the $m = 4$ terms clearly satisfy the aimed bound (17.30) and thus can be neglected.

First, we prove that the quadratic term (as well as all higher order terms) in the Taylor expansion (17.32) is negligible. Here the cancellation between the $v$ and $w$ contributions is essential. It is therefore sufficient to consider

$$(N\eta)^2 \sum_{m,m'=1}^{3} N^{-m/2 - m'/2} \left[ \hat{R}_v^{(m)} \hat{R}_v^{(m')} - \hat{R}_w^{(m)} \hat{R}_w^{(m')} \right]$$

instead of $[N\eta \xi_v]^2 - [N\eta \xi_w]^2$. Since $F''$ is bounded, any term with $m + m' \geq 4$ is negligible similarly to the previous by power counting. The contribution of the $m = m' = 1$ term to the quadratic term in (17.32) is explicitly given by

$$EF''(\Xi) \left( \frac{1}{N} \sum_k \left[ R_{ki} v_{ij} R_{jk} + R_{kj} v_{ji} R_{ik} \right] \right) \left( \frac{1}{N} \sum_{k'} \left[ R_{k'i} v_{ij} R_{jk'} + R_{kj} v_{ji} R_{ik'} \right] \right) = EF''(\Xi) \left( v \leftrightarrow w \right).$$

Since $R$ and $\Xi$ are independent of $v_{ij}$ and $w_{ij}$, the expectations with respect to $v_{ij}$ and $w_{ij}$ can be computed. Since their second moments coincide, the two expectations above exactly cancel each other.

We are left with the $m + m' = 3$ terms, we may assume $m = 2$ and $m' = 1$. We do not expect any more cancellation from the $v$ and $w$ terms, so we consider them separately. A typical such term contributing to (17.30) with all the prefactors is of the form

$$N^2 (N\eta)^2 N^{-3/2} F''(\Xi) \left( \frac{1}{N} \sum_k \left[ R_{ki} v_{ij} R_{jj} v_{ji} R_{ik} \right] \right) \left( \frac{1}{N} \sum_{k'} R_{k'i} v_{ij} R_{jk'} \right).$$

The key point is that generically there are at least four off-diagonal elements, $R_{ki}$, $R_{ik}$, $R_{k'i}$ and $R_{jk'}$ (we chose the "worst" term, where the resolvent $R_{jj}$ in the middle is diagonal). So together with the boundedness of $F''$, the direct power counting gives

$$N^2 (N\eta)^2 N^{-3/2} (N^{-1/3 + C\varepsilon})^4 = N^{-1/6 + C\varepsilon},$$

which is exactly the precision required in (17.30). In case of each coinciding index (for example $k = i$), two off-diagonal elements may be "lost", but we gain a factor $1/N$ from the restriction of the summation. So terms with coinciding indices are negligible, as before. Here we discussed only the quadratic term in the Taylor expansion (17.32), but it is easy to see that higher order terms are even smaller, without any cancellation effect.

So we are left with estimating the linear term in (17.32), i.e., we need to bound

$$N^2 (N\eta) \sum_{m=1}^{3} N^{-m/2} EF'(\Xi) \left[ \hat{R}_v^{(m)} - \hat{R}_w^{(m)} \right],$$

148
where we already took into account that the $m = 4$ term as well as the $\Omega_v$ error term from (17.31) are negligible by direct power counting. If $F'(\Xi)$ were deterministic, then the previous argument leading to (17.15) would also prove the $N^{-1/6+C\varepsilon}$ bound for the linear term in (17.32). In fact, the exact cancellation between the expectations of the $R_{\nu}^{(m)}$ and $R_{\nu}^{(m)}$ terms for $m = 1$ and $m = 2$ relied only on computing the expectation with respect to $v_{ij}$ and $w_{ij}$ and on the matching of the first and second moments. Since $F'(\Xi)$ is independent of $v_{ij}$ and $w_{ij}$, this argument remains valid even with the $F'(\Xi)$ factor included. We thus need to control the $m = 3$ term, i.e., show that

$$N^2(N\eta)N^{-3/2}\varepsilon F'(\Xi)R_{\nu}^{(3)} \leq N^{-1/6+C\varepsilon}. \quad (17.33)$$

The same bound then would also hold if $v$ is replaced with $w$.

Using the boundedness of $F'$ and the argument between (17.24) and (17.28) in the previous proof, we see that for (17.33) it is sufficient to show that

$$\left| N^2(N\eta)N^{-3/2}N^{-3/2} \sum_k \sum_{\ell} \mathbb{E}F'(\Xi) \left[ R_{k\ell}^{(i)} v_{\ell i} v_{\ell j} v_{ij} v_{ij} R_{j\ell}^{(i)} \right] \right| \leq N^{-1/6+C\varepsilon}. \quad (17.34)$$

Without the $F'(\Xi)$ term, the expectation with respect to $v_{\ell i}$ would render this term zero, in the generic case when $\ell \neq j$. In order to exploit this effect we need to expand $F'(\Xi)$ in the single variable $v_{\ell i}$.

Fix an index $\ell \neq j$ and let $Q'$ be the matrix identical to $Q$ except the $(\ell, i)$ and $(i, \ell)$ matrix elements are zero, and let $T = T' = (Q' - z)^{-1}$ be its resolvent. Clearly, $T$ is independent of $v_{ij}$ and $v_{\ell i}$ and the local law (17.19) holds for the matrix elements of $T$ as well. Using the resolvent expansion, we find

$$\Xi = (N\eta) \mathbb{I} \frac{1}{N} \sum_n \left[ T_{nn} + N^{-1/2}(T_n v_{\ell i} T_n + T_{ni} v_{\ell i} T_{ln} + \ldots) \right], \quad (17.35)$$

where the higher order terms carry a prefactor $N^{-1}$. We Taylor expand $F'(\Xi)$ around

$$\Xi_0 := (N\eta) \mathbb{I} \frac{1}{N} \sum_n T_{nn},$$

and we obtain

$$F'(\Xi) = F'(\Xi_0) + \frac{1}{2} F''(\Xi_0)(N\eta) \mathbb{I} \left[ \frac{1}{N^{3/2}} \sum_n \left[ T_{n\ell} v_{\ell i} T_{ni} + T_{ni} v_{\ell i} T_{ln} + \ldots \right] \right] + \ldots. \quad (17.36)$$

Since $\Xi_0$ is independent of $v_{\ell i}$, when we insert the above expansion for $F'(\Xi)$ into (17.34), the contribution of the leading term $F'(\Xi_0)$ gives zero after taking the expectation for $v_{\ell i}$.

The two subleading terms in (17.36), that are linear in $v_{\ell i}$, have generically two off-diagonal elements, $T_{n\ell}$ and $T_{ni}$ (or $T_{ni}$ and $T_{ln}$) which have size $N^{-1/3+C\varepsilon}$ each. So the contribution of this term to (17.34) by simple power counting is

$$N^2(N\eta)N^{-3/2}N^{-3/2}NN(N\eta)N^{-1/2}(N^{-1/3+C\varepsilon})^4 \leq N^{-1/6+C\varepsilon}.$$ 

Similar argument shows that the higher order terms in the Taylor expansion (17.36) as well as higher order terms in the resolvent expansion (17.35) have a negligible contribution.

As before, we presented the argument for generic indices; coinciding indices give smaller contributions as we explained before. We leave the details to the reader. This completes the proof of (17.30) and thus the proof of Theorem 17.4.
18 Further results and historical notes

Spectral statistics of large random matrices have been studied from many different perspectives. The main focus of this book was on one particular result: the Wigner-Dyson-Mehta bulk universality for Wigner matrices in the average energy sense (Theorem 5.1). In this last section we collect a few other recent results and open questions of this very active research area. We also give references to related results and summarize the history of the development.

18.1 Wigner matrices: bulk universality in different senses

Theorem 5.1, which is also valid for complex Hermitian ensembles, settles the average energy version of the Wigner-Dyson-Mehta conjecture for generalized Wigner matrices under the moment assumption (5.6) on the matrix elements. Theorem 5.1 was proved in increasing order of generality in [63], [69], [68] and [70]; see also [130] for the complex Hermitian Wigner ensemble and for a restricted class of real symmetric matrices (see comments about this class later on). The following theorem reduces the moment assumption to 4 + $\varepsilon$ moments. The proof in [53] also provides an effective speed of convergence in (18.2). We also point out that the condition (2.6) can be somewhat relaxed, see Corollary 8.3 [55].

Theorem 18.1 (Universality with averaged energy). [53, Theorem 7.2] Suppose that $H = (h_{ij})$ is a complex Hermitian (respectively, real symmetric) generalized Wigner matrix. Suppose that for some constants $\varepsilon > 0$, $C > 0$,

$$\mathbb{E} \left| \sqrt{Nh_{ij}} \right|^{4+\varepsilon} \leq C.$$

(18.1)

Let $n \in \mathbb{N}$ and $O : \mathbb{R}^n \to \mathbb{R}$ be a test function (i.e., compactly supported and continuous). Fix $|E_0| < 2$ and $\xi > 0$, then with $b_N = N^{-1+\xi}$ we have

$$\lim_{N \to \infty} \int_{E_0-b_N}^{E_0+b_N} \frac{dE}{2b_N} \int_{\mathbb{R}^n} d\alpha \, O(\alpha) \frac{1}{\varrho_{sc}(E)^n} \left( p_N^{(n)} - p_{G,N}^{(n)} \right) \left( E + \frac{\alpha}{N\varrho_{sc}(E)} \right) = 0. \tag{18.2}$$

Here $\varrho_{sc}$ is the semicircle law defined in (3.1), $p_N^{(n)}$ is the n-point correlation function of the eigenvalue distribution of $H$ (4.20), and $p_{G,N}^{(n)}$ is the n-point correlation function of an $N \times N$ GUE (respectively, GOE) matrix.

A relatively straightforward consequence of Theorem 18.1 is the average gap universality formulated in the following corollary. The gap distributions of the comparison ensembles, i.e., the Gaussian ensembles, can be explicitly expressed in terms of a Fredholm determinant [37,40]. Recall the notation $[A,B] := \mathbb{Z} \cap [A,B]$ for any real numbers $A < B$.

Corollary 18.2 (Gap universality with averaged label). Let $H$ be as in Theorem 18.1 and $O$ be a test function of $n$ variables. Fix small positive constants $\xi, \alpha > 0$. Then for any integer $j_0 \in [\alpha N, (1-\alpha)N]$ we have

$$\lim_{N \to \infty} \frac{1}{2N^\xi} \sum_{|j-j_0| \leq N^\xi} \left[ \mathbb{E} - \mathbb{E}^G \right] O(N(\lambda_j - \lambda_{j+1}), N(\lambda_j - \lambda_{j+2}), \ldots, N(\lambda_j - \lambda_{j+n})) = 0. \tag{18.3}$$

Here $\lambda_j$'s are the ordered eigenvalues and for brevity we omitted the rescaling factor $g(\lambda_j)$ from the argument of $O$. Moreover, $\mathbb{E}$ and $\mathbb{E}^G$ denote the expectation with respect to the Wigner ensemble $H$ and the Gaussian (GOE or GUE) ensemble, respectively.

The next result shows that Theorem 18.1 holds under the same conditions without energy averaging.

Theorem 18.3 (Universality at fixed energy). [26] Consider a complex Hermitian (respectively, real symmetric) generalized Wigner matrix with moment condition (18.1). Then for any $E$ with $|E| < 2$ we have

$$\lim_{N \to \infty} \int_{\mathbb{R}^n} d\alpha \, O(\alpha) \frac{1}{\varrho_{sc}(E)^n} \left( p_N^{(n)} - p_{G,N}^{(n)} \right) \left( E + \frac{\alpha}{N\varrho_{sc}(E)} \right) = 0. \tag{18.4}$$
The fixed energy universality (18.4) for the $\beta = 2$ (complex Hermitian) case was already known earlier, see [57, 66, 131]. This case is exceptional since the Harish-Chandra/Itzykson/Zuber identity allows one to compute correlation functions for Gaussian divisible ensembles. This method relies on an algebraic identity and has no generalization to other symmetry classes.

Finally, the gap universality with fixed label asserts that (18.3) holds without averaging.

**Theorem 18.4** (Gap universality with fixed label). [67, Theorem 2.2] Consider a complex Hermitian (respectively, real symmetric) generalized Wigner matrix and assume subexponential decay of the matrix elements instead of (18.1). Then Corollary 18.2 holds without averaging:

$$\lim_{N \to \infty} [E - E^G] O\left(N(\lambda_j - \lambda_{j+1}), N(\lambda_j - \lambda_{j+2}), \ldots, N(\lambda_j - \lambda_{j+n})\right) = 0, \quad (18.5)$$

for any $j \in [\alpha N, (1 - \alpha) N]$ with a fixed $\alpha > 0$. More generally, for any $k, m \in [\alpha N, (1 - \alpha) N]$ we have

$$\lim_{N \to \infty} \left| [E - E^G] O\left((N\varrho_k)(\lambda_k - \lambda_{k+1}), (N\varrho_k)(\lambda_k - \lambda_{k+2}), \ldots, (N\varrho_k)(\lambda_k - \lambda_{k+n})\right) - [E - E^G] O\left((N\varrho_m)(\lambda_m - \lambda_{m+1}), \ldots, (N\varrho_m)(\lambda_m - \lambda_{m+n})\right) \right| = 0, \quad (18.6)$$

where the local density $\varrho_k$ is defined by $\varrho_k := \varrho_{sc}(\gamma_k)$ with $\gamma_k$ from (11.31).

The second part (18.6) of this theorem asserts that the gap distribution is not only independent of the specific Wigner ensemble, but it is also universal throughout the bulk spectrum. This is the counterpart of the statement that the appropriately rescaled correlation functions (18.4) have a limit that is independent of $E$, see (4.38).

Prior to [67], universality for a single gap was only achieved in the special case of the Gaussian unitary ensemble (GUE) in [128]; this statement then immediately implies the same result for complex Hermitian Wigner matrices satisfying the four moment matching condition.

### 18.2 Historical overview of the three-step strategy for bulk universality

We summarize the recent history related to the bulk universality results in Section 18.1. The three step strategy first appeared implicitly in [57] in the context of complex Hermitian matrices and was conceptualized for all symmetry classes in [63] by linking it to the DBM. Several key components of this strategy appeared in the later work [63, 64, 69]. We will review the history of Steps 1-3 separately.

**History of Step 1.** The semicircle law was proved by Wigner for energy windows of order one [137]. Various improvements were made to shrink the spectral windows; in particular, results down to scale $N^{-1/2}$ were obtained by combining the results of [11] and [79]. The result at the optimal scale, $N^{-1}$, referred to as the local semicircle law, was established for Wigner matrices in a series of papers [60–62]. The method was based on a self-consistent equation for the Stieltjes transform of the eigenvalues, $m(z)$, and the continuity in the imaginary part of the spectral parameter $z$. As a by-product, the optimal eigenvector delocalization estimate was proved. In order to deal with the generalized Wigner matrices, one needed to consider the self-consistent equation of the vector $(G_{ii}(z))_{i=1}^{N}$, the diagonal matrix elements of the Green function, since in this case there is no closed equation for $m(z) = N^{-1} \text{Tr } G(z)$ [68, 69]. Together with the fluctuation averaging lemma, this method implied the optimal rigidity estimate of eigenvalues in the bulk in [68] and up to the edges in [70]. Furthermore optimal control was obtained for individual matrix elements of the resolvent $G_{ij}$ and not only for its trace. The estimate on $G_{ii}$ also provided a simple alternative proof of the eigenvector delocalization estimate. A comprehensive summary of these results can be found in [55]. Several further extensions concern weaker moment conditions and improvements of $(\log N)$-powers, see e.g. [32, 77, 129] and references therein.

**History of Step 2.** The universality for Gaussian divisible ensembles was proved by Johansson [85] for complex Hermitian Wigner ensembles in a certain range of the bulk spectrum. This was the first result beyond
exact Gaussian calculations going back to Gaudin, Dyson and Mehta. It was later extended to complex
sample covariance matrices on the entire bulk spectrum by Ben Arous and Pécé [19]. There were two major
restrictions of this method: (i) the Gaussian component was required to be of order one independent of \( N \);
(ii) it relies on an explicit formula by Brézin-Hikami [30,31] for the correlation functions of eigenvalues which
is valid only for Gaussian divisible ensembles with unitary invariant Gaussian component. The size of the
Gaussian component was reduced to \( N^{-1+\varepsilon} \) in [57] by using an improved formula for correlation functions
and the local semicircle law from [60–62].

To eliminate the usage of an explicit formula, a conceptual approach for Step 2 via the local ergodicity
of Dyson Brownian motion was initiated in [63]. In [64], a general theorem for the bulk universality was
formulated which applies to all classical ensembles, i.e., real and complex Wigner matrices, real and complex
sample covariance matrices and quaternions Wigner matrices. The relaxation time to local equilibrium proved
in these two papers were not optimal; the optimal relaxation time, conjectured by Dyson, was obtained later
in [70]. The DBM approach in these papers yields the average gap and hence the average energy universality,
while the method via the Brézin-Hikami formula gives universality in a fixed energy sense, but is strictly
restricted to the complex Hermitian case. The fixed energy universality for all symmetry classes was recently
achieved in [26]. The method in this paper was still based on DBM, but the analysis of DBM was very different
from the one discussed in this book.

**History of Step 3.** Once universality was established for Gaussian divisible ensembles \( e^{-t/2}H_0 + (1
- e^{-t})^{1/2}H^G \) (here we used the convention used in Chapter 5) for a sufficiently large class of \( H_0 \) and sufficiently
small \( t \), the final step is to approximate the local eigenvalue distribution of a general Wigner matrix by that
of a Gaussian divisible one. The first approximation result was obtained via the reversal heat flow in [57]
which required smoothness of the distribution of matrix elements. Shortly after, Tao and Vu [130] proved the
four moment theorem which asserts that the local statistics of two Wigner ensembles are the same provided
that the first four moments of their matrix elements are identical. The third comparison method is the
Green function comparison theorem, Theorem 16.1 [63,68,69], which was the main content of Section 16.
It uses similar moment conditions as those appeared in [130], but it provides detailed information on the
matrix elements of Green functions as well. Both [130] and the proof of the Green function comparison
theorem [63,68,69] use the local semicircle law [60,62] and the Lindeberg’s idea, which had appeared in
the proof of the Wigner semicircle law by Chatterjee [34]. The approach [130] requires additional difficult
estimates on level repulsions, while Theorem 16.1 follows directly from the estimates on the matrix elements
of the Green function from Step 1 via standard resolvent expansions. Another comparison method reviewed
in this book, Theorem 15.2, can be viewed as a variation of the GFT.

In summary, the WDM conjecture was first solved in the complex Hermitian case and then for the real
symmetric case. In the first case, the bulk universality for Hermitian matrices with smooth distribution was
proved in [57]. The smoothness requirement was removed in [130] but the complex Bernoulli distribution
was still excluded; this was finally added in [58], see also [66,131] for related discussions.

The result of [130] also implies that bulk universality of symmetric Wigner matrices under the restriction
that the first four moments of the matrix elements match those of GOE. The major roadblock to the WDM
conjecture for real symmetric matrices was the lack of Step 2 in this case. For symmetric matrices, we do
not know any Brézin-Hikami type formula, which was the backbone for the Step 2 in the Hermitian case.
The resolution of this difficulty was to abandon any explicit formula and to prove Dyson’s conjecture on
the relaxation of the Dyson Brownian Motion, see Step 2. Together with the Green function comparison
theorem [63], bulk universality was proved for real symmetric matrices without any smoothness requirement.
In particular, this result covered the real Bernoulli case [68] that has been a major objective in this subject
due to the application to the adjacency matrices for random graphs. The flexibility of the DBM approach
also allowed one to establish universality beyond the traditional Wigner ensembles. The first such result was
bulk universality for generalized Wigner matrices [69], followed by many other models, see Section 18.6.

Finally, the technical condition assumed in all these papers, i.e., that the probability distributions of the
matrix elements decay subexponentially, was reduced to the \( (4+\varepsilon) \)-moment assumption (18.1) by using the
universality of Erdős-Rényi matrices [53]. This yielded the bulk universality as formulated in Theorem 18.1.
18.3 Invariant ensembles and log-gases

The explicit formula (4.3) for the joint density function of the eigenvalues for the invariant matrix ensemble defined in (4.2) gives rise to a direct statistical physics interpretation: it can be viewed as a measure $\mu^{(N)}$ on $N$ point particles on the real line. Inspired by the Gibbs formalism, we may write it as

$$
\mu^{(N)}(d\lambda) = p_N(\lambda)d\lambda = \text{const} \cdot e^{-\beta NH(\lambda)}d\lambda, \quad \lambda = (\lambda_1, \ldots, \lambda_N),
$$

where

$$
\mathcal{H}(\lambda) := \frac{1}{N} \sum_{j=1}^{N} \frac{1}{2} V(\lambda_j) - \frac{1}{N} \sum_{1 \leq i < j \leq N} \log(\lambda_j - \lambda_i).
$$

Clearly, $\mathcal{H}$ can be interpreted as the classical Hamiltonian function of a log-gas, i.e., a system of $N$ interacting particles subject to a potential $V$ and a two-body logarithmic interaction. While a direct relation to invariant ensembles exists only for specific values $\beta = 1, 2, 4$, log-gases may be studied for any $\beta > 0$, i.e., at any inverse temperature.

In the case of invariant ensembles, it is well-known that for $V$ satisfying certain mild conditions the sequence of one-point correlation function, or density, associated with (18.7) has a limit as $N \to \infty$ and the limiting equilibrium density $\varrho_V(s)$ can be obtained as the unique minimizer of the functional

$$
I(\nu) = \int_{\mathbb{R}} V(t)\nu(t)dt - \int_{\mathbb{R}} \int_{\mathbb{R}} \log |t-s|\nu(s)\nu(t)dt ds.
$$

This is the analogue of the Wigner semicircle law for log-gases.

We assume that $\varrho = \varrho_V$ is supported on a single compact interval, $[A, B]$ and $\varrho \in C^2(A, B)$. Moreover, we assume that $V$ is regular in the sense that $\varrho$ is strictly positive on $(A, B)$ and vanishes as a square root at the endpoints, i.e.,

$$
\varrho(t) = s_A \sqrt{t-A} \left(1 + O(t-A)\right), \quad t \to A^+,
$$

for some constant $s_A > 0$ and a similar condition holds at the upper edge.

It is known that these conditions are satisfied if, for example, $V$ is strictly convex. In this case $\varrho_V$ satisfies the equation

$$
\frac{1}{2} V'(t) = \int_{\mathbb{R}} \frac{\varrho_V(s)ds}{t-s}
$$

for any $t \in (A, B)$. For the Gaussian case, $V(x) = x^2/2$, the equilibrium density is given by the semicircle law, $\varrho_V = \varrho_{sc}$, see (3.1).

Under these conditions, the density of $\mu^{(N)}$ converges to $\varrho_V$ even locally down to the optimal scale and rigidity analogous to Theorem 11.5 holds:

$$
\mathbb{P}^{\mu^{(N)}}\left(|\lambda_j - \gamma_{j,V}| \geq N^{-\frac{2}{3} + \xi} \min\{k, N-k+1\}^{-\frac{1}{3}}\right) \leq e^{-N^c},
$$

where the quantiles $\gamma_{j,V}$ of the density $\varrho_V$ are defined by

$$
\frac{j}{N} = \int_{A}^{\gamma_{j,V}} \varrho_V(x)dx,
$$

similarly to (11.31), see Theorem 2.4 [24].

The higher order correlation functions, given in Definition 4.2 and rescaled by the local density $\varrho_{\beta,V}(E)$ around an energy $E$ in the bulk, exhibit a universal behavior in a sense that they depend only on $\beta$ but independent of the potential $V$. Similar result holds at the edges of the support of the density $\varrho_{\beta,V}$. This can be viewed as a special realization of the universality hypothesis for the invariant matrix ensemble. Historically, the universality for the special classical cases $\beta = 1, 2, 4$ had been extensively studied via orthogonal polynomial methods. These methods rely on explicit expressions of the correlation functions in
terms of Fredholm determinants involving orthogonal polynomials. The key advantage of this method is that it yields explicit formulas for the limiting gap distributions or correlation functions. However, these explicit expressions are available only for the specific values $\beta = 1, 2, 4$. Within the framework of this approach, the $\beta = 1, 2, 4$ cases of Theorem 18.5 below were proved for very general potentials for $\beta = 2$ and for analytic potentials with some additional conditions for $\beta = 1, 4$. This is a whole subject by itself and we can only refer the readers to some reviews or books [39,90,115].

The first general result for nonclassical $\beta$ values is the following theorem proved in [23–25]. In these works, a new approach to the universality of $\beta$ ensemble was initiated. It departed from the previously mentioned traditional approach using explicit expressions for the correlation functions. Instead, it relies on comparing the probability measures of the $\beta$ ensembles to those of the Gaussian ones. This basic idea of comparing two probability measures directly was also used in the later works, [118] and [18], albeit in a different way, where Theorem 18.5 were reproved under different sets of conditions on the potential $V$.

**Theorem 18.5** (Universality with averaged energy). Assume $V \in C^4(\mathbb{R})$, regular and let $\beta > 0$. Consider the $\beta$-ensemble $\mu^{(N)} = \mu^{(N)}_{\beta,V}$ given in (18.7) with correlation functions $p^{(n)}_{V,N}$ defined analogously to (4.20). For the Gaussian case, $V(x) = x^2/2$, the correlation functions are denoted by $p^{(n)}_{G,N}$. Let $E_0 \in (A,B)$ lie in the interior of the support of $\varrho$ and similarly let $E_0 \in (-2, 2)$ be inside the support of $\varrho_{sc}$. Then for $b_N = N^{-1+\xi}$ with some $\xi > 0$ we have

$$
\lim_{N \to \infty} \int_{\mathbb{R}^n} d\alpha O(\alpha) \left[ \int_{E_0-b_N}^{E_0+b_N} \frac{dE'}{2b_N} \frac{1}{\varrho(E')} p^{(n)}_{V,N} \left( E + \frac{\alpha}{N \varrho(E)} \right) \right] = 0 \tag{18.11}
$$

i.e., the correlation functions of $\mu^{(N)}_{\beta,V}$ averaged around $E_0$ asymptotically coincide with those of the Gaussian case. In particular, they are independent of $E_0$.

Unversality at fixed energy, i.e., (18.11) without averaging in $E$, was proven by M. Shcherbina in [118] for any $\beta > 0$ under certain regularity assumptions on the potentials. We remark that the gap distribution in the Gaussian case for any $\beta > 0$ can be described by an interesting stochastic process [135].

Theorem 18.5 immediately implies gap universality with averaged labels, exactly in the same way as Corollary 18.2 was deduced from Theorem 18.1. To formulate the gap universality with a fixed label, we recall the quantiles $\gamma_{j,V}$ from (18.10) and we set

$$
\varrho_j^V := \varrho V(\gamma_{j,V}), \quad \text{and} \quad \varrho_j := \varrho_{sc}(\gamma_j) \tag{18.12}
$$

to be the limiting densities at the $j$-th quantiles. Let $E^{\mu_V}$ and $E^G$ denote the expectation w.r.t. the measure $\mu_V$ and its Gaussian counterpart for $V(\lambda) = \frac{1}{2}\lambda^2$.

**Theorem 18.6** (Gap universality with fixed label). [67, Theorem 2.3] Consider the setup of Theorem 18.5 and we also assume $\beta \geq 1$. Set some $\alpha > 0$, then

$$
\lim_{N \to \infty} \left| E^{\mu_V} \left( (N\varrho_k^V)(\lambda_k - \lambda_{k+1}), (N\varrho_k^V)(\lambda_k - \lambda_{k+2}), \ldots, (N\varrho_k^V)(\lambda_k - \lambda_{k+n}) \right) \right. \tag{18.13}

$$

$$
- E^{\mu_G} \left( (N\varrho_m)(\lambda_m - \lambda_{m+1}), \ldots, (N\varrho_m)(\lambda_m - \lambda_{m+n}) \right) \right| = 0
$$

for any $k, m \in [\alpha N, (1-\alpha)N]$. In particular, the distribution of the rescaled gaps w.r.t. $\mu_V$ does not depend on the index $k$ in the bulk.

We point out that Theorem 18.5 holds for any $\beta > 0$, but Theorem 18.6 requires $\beta \geq 1$. This is partly due to that the De Giorgi-Nash-Moser regularity theory was used in [67]. This restriction was later removed by Bekerman-Figalli-Guionnet [18] under a higher regularity assumption on $V$ and some additional hypothesis.
18.4 Universality at the spectral edge

So far we stated results for the bulk of the spectrum. Similar results hold at the edge; in this case the “averaged” results are meaningless. In Theorem 17.1 we already presented a universality result for the largest eigenvalue for the Wigner ensemble. That proof easily extends to the joint distribution of finitely many extreme eigenvalues. The following theorem gives a bit more: they control finite dimensional distributions of the first $N^{1/4}$ eigenvalues and they also identify the limit distribution. We then give the analogous result for log gases.

**Theorem 18.7** (Universality at the edge for Wigner matrices). [24] Let $H$ be a generalized Wigner ensemble with subexponentially decaying matrix elements. Fix $n \in \mathbb{N}$, $\kappa < 1/4$ and a test function $O$ of $n$ variables. Then for any $\Lambda \subset [1, N^s]$ with $|\Lambda| = n$, we have

$$
\left| \mathbb{E} - \mathbb{E}^G \right| O \left( \left( N^{2/3} j^{1/3} (\lambda_j - \gamma_j) \right)_{j \in \Lambda} \right) \leq N^{-\chi},
$$

with some $\chi > 0$, where $\mathbb{E}^G$ is expectation w.r.t. the standard GOE or GUE ensemble depending on the symmetry class of $H$ and $\gamma_j$’s are semicircle quantiles (11.31).

The edge universality for Wigner matrices was first proved by Soshnikov [124] under the assumption that the distribution of the matrix elements is symmetric and has finite moments for all orders. Soshnikov used an elaborated moment matching method and the conditions on the moments are not easy to improve. A completely different method based on the Green function comparison theorem was initiated in [70]. This method is more flexible in many ways and it removes essentially all restrictions in previous works. In particular, the optimal moment condition on the distribution of the matrix elements was obtained by Lee and Yin [97]. All these works assume that the variances of the matrix elements are identical. The main point of Theorem 18.7 is to consider generalized Wigner matrices, i.e., matrices with non-constant variances. It was proved in [70, Theorem 17.1] that the edge statistics for any generalized Wigner matrix coincide with those of a generalized Gaussian Wigner matrix with the same variances, but it was not shown that the statistics are independent of the variances themselves. Theorem 18.7 provides this missing step and thus it proves the edge universality for the generalized Wigner matrices.

**Theorem 18.8** (Universality at the edge for log-gases). [24] Let $\beta \geq 1$ and $V$ (resp. $\tilde{V}$) be in $C^4(\mathbb{R})$, regular such that the equilibrium density $\varrho_V$ (resp. $\varrho_{\tilde{V}}$) is supported on a single interval $[\lambda, \beta]$ (resp. $[\lambda, \beta]$). Without loss of generality we assume that for both densities (18.8) holds with $A = 0$ and with the same constant $s_A$. Fix $n \in \mathbb{N}$, $\kappa < 2/5$. Then for any $\Lambda \subset [1, N^s]$ with $|\Lambda| = n$, we have

$$
\left| \mathbb{E}^{\varrho_V} - \mathbb{E}^{\varrho_{\tilde{V}}} \right| O \left( \left( N^{2/3} j^{1/3} (\lambda_j - \gamma_j) \right)_{j \in \Lambda} \right) \leq N^{-\chi}
$$

with some $\chi > 0$. Here $\gamma_j$ are the quantiles w.r.t. the density $\varrho_V$ (18.10).

The history of edge universality runs in parallel to the bulk one in the sense that most bulk methods were extended to the edge case. Similar to the bulk case, the edge universality was first proved via orthogonal polynomial methods for the classical values of $\beta = 1, 2, 4$ in [38,42,110,116]. The first edge universality results were given independently in [24] (valid for general potentials and $\beta \geq 1$) and, with a completely different method, in [91] (valid for strictly convex potentials and any $\beta > 0$). The later work [18] also covered the edge case and proved universality to $\beta > 0$ under some higher regularity assumption on $V$ and an additional condition that can be checked for convex $V$. Some open questions concern the universal behavior at the non-regular edges where even the scaling may change.

18.5 Eigenvectors

Universality questions have been traditionally formulated for eigenvectors, but they naturally extend to eigenvectors as well. They are closely related to two different important physical phenomena: delocalization and quantum ergodicity.
The concept of delocalization stems from the basic theory of random Schrödinger operators describing a single quantum particle subject to a random potential $V$ in $\mathbb{R}^d$. If the potential decays at infinity, then the operator $-\Delta + V$ acting on $L^2(\mathbb{R}^d)$ has a discrete pure point spectrum with $L^2$-eigenfunctions, localized or bound states, in the low energy regime. In the high energy regime it has absolutely continuous spectrum with bounded but not $L^2$-normalizable generalized eigenfunctions. The latter are also called delocalized or extended states and they correspond to scattering and conductance. If the potential $V$ is a stationary, ergodic random field, then celebrated Anderson localization \cite{AM75} occurs: at high disorder a dense pure point spectrum with exponentially decaying eigenfunctions appears even in the energy regime that would classically correspond to scattering states. The localized regime has been mathematically well understood for both the high disorder regime and the regime of low density of states (near the spectral edges) since the fundamental works of Fröhlich and Spencer \cite{FS79} and later by Aizenman and Molchanov \cite{AM80}. The local spectral statistics is Poisson \cite{P93}, reflecting the intuition that exponentially decaying eigenfunctions typically have no overlap, so their energies are independent.

The low disorder regime, however, is mathematically wide open. In $d \geq 3$ or higher dimensions it is conjectured that $-\Delta + V$ has absolutely continuous spectrum (away from the spectral edges), the corresponding eigenfunctions are delocalized and the local eigenvalue statistics of the restriction of $-\Delta + V$ to a large finite box follow the GOE local statistics. In other words, a phase transition is believed to occur as the strength of the disorder varies; this is called the Anderson metal-insulator transition. Currently even the existence of this extended states regime is not known; this is considered one of the most important open question in mathematical physics. Only the Bethe lattice (regular tree) is understood \cite{A95,GS93} but this model exhibits Poisson local statistics \cite{D93} due to its exponentially growing boundary.

The strong link between local eigenvalue statistics and delocalization for random Schrödinger operators naturally raises the question about the structure of the eigenvectors for an $N \times N$ Wigner matrix $H$. Complete delocalization in this context would mean that any $\ell^2$ normalized eigenvector $u = (u_1, \ldots, u_N)$ of $H$ is substantially supported on each coordinate; ideally $|u_i|^2 \approx N^{-1}$ for each $i$. Due to fluctuations, this can hold only in a somewhat weaker sense.

The first type of results provide an upper bound on the $\ell^\infty$ norm in very high probability:

**Theorem 18.9.** Let $u_{\alpha}$, $\alpha = 1, 2, \ldots, N$, be the $\ell^2$-normalized eigenvectors of a generalized Wigner matrix $H$ satisfying the moment condition (5.6). Then for any (small) $\varepsilon > 0$ and (large) $D > 0$ we have

$$
P\left( \exists \alpha : ||u_{\alpha}||_{\infty} \geq \frac{N^{\varepsilon}}{N} \right) \leq N^{-D}. \tag{18.15}$$

Under stronger decay conditions on $H$, the $N^{\varepsilon}$ tolerance factor may be changed to a log-power and the probability estimate improved to sub exponential. This theorem directly follows from the local semicircle law, Theorem 6.7. Indeed, by spectral decomposition

$$\Im G_{ii}(z) = \eta \sum_{\alpha} \frac{| u_{\alpha}(i) |^2}{(E - \lambda_{\alpha})^2 + \eta^2} \geq \eta^{-1} | u_{\alpha}(i) |^2, \quad z = E + i \eta,$$

where we chose $E$ in the $\eta$-vicinity of $\lambda_{\alpha}$. Since from (6.31) the left hand side is bounded by $N^{\varepsilon}(N\eta)^{-1}$ with very high probability, we obtain (18.15). Under stronger decay conditions on $H$, the local semicircle law can be slightly improved and thus the $N^{\varepsilon}$ tolerance factor may be changed to a log-power and the probability estimate improved to sub exponential.

Although this bound prevents concentration of eigenvectors onto a set of size less than $N^{1-\varepsilon}$, it does not imply the “complete flatness” of eigenvectors in the sense that $|u_{\alpha}(i)| \approx N^{-1/2}$. Note that by a complete $U(N)$ or $O(N)$ symmetry, the eigenfunctions of a GUE or GOE matrix are distributed according to the Haar measure on the $N$-dimensional unit sphere, moreover the distribution of a particular coordinate $u_{\alpha}(i)$ of a fixed eigenvector $u_{\alpha}$ is asymptotically Gaussian. The same result holds for generalized Wigner matrices by Theorem 1.2 of \cite{B89} (a similar result was also obtained in \cite{S78,IT85} under the condition that the first four moments match with those of the Gaussian):

**Theorem 18.10** (Asymptotic normality of eigenvectors). Let $u_{\alpha}$ be the eigenvectors of a generalized Wigner matrix with moment condition (5.6). Then there is a $\delta > 0$ such that for any $m \in \mathbb{N}$ and $I \subset T_N :=$
\[ [1, N^{1/4}] \cup [N^{1-\delta}, N - N^{1-\delta}] \cup [N - N^{1/4}, N] \] with \(|I| = m\) and for any unit vector \(q \in \mathbb{R}^N\), the vector 
\[
(\sqrt{N}((q, \alpha_\alpha)))_{\alpha \in I} \in \mathbb{R}^m
\]
are asymptotically normal in the sense of finite moments.

Moreover, the study of eigenvectors leads to another fundamental problem of mathematical physics: quantum ergodicity. Recall the quantum ergodicity theorem (Shnirel’man [122], Colin de Verdière [35] and Zelditch [140]) asserts that “most” eigenfunctions for the Laplacian on a compact Riemannian manifold with ergodic geodesic flow are completely flat. For \(d\)-regular graphs under certain assumptions on the injectivity radius and spectral gap of the adjacency matrices, similar results were proven for eigenvectors of the adjacency matrices [7]. A stronger notion of quantum ergodicity, the quantum unique ergodicity (QUE) proposed by Rudnick-Sarnak [114] demands that all high energy eigenfunctions become completely flat, and it supposedly holds for negatively curved compact Riemannian manifolds. One case for which QUE was rigorously proved concerns arithmetic surfaces, thanks to tools from number theory and ergodic theory on homogeneous spaces [81, 82, 100]. For Wigner matrices, a probabilistic version of QUE was settled in Corollary 1.4 of [28]:

**Theorem 18.11** (Quantum unique ergodicity for Wigner matrices). Let \(u_\alpha\) be the eigenvectors of a generalized Wigner matrix with moment condition (5.6). Then there exists \(\varepsilon > 0\) such that for any deterministic \(1 \leq \alpha \leq N\) and \(I \subset [1, N]\), for any \(\delta > 0\) we have

\[
P\left( \left| \sum_{i \in I} |u_\alpha(i)|^2 - \frac{|I|}{N} \right| \geq \delta \right) \leq \frac{N^{-\varepsilon}}{\delta^2}. \tag{18.16}
\]

### 18.6 General mean field models

Wigner matrices and their generalizations in Definition 2.1 are the most prominent mean field random matrix ensembles, but there are several other models of interest. Here we give a partial list. More detailed account of the results and citations can be found later on in the few representative references.

The three step strategy opens up a path to investigate local spectral statistics of a large class of matrices. In many examples, the limiting density of the eigenvalues is not the semicircle any more, so the Dyson Brownian motion is not in global equilibrium and Step 2 is more complicated. In the spirit of Dyson’s conjecture, however, gap universality already follows from local equilibration of DBM. Indeed, a local version of the DBM was used as a basic tool in [25]. A recent development to compare local and global dynamics of DBM was implemented by a coupling and homogenization argument in [26]. It turns out that as long as the initial condition of the DBM is regular on a certain scale, the local gap statistics reaches its equilibrium, i.e., the Wigner-Dyson-Mehta distribution, on the corresponding time scale [92] (see also [65] for a similar result). This concept reduces the proof of universality to a proof of the corresponding local law.

One natural class of mean field ensembles are the deformed Wigner matrices; these are of the form \(H = V + W\), where \(V\) is a diagonal matrix and \(W\) is a standard Wigner matrix. In fact, after a trivial rescaling, these matrices may be viewed as instances of the DBM with initial condition \(H_0 = V\), recalling that the law of the DBM at time \(t\) is given by \(H_t = e^{-t/2}V + (1 - e^{-t})^{1/2}W\). Assuming that the diagonal elements of \(V\) have a limiting density \(\varrho_V\), the limiting density of \(H_t\) is given by the free convolution of \(\varrho_V\) with the semicircle density. The corresponding local law was proved in [93] followed by the proof of bulk universality in [96] and the edge universality [95, 96].

The free convolution of two arbitrary densities, \(\varrho_\alpha \boxplus \varrho_\beta\), naturally appear in random matrix theory as the limiting density of the sum of two asymptotically free matrices \(A\) and \(B\) whose limiting densities are given by \(\varrho_\alpha\) and \(\varrho_\beta\). Moreover, for any deterministic \(A\) and \(B\), the matrices \(A\) and \(UBU^*\) are asymptotically free [136], where \(U\) is a Haar distributed unitary matrix. Thus the eigenvalue density of \(A + UBU^*\) is given by the free convolution. The corresponding local law on the optimal scale in the bulk was given in [15]. Another way to generate a matrix ensemble with a limiting density different from the semicircle law is to consider Wigner-type matrices that are further generalizations of the universal Wigner matrices from Definition 2.1. These are complex Hermitian or real symmetric matrices with centered independent (up to symmetry) matrix elements but without the condition that the sum of the variances in each row is constant.
The limiting density is determined by the matrix of variances $S$ by solving the corresponding Dyson-Schwinger equation. Depending on $S$, the density may exhibit square root singularity similar to the edge of the semicircle law or a cubic root singularity [5]. The corresponding optimal local law and bulk universality were proved in [4].

Different complications arise for adjacency matrices of sparse graphs. Each edge of the Erdős-Rényi graph is chosen independently with probability $p$, so it is a mean field model with a semicircle density, but a typical realization of the matrix contains many zero elements if $p \ll 1$. Optimal local law was proved in [56], bulk universality for $p \gg N^{-1/3}$ in [53] and for $p \gg N^{-4}$ in [83]. Another prominent model of random graphs is the uniform measure on $d$-regular graphs. The elements of the adjacency matrix are weakly dependent since their sum in every row is $d$. For $d \gg 1$ the limiting density is the semicircle law and optimal local law was obtained in [17]. Bulk universality was shown in the regime $1 \ll d \ll N^{2/3}$ in [16].

Sample covariance matrices (2.9) and especially their deformations with finite rank deterministic matrices play an important role in statistics. The first application of the three step strategy to prove bulk universality was presented in [64]. The edge universality was achieved in [89,94]. For applications in principal component analysis, the main focus is on the extreme eigenvalues and the outliers whose detailed analysis was given in [22]. Here we have listed references for papers using methods closely related to this book. There are many references in this subject which we are unable to include here.

### 18.7 Beyond mean field models: band matrices

Wigner predicted that universality should hold for any large quantum system, described by a Hamiltonian $H$, of sufficient complexity. After discretization of the underlying physical state space, the Hamilton operator is given by a matrix $H$ whose matrix elements $H_{xy}$ describe the quantum transition rate from state $x$ to $y$. Generalized Wigner matrices correspond to a mean-field situation where transition from any state $x$ to $y$ is possible. In contrast, random Schrödinger operators of the form $-\Delta + V$ defined on $\mathbb{Z}^d$ allow transition only to neighboring lattice sites with a random on-site interaction, so they are prototypes of disordered quantum systems with a non-trivial spatial structure.

As explained in Section 18.5, from the point of view of Anderson phase transition, generalized Wigner matrices are always in the delocalized regime. Random Schrödinger operators in $d = 1$ dimension, or more general random matrices $H$ representing one dimensional Hamiltonians with short range random hoppings, are in the localized regime. It is therefore natural to vary the range of interaction to detect the phase transition.

One popular model interpolating between the Wigner matrices and random Schrödinger operator is the random band matrix, see Example 6.1. In this model the physical state space, that labels the matrix elements, is equipped with a distance. Band matrices are characterized by the property that $H_{ij}$ becomes negligible if $\text{dist}(i, j)$ exceeds a certain parameter, $W$, called the band width. A fundamental conjecture [75] states that the local spectral statistics of a band matrix $H$ are governed by random matrix statistics for large $W$ and by Poisson statistics for small $W$. The transition is conjectured to be sharp [75,125] for the band matrices in one spatial dimension around the critical value $W = \sqrt{N}$. In other words, if $W \gg \sqrt{N}$, we expect the universality results of [26,57,63,67] to hold. Furthermore, the eigenvectors of $H$ are expected to be completely delocalized in this range. For $W \ll \sqrt{N}$, one expects that the eigenvectors are exponentially localized. This is the analogue of the celebrated Anderson metal-insulator transition for random band matrices. The only rigorous work indicating the $\sqrt{N}$ threshold concern the second mixed moments of the characteristic polynomial for a special class of Gaussian band matrices [119,121].

The localization length for band matrices in one spatial dimension was recently investigated in numerous works. For general distribution of the matrix entries, eigenstates were proved to be localized [115] for $W \ll N^{1/8}$, and delocalization of most eigenvectors in a certain averaged sense holds for $W \gg N^{6/7}$ [50,51], improved to $W \gg N^{4/5}$ [54]. The Green’s function $(H - z)^{-1}$ was controlled down to the scale $\text{Im } z \gg W^{-1}$ in [69], implying a lower bound of order $W$ for the localization length of all eigenvectors. When the entries are Gaussian with some specific covariance profiles, supersymmetry techniques are applicable to obtain stronger results. This approach has first been developed by physicists (see [49] for an overview); the rigorous analysis was initiated by Spencer (see [125] for an overview), with an accurate estimate on the expected density of
states on arbitrarily short scales for a three-dimensional band matrix ensemble in [44]. More recent works include universality for $W \geq cN$ [120], and the control of the Green’s function down to the optimal scale $\text{Im} z \gg N^{-1}$, hence delocalization in a strong sense for all eigenvectors, when $W \gg N^{6/7}$ [14] with first four moments matching the Gaussians ones (both results require a block structure and hold in part of the bulk spectrum).

While delocalization and Wigner-Dyson-Mehta spectral statistics are expected to occur simultaneously, there is no rigorous argument directly linking them. The Dyson Brownian Motion, the cornerstone of the three step strategy, proves universality for matrices where each entry has a non trivial Gaussian component and the comparison ideas require that second moments match exactly. Therefore this approach cannot be applied to matrices with many zero entries. However, a combination of the quantum unique ergodicity with a mean field reduction strategy in [27] yields Wigner-Dyson-Mehta bulk universality for a large class of band matrices with general distribution in the large band width regime $W \geq cN$. In contrast to the bulk, universality at the spectral edge is much better understood: extreme eigenvalues follow the Tracy-Widom law for $W \gg N^{5/6}$, an essentially optimal condition [123].
References


166
Index

β-ensemble, 17

Admissible control parameter, 50
Airy function, 21
Airy kernel, 21
Anderson localization, 154
Anderson transition, 154, 156

Backtracking sequence, 11
Bakry-Émery argument, 95
Band width, 156
Brascamp-Lieb inequality, 102, 107

Catalan numbers, 12
Christoffel-Darboux formula, 19
Classical exponent, 17
Classical location of eigenvalues, 79
Concentration inequality, 97
Correlation function, 18

Deformed Wigner matrix, 155
Delocalization, 24, 154
Determinantal correlation, 19
Dirichlet form, 83, 94, 111
Dirichlet form inequality, 114, 119
Dyson Brownian motion (DBM), 7, 25, 80
Dyson conjecture, 7, 25, 84, 112

Edge universality, 140, 153
Empirical distribution function, 77
Empirical eigenvalue distribution, 77
Empirical eigenvalue measure, 13
Energy parameter, 13
Energy universality; averaged, fixed, 23, 148
Entropy inequality, 91

Fluctuation averaging, 31, 46, 56, 63
Free convolution, 155

Gap universality; averaged, fixed, 23, 148
Gaudin distribution, 5
Gaussian β-ensemble, 17, 25
Gaussian divisible ensemble, 6, 24, 25
Gaussian Orthogonal Ensemble (GOE), 5, 10
Gaussian Unitary Ensemble (GUE), 5, 10
Generalized Wigner matrix, 6, 9
Generator, 84
Gibbs inequality, 90
Gibbs measure, 16
Green function, 27
Green function comparison, 134, 140, 142

Green function continuity, 125
Gross’ hypercontractivity bound, 100
Haar measure, 16, 155
Harish-Chandra-Itzykson-Zuber formula, 6
Helffer-Sjöstrand formula, 75
Herbst bound, 97
Hermite polynomial, 18

Integrated density of states, 77
Interlacing of eigenvalues, 36
Invariant ensemble, 6, 15, 151
Invariant measure, 25
Isotropic local law, 31, 31
Level repulsion, 81
Local equilibrium, 7, 25
Local relaxation flow, 113
Local relaxation measure, 113
Local semicircle law, 6, 24, 30
Log gas, 17, 151
Logarithmic Sobolev inequality (LSI), 94

Marchenko-Pastur law, 11
Marcinkiewicz-Zygmund inequality, 44
Mean field model, 10
Moment matching, 138
Moment method, 11

Ornstein-Uhlenbeck process, 6, 24, 80, 125

Partial expectation, 35, 46
Pinsker inequality, 91
Plancherel-Rotach asymptotics, 20, 21

Quadratic large deviation, 44
Quantum (unique) ergodicity, 155, 157
Quaternion determinant, 21

Random band matrix, 26, 156
Random Schrödinger operator, 26, 154
Regular graph, 156
Regular potential, 151
Relative entropy, 90
Resolvent, 13, 27
Resolvent decoupling identities, 47
Riemann-Hilbert method, 6
Rigidity of eigenvalues, 24, 79, 151

Sample covariance matrix, 10, 156
Schur formula, 34

167
Self-consistent equation, 31, 37, 46
Sine-kernel, 20
Single entry distribution, 10
Sparse graph, 156
Spectral domain, 29, 50
Spectral edge, 26
Spectral gap, 98
Spectral parameter, 13
Stieltjes transform, 13, 26, 74
Stochastic domination, 29

Three-step strategy, 6, 24, 149
Tracy-Widom law, 140

Universal Wigner matrix, 9

Vandermonde determinant, 15

Ward identity, 47
Weak local law, 31, 34
Wigner ensemble, 5, 9
Wigner semicircle law, 10, 26
Wigner surmise, 5
Wigner-Dyson-Mehta (WDM) conjecture, 5, 157
Wigner-type matrix, 155
Wishart matrix, 10