Lecture Notes on the Matrix Dyson Equation and its Applications for Random Matrices

László Erdős*

Institute of Science and Technology, Austria

June 19, 2017

Abstract

These lecture notes are a concise introduction of recent techniques to prove local spectral universality for a large class of random matrices. The general strategy is presented following the recent book with H.T. Yau [39]. We extend the scope of this book by focusing on new techniques developed to deal with generalizations of Wigner matrices that allow for non-identically distributed entries and even for correlated entries. This requires to analyze a system of nonlinear equations, or more generally a nonlinear matrix equation called the *Matrix Dyson Equation (MDE)*. We demonstrate that stability properties of the MDE play a central role in random matrix theory. The analysis of MDE is based upon joint works with J. Alt, O. Ajanki, D. Schröder and T. Krüger that are supported by the ERC Advanced Grant, RANMAT 338804 of the European Research Council.

AMS Subject Classification (2010): 15B52, 82B44

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

*Partially supported by ERC Advanced Grant, RANMAT 338804

Contents

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

The cornerstone of probability theory is the fact that the collective behavior of many independent random variables exhibits universal patterns; the obvious examples are the *law of large numbers (LLN)* and the *central limit theorem (CLT)*. They assert that the normalized sum of N independent, identically distributed (i.i.d.) random variables $X_1, X_2, \dots, X_N \in \mathbb{R}$ converge to their common expectation value:

$$\frac{1}{N}\big(X_1 + X_2 + \ldots + X_N\big) \to \mathbb{E}X \tag{1.1}$$

as $N \to \infty$, and their centered average with a $\sqrt{N}$ normalization converges to the centered Gaussian distribution with variance $\sigma^2 = \mathrm{Var}(X)$:

$$S_N := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big(X_i - \mathbb{E}X\big) \Longrightarrow \mathcal{N}(0, \sigma^2).$$

The convergence in the latter case is understood in distribution, i.e. tested against any bounded continuous function Φ:

$$\mathbb{E}\Phi(S_N) \to \mathbb{E}\Phi(\xi),$$

where ξ is an $\mathcal{N}(0, \sigma^2)$ distributed normal random variable.

These basic results directly extend to random vectors instead of scalar valued random variables. The main question is: what are their analogues in the non-commutative setting, e.g. for matrices? Focusing on their spectrum, what do eigenvalues of typical large random matrices look like? Is there a deterministic limit of some relevant random quantity, like the average in case of the LLN (1.1). Is there some stochastic *universality pattern* arising, similarly to the ubiquity of the Gaussian distribution in Nature owing to the central limit theorem?

These natural questions could have been raised from pure curiosity by mathematicians, but historically random matrices first appeared in statistics (Wishart in 1928 [77]), where empirical covariance matrices of measured data (samples) naturally form a random matrix ensemble and the eigenvalues play a crucial role in principal component analysis. The question regarding the universality of eigenvalue statistics, however, appeared only in the 1950's in the pioneering work [76] of Eugene Wigner. He was motivated by a simple observation looking at data from nuclear physics, but he immediately realized a very general phenomenon in the background. 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. Instead of pursuing a direct solution of this problem, Wigner appealed to a phenomenological model to explain his observation. His 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. Eigenvalue gaps carry basic information about possible excitations of the quantum systems. In fact, beyond nuclear physics, random matrices enjoyed a renaissance in the theory of disordered quantum systems, where the spectrum of a non-interacting electron in a random impure environment was studied. It turned out that eigenvalue statistics is one of the basic signatures of the celebrated *metal-insulator*, or *Anderson* transition in condensed matter physics [11].

1.1 Random matrix ensembles

Throughout these notes we will consider $N \times N$ square matrices of the form

$$H = H^{(N)} = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1N} \\ h_{21} & h_{22} & \dots & h_{2N} \\ \vdots & \vdots & & \vdots \\ h_{N1} & h_{N2} & \dots & h_{NN} \end{pmatrix}. \tag{1.2}$$

The entries are real or complex random variables, subject to the symmetry constraint

$$h_{ij} = \bar{h}_{ji}, \qquad i,j = 1,\dots,N,$$

so that $H = H^*$ is either Hermitian (complex) or symmetric (real). In particular, the eigenvalues of H, $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N$ are real and we will be interested in their statistical behavior induced by the randomness of H as the size of the matrix N goes to infinity. Hermitian symmetry is very natural from the point of view of physics applications and it makes the problem much more tractable mathematically. Nevertheless, there has recently been an increasing interest in non-hermitian random matrices as well motivated by systems of ordinary differential equations with random coefficients arising in biological networks (see, e.g. [9] and references therein).

There are essentially two customary ways to define a probability measure on the space of $N \times N$ random matrices that we now briefly introduce. The main point is that either one specifies the distribution of the matrix elements directly or one aims at a basis-independent measure. The prototype of the first case is the Wigner ensembles and we will be focusing on its natural generalizations in these notes. The typical example of the second case are the invariant ensembles. We will briefly introduce them now.

1.1.1 Wigner ensemble

The most prominent example of the first class is the traditional **Wigner matrix**, where the matrix elements h_{ij} are i.i.d. random variables subject to the symmetry constraint $h_{ij} = \overline{h_{ji}}$. More precisely, Wigner matrices are defined by assuming that

$$\mathbb{E} h_{ij} = 0, \qquad \mathbb{E}|h_{ij}|^2 = \frac{1}{N}, \tag{1.3}$$

and in the *real symmetric case*, the collection of random variables $\{h_{ij} \ : \ i \leqslant j\}$ are independent, identically distributed, while in the *complex hermitian case* the distributions of $\{\operatorname{Re} h_{ij}, \operatorname{Im} h_{ij} \ : \ 1 \leqslant i < j \leqslant N\}$ and $\{\sqrt{2} h_{ii} \ : \ i = 1, 2, \dots, N\}$ are independent and identical.

The common variance of the matrix elements is the single parameter of the model; by a trivial rescaling we may fix it conveniently. The normalization $1/N$ chosen in (1.3) guarantees that the typical size of the eigenvalues remain of order 1 even as N tends to infinity. To see this, we may compute the expectation of the trace of H^2 in two different ways:

$$\mathbb{E}\sum_i \lambda_i^2 = \mathbb{E}\,\mathrm{Tr}\,H^2 = \mathbb{E}\sum_{ij}|h_{ij}|^2 = N \tag{1.4}$$

indicating that $\lambda_i^2 \sim 1$ on average. In fact, much stronger bounds hold and one can prove that

$$\|H\| = \max_i |\lambda_i| \to 2, \qquad N \to \infty,$$

in probability.

In these notes we will focus on Wigner ensembles and their extensions, where we will drop the condition of identical distribution and we will weaken the independence condition. We will call them *Wigner type* and *correlated ensembles*. Nevertheless, for completeness we also present the other class of random matrices.

1.1.2 Invariant ensembles

The ensembles in the second class are defined by the measure

$$\mathbb{P}(H)\mathrm{d}H := Z^{-1}\exp\big(-\frac{\beta}{2}N\,\mathrm{Tr}\,V(H)\big)\mathrm{d}H. \tag{1.5}$$

Here $\mathrm{d}H = \prod_{i\leqslant j}\mathrm{d}h_{ij}$ is the flat Lebesgue measure on $\mathbb{R}^{N(N+1)/2}$ (in case of complex Hermitian matrices and $i<j$, $\mathrm{d}H_{ij}$ is the Lebesgue measure on the complex plane $\mathbb{C}$ instead of $\mathbb{R}$). The (potential) 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 (1.5) is finite. The parameter β distinguishes between the two symmetry classes: $\beta=1$ for the real symmetric case, while $\beta=2$ for the complex hermitian case – for traditional reason we factor this parameter out of the potential. Finally, Z is the normalization factor to make $P(H)\mathrm{d}H$ a probability measure. Similarly to the normalization of the variance in (1.3), the factor N in the exponent in (1.5) guarantees that the eigenvalues remain order one even as $N\to\infty$. This scaling also guarantees that empirical density of the eigenvalues will have a deterministic limit without further rescaling.

Probability distributions of the form (1.5) 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 (1.5) invariant thanks to $\mathrm{Tr}\,V(U^*HU) = \mathrm{Tr}\,V(H)$.

An important special case is when V is a quadratic polynomial, after shift and rescaling we may assume that $V(x)=\frac{1}{2}x^2$. In this case

$$\mathbb{P}(H)\mathrm{d}H = Z^{-1}\exp\big(-\frac{\beta}{4}N\sum_{ij}|h_{ij}|^2\big)\mathrm{d}H = Z^{-1}\prod_{i<j}\exp\big(-\frac{\beta}{2}N|h_{ij}|^2\big)\mathrm{d}h_{ij}\prod_i\exp\big(-\frac{\beta}{4}Nh_{ii}^2\big)\mathrm{d}h_{ii},$$

i.e. the measure factorizes and it is equivalent to independent Gaussians for the matrix elements. The factor N in the definition (1.5) and the choice of β ensure that we recover the normalization (1.3). (A pedantic reader may notice that the normalization of the diagonal element for the real symmetric case is off by a factor of 2, but this small discrepancy plays no role.) The invariant Gaussian ensembles, i.e. (1.5) with $V(x)=\frac{1}{2}x^2$, are called **Gaussian orthogonal ensemble (GOE)** for the real symmetric case ($\beta=1$) and **Gaussian unitary ensemble (GUE)** for the complex hermitian case ($\beta=2$).

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 1.1 ([26] or Theorem 2.6.3 [63]). *Suppose that the real symmetric or complex Hermitian matrix ensembles given in* (1.5) *have independent entries h_{ij}, $i \leqslant j$. 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.*

The significance of the Gaussian ensembles is that they allow for explicit calculations that are not available for Wigner matrices with general non-Gaussian single entry distribution. In particular the celebrated **Wigner-Dyson-Mehta correlation functions** can be explicitly obtained for the GOE and GUE ensembles. Thus the typical proof of identifying the eigenvalue correlation function for a general matrix ensemble goes through universality: one first proves that the correlation function is independent of the distribution, hence it is the same as GUE/GOE, and then, in the second step, one computes the GUE/GOE correlation functions. This second step has been completed by Gaudin, Mehta and Dyson in the 60's by an ingenious calculation, see e.g. the classical treatise by Mehta [63].

One of the key ingredients of the explicit calculations is the surprising fact that the joint (symmetrized) density function of the eigenvalues, $p(\lambda_1, \lambda_2, \ldots, \lambda_N)$ can be computed explicitly for any invariant ensemble. It is given by

$$p_N(\lambda_1, \lambda_2, \ldots, \lambda_N) = \text{const.} \prod_{i<j} (\lambda_i - \lambda_j)^\beta e^{-\frac{\beta}{2} N \sum_{j=1}^N V(\lambda_j)}. \tag{1.6}$$

where the constant ensures the normalization, but its exact value is typically unimportant.

Remark 1.2. *In other sections of these notes we usually label the eigenvalues in increasing order so that their probability density, denoted by $\widetilde{p}_N(\boldsymbol{\lambda})$, is defined on the set*

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

For the purpose of (1.6)*, 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, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ on $\mathbb{R}^N$. The relation between the ordered and unordered densities is clearly $\widetilde{p}_N(\boldsymbol{\lambda}) = N!\, p_N(\boldsymbol{\lambda}) \cdot \mathbf{1}(\boldsymbol{\lambda} \in \Xi^{(N)})$.*

The emergence of the *Vandermonde determinant* in (1.6) is a result of integrating out the "angle" variables in (1.5), i.e., the unitary matrix in the diagonalization of $H = U \Lambda U^*$. This is a remarkable formula since it gives a direct access to the eigenvalue distribution. In particular, it shows that the eigenvalues are strongly correlated. For example, no two eigenvalues can be too close to each other since the corresponding probability is suppressed by the factor $\lambda_j - \lambda_i$ for any $i \neq j$; this phenomenon is called the **level repulsion**. We remark that level repulsion also holds for Wigner matrices with smooth distribution [34] but its proof is much more involved.

In fact, one may view the ensemble (1.6) as a statistical physics question by rewriting p_N as a classical Gibbs measure of a N point particles on the line with a logarithmic mean field interaction:

$$p_N(\boldsymbol{\lambda}) = (\text{const.}) e^{-\beta N \mathcal{H}(\boldsymbol{\lambda})} \tag{1.7}$$

with a Hamiltonian

$$\mathcal{H}(\boldsymbol{\lambda}) = \frac{1}{2} \sum_i V(\lambda_i) - \frac{1}{N} \sum_{i<j} \log |\lambda_j - \lambda_i|.$$

This ensemble of point particles with logarithmic interactions is also called **log-gas**. We remark that viewing the Gibbs measure (1.7) as the starting point and forgetting about the matrix ensemble behind, the parameter β does not have to be 1 or 2; it can be any positive number, $\beta > 0$, and it has the interpretation of the inverse temperature. We will not pursue general invariant ensembles in these notes.

1.2 Eigenvalue statistics on different scales

The normalization both in (1.3) and (1.5) is chosen in such a way that the typical eigenvalues remain of order 1 even in the large N limit. In particular, the typical distance between neighboring eigenvalues is of order $1/N$. We distinguish two different scales for studying eigenvalues: *macroscopic* and *microscopic* scales. With our scaling, the macroscopic scale is order one and on this scale we detect the cumulative effect of cN eigenvalues with some positive constant c. In contrast, on the microscopic scales individual eigenvalues are detected; this scale is typically of order $1/N$. However, near the spectral edges, where the density of eigenvalues goes to zero, the typical eigenvalue spacing hence the microscopic scale may be larger. Some phenomena (e.g. fluctuations of linear statistics of eigenvalues) occur on various *mesoscopic* scales that lie between the macroscopic and the microscopic scales.

1.2.1 Eigenvalue density on macroscopic scales: global laws

The first and simplest question is to determine the eigenvalue density, i.e. the behavior of the *empirical eigenvalue density* or **empirical density of states**

$$\mu_N(\mathrm{d}x) := \frac{1}{N}\sum_i \delta(x-\lambda_i) \tag{1.8}$$

in the large N limit. This is a random measure, but under very general conditions it converges to a deterministic measure, similarly to self-averaging property encoded in the law of large numbers (1.1).

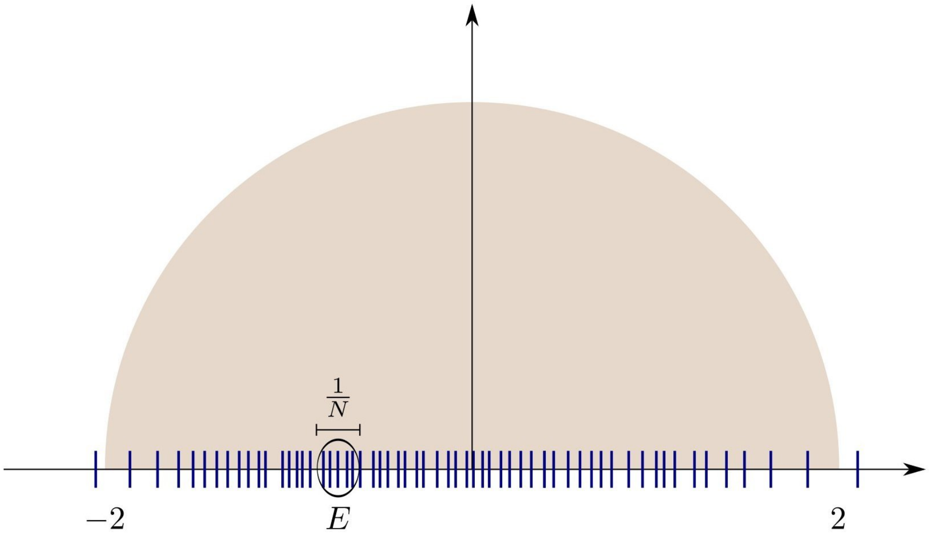


Figure 1: Semicircle law and eigenvalues of a GUE random matrix of size $N = 60$.

For Wigner ensemble, the empirical distribution of eigenvalues converges to 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 [76]) that for any fixed $a \leqslant b$ real numbers,

$$\lim_{N\to\infty} \frac{1}{N}\#\big\{i \,:\, \lambda_i \in [a,b]\big\} = \int_a^b \varrho_{sc}(x)\mathrm{d}x, \qquad \varrho_{sc}(x) := \frac{1}{2\pi}\sqrt{(4-x^2)_+}, \tag{1.9}$$

where $(a)_+ := \max\{a,0\}$ denotes the positive part of the number a. Alternatively, one may formulate the Wigner semicircle law as the weak convergence of the empirical distribution μ_N to the semicircle distribution, $\varrho_{sc}(x)\mathrm{d}x$, i.e.

$$\int_{\mathbb{R}} f(x)\mu_N(\mathrm{d}x) = \frac{1}{N}\sum_i f(\lambda_i) \to \int f(x)\varrho_{sc}(x)\mathrm{d}x, \qquad N\to\infty$$

for any continuous function f.

Note that the emergence of the semicircle density is already a certain form of universality: the common distribution of the individual matrix elements is "forgotten"; the density of eigenvalues is asymptotically always the same, independently of the details of the distribution of the matrix elements.

We will see that for a more general class of Wigner type matrices with zero expectation but not identical distribution a similar limit statement holds for the empirical density of eigenvalues, i.e. there is a deterministic density function $\varrho(x)$ such that

$$\int_{\mathbb{R}} f(x)\mu_N(\mathrm{d}x) = \frac{1}{N}\sum_i f(\lambda_i) \to \int f(x)\varrho(x)\mathrm{d}x, \qquad N \to \infty \tag{1.10}$$

holds. The density function ϱ thus approximates the empirical density, so we will call it **asymptotic density of states.** In general it is not the semicircle density, but is determined by the second moments of the matrix elements and it is independent of other details of the distribution. For independent entries, the variance matrix

$$S = (s_{ij})_{i,j=1}^N, \qquad s_{ij} := \mathbb{E}|h_{ij}|^2 \tag{1.11}$$

contains all necessary information. For matrices with correlated entries, all relevant second moments are encoded in the linear operator

$$\mathcal{S}[R] := \mathbb{E}HRH, \qquad R \in \mathbb{C}^{N\times N}$$

acting on $N \times N$ matrices. It is one of the key question in random matrix theory to compute the asymptotic density ϱ from the second moments; we will see that the answer requires solving a system of nonlinear equations, that will be commonly called the **Dyson equation**. The explicit solution leading to the semicircle law is available only for Wigner matrices, or a little bit more generally, for ensembles with the property

$$\sum_j s_{ij} = 1 \qquad \text{for any } i. \tag{1.12}$$

These are called **generalized Wigner ensembles** and have been introduced in [44].

For invariant ensembles, the asymptotic density $\varrho = \varrho_V$ depends on the potential function V. It can be computed by solving a convex minimization problem, namely it is the the unique minimizer of the functional

$$I(\nu) = \int_{\mathbb{R}} V(t)\nu(t)\mathrm{d}t - \int_{\mathbb{R}}\int_{\mathbb{R}} \log|t-s|\nu(s)\nu(t)\mathrm{d}t\mathrm{d}s.$$

In both cases, under some mild conditions on the variances S or on the potential V, respectively, the asymptotic density ϱ is compactly supported.

1.2.2 Eigenvalues on mesoscopic scales: local laws

The Wigner semicircle law in the form (1.9) asymptotically determines the number of eigenvalues in a fixed interval $[a,b]$. The number of eigenvalues in such intervals is comparable with N. However, keeping in mind the analogy with the law of large numbers, it is natural to raise the question whether the same asymptotic relation holds if the length of the interval $[a,b]$ shrinks to zero as $N \to \infty$. To expect a deterministic answer, the interval should still contain many eigenvalues, but this would be guaranteed by $|b-a| \gg 1/N$. This turns out to be correct and the local semicircle law asserts that

$$\lim_{N\to\infty} \frac{1}{2N\eta}\#\big\{i \,:\, \lambda_i \in [E-\eta, E+\eta]\big\} = \varrho_{sc}(E) \tag{1.13}$$

uniformly in $\eta = \eta_N$ as long as $N^{-1+\varepsilon} \leqslant \eta_N \leqslant N^{-\varepsilon}$ for any $\varepsilon > 0$ and E is not at the edge, $|E| \neq 2$. Here we considered the interval $[a,b] = [E-\eta, E+\eta]$, i.e. we fixed its center and viewed its length as an N-dependent parameter. (The N^ε factors can be improved to some $(\log N)$-power.)

1.2.3 Eigenvalues on microscopic scales: universality of local eigenvalue statistics

Wigner's original observation concerned the distribution of the distances between consecutive (ordered) eigenvalues, or **gaps**. In the bulk of the spectrum, i.e. in the vicinity of a fixed energy level E with $|E| < 2$ in case of the semicircle law, the gaps have a typical size of order $1/N$ (at the spectral edge, $|E| = 2$, the relevant microscopic scale is of order $N^{-2/3}$, but we will not pursue edge behavior in these notes). Thus the corresponding rescaled gaps have the form

$$g_i := N\varrho(\lambda_i)\big(\lambda_{i+1} - \lambda_i\big), \tag{1.14}$$

where ϱ is the asymptotic density, e.g. $\varrho = \varrho_{sc}$ for Wigner matrices. Wigner predicted that the fluctuations of the gaps are universal and their distribution is given by a new law, the *Wigner surmise.* Thus there exists a random variable ξ, depending only on the symmetry class $\beta = 1, 2$, such that

$$g_i \Longrightarrow \xi$$

in distribution, for any gap away from the edges, i.e., if $\varepsilon N \leqslant i \leqslant (1-\varepsilon)N$ with some fixed $\varepsilon > 0$. This might be viewed as the random matrix analogue of the central limit theorem. Note that universality is twofold. First, the distribution of g_i is independent of the index i (as long as λ_i is away from the edges). Second, more importantly, the limiting gap distribution is independent of the distribution of the matrix elements, similarly to the universal character of the central limit theorem.

However, the gap universality holds much more generally than the semicircle law: the rescaled gaps (1.14) follow the same distribution as the gaps of the GUE or GOE (depending on the symmetry class) essentially for any random matrix ensemble with "sufficient" amount of randomness. In particular, it holds for invariant ensembles, as well as for Wigner type and correlated random matrices, i.e. for very broad extensions of the original Wigner ensemble. In fact, it holds much beyond the traditional realm of random matrices; it is conjectured to hold for any random matrix describing a disordered quantum system in the *delocalized regime*, see Section 5.2 later.

The universality on microscopic scales can also be expressed in terms of the appropriately rescaled **correlation functions**, in fact, in this way the formulas are more explicit. First we define the correlation functions.

Definition 1.3. *Let $p_N(\lambda_1, \lambda_2, \dots, \lambda_N)$ be the joint symmetrized probability distribution of the eigenvalues. For any $n \geqslant 1$, the n-point correlation function is defined by*

$$p_N^{(n)}(\lambda_1, \lambda_2, \dots, \lambda_n) := \int_{\mathbb{R}^{N-n}} p_N(\lambda_1, \dots, \lambda_n, \lambda_{n+1}, \dots \lambda_N)\mathrm{d}\lambda_{n+1} \dots \mathrm{d}\lambda_N. \tag{1.15}$$

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

$$\frac{1}{N(N-1)}\mathbb{E}\sum_{i\neq j} O(\lambda_i, \lambda_j) = \int_{\mathbb{R}\times\mathbb{R}} O(\lambda_1, \lambda_2) p_N^{(2)}(\lambda_1, \lambda_2)\mathrm{d}\lambda_1\mathrm{d}\lambda_2, \tag{1.16}$$

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. In particular, the global law (1.10) is equivalent to the statement that the one point correlation function converges to the asymptotic density

$$p_N^{(1)}(x)\mathrm{d}x \to \varrho(x)\mathrm{d}x$$

weakly, since

$$\int_{\mathbb{R}} O(x)p_N^{(1)}(x)\mathrm{d}x = \frac{1}{N}\sum_i O(\lambda_i) \to \int O(x)\varrho(x)\mathrm{d}x.$$

Correlation functions are difficult to compute in general, even if the joint density function p_N is explicitly given as in the case of the invariant ensembles (1.6). Naively one may think that computing the correlation functions in this latter case boils down to an elementary calculus exercise by integrating out all but a few variables. However, that task is complicated.

As mentioned, one may view the joint density of eigenvalues of invariant ensembles (1.6) as a Gibbs measure of a log-gas and here β can be any positive number (inverse temperature). The universality of correlation functions is a valid question for all β-log-gases that has been positively answered in [15,19–21,67] by showing that for a sufficiently smooth potential V (in fact $V \in C^4$ suffices) the correlation functions depend only on β and are independent of V. We will not pursue general invariant ensembles in these notes.

The logarithmic interaction is of long range, hence the system (1.7) is strongly correlated and standard methods of statistical mechanics to compute correlation functions cannot be applied. The computation is quite involved even for the simplest Gaussian case, and it relies on sophisticated identities involving Hermite orthogonal polynomials. These calculations have been developed by Gaudin, Mehta and Dyson in the 60's and can be found, e.g. in Mehta's book [63]. Here we just present the result for the most relevant $\beta = 1, 2$ cases.

We fix an energy E in the bulk, i.e., $|E| < 2$, and we rescale the correlation functions by a factor $N\varrho$ around E to make the typical distance between neighboring eigenvalues 1. These rescaled correlation functions then have a universal limit:

Theorem 1.4. *For GUE ensembles, the rescaled correlation functions converge to the determinantal formula with the sine kernel, $S(x) := \frac{\sin \pi x}{\pi x}$, i.e.*

$$\frac{1}{[\varrho_{sc}(E)]^n} p_N^{(n)}\Big(E + \frac{\alpha_1}{N\varrho_{sc}(E)}, E + \frac{\alpha_2}{N\varrho_{sc}(E)}, \ldots, E + \frac{\alpha_n}{N\varrho_{sc}(E)}\Big) \rightharpoonup q_{GUE}^{(n)}(\boldsymbol{\alpha}) := \det\big(S(\alpha_i - \alpha_j)\big)_{i,j=1}^n \quad (1.17)$$

as weak convergence of functions in the variables $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$.

Formula (1.17) holds for the GUE case. The corresponding expression for GOE is more involved [10,63]

$$q_{\rm GOE}^{(n)}(\boldsymbol{\alpha}) := \det\big(K(\alpha_i - \alpha_j)\big)_{i,j=1}^n, \qquad K(x) := \begin{pmatrix} S(x) & S'(x) \\ -\frac{1}{2}\,\mathrm{sgn}(x) + \int_0^x S(t)\mathrm{d}t & S(x) \end{pmatrix}. \quad (1.18)$$

Here the determinant is understood as the trace of the *quaternion determinant* after the canonical correspondance between quaternions $a \cdot \mathbf{1} + b \cdot \mathbf{i} + c \cdot \mathbf{j} + d \cdot \mathbf{k}$, $a, b, c, d \in \mathbb{C}$, and 2×2 complex matrices given by

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

Note that the limit in (1.17) is universal in the sense that it is independent of the energy E. However, universality also holds in a much stronger sense, namely that the local statistics (limits of rescaled correlation functions) depend only on the symmetry class, i.e. on β, and are independent of any other details. In particular, they are always given by the sine kernel (1.17) or (1.18) not only for the Gaussian case but for any Wigner matrices with arbitrary distribution of the matrix elements, as well as for any invariant ensembles with arbitrary potential V. This is the **Wigner-Dyson-Mehta (WDM) universality conjecture**, formulated precisely in Mehta's book [63] in the late 60's.

The WDM conjecture for invariant ensembles has been in the focus of very intensive research on orthogonal polynomials with general weight function (the Hermite polynomials arising in the Gaussian setup have Gaussian weight function). It motivated the development of the Riemann-Hilbert method [45], that was originally brought into this subject by Fokas, Its and Kitaev [45], and the universality of eigenvalue statistics was established for large classes of invariant ensembles by Bleher-Its [17] and by Deift and collaborators [26–28]. The key element of this success was that invariant ensembles, unlike Wigner matrices, have explicit formulas (1.6) for the joint densities of the eigenvalues. With the help of the Vandermonde structure of these formulas, one may express the eigenvalue correlation functions as determinants whose entries are given by functions of orthogonal polynomials.

For Wigner ensembles, there are no explicit formulas for the joint density of eigenvalues or for the correlation functions statistics and the WDM conjecture was open for almost fifty years with virtually no progress. The first significant advance in this direction was made by Johansson [56], who proved the universality for *complex* Hermitian matrices under the assumption that the common distribution of the matrix entries has a substantial Gaussian component, i.e., the random matrix H is of the form $H = H_0 + aH^G$ where H_0 is a general Wigner matrix, H^G is the GUE matrix, and a is a certain, not too small, positive constant independent of N. His proof relied on an explicit formula by Brézin and Hikami [23, 24] that uses a certain version of the Harish-Chandra-Itzykson-Zuber formula [55]. These formulas are available for the complex Hermitian case only, which restricted the method to this symmetry class.

Exercise 1.5. *Verify formula* (1.16)*.*

1.2.4 The three step strategy

The WDM conjecture in full generality has recently been resolved by a new approach called the **three step strategy** that has been developed in a series of papers by Erdős, Schlein, Yau and Yin between 2008 and 2013 with a parallel development by Tao and Vu. A detailed presentation of this method can be found in [39], while a shorter summary was presented in [43].

This approach consists of the following three steps:

Step 1. Local semicircle law: It provides an a priori estimate showing that the density of eigenvalues of generalized Wigner matrices is given by the semicircle law at very small microscopic scales, i.e., down to spectral intervals that contain N^ε eigenvalues.

Step 2. Universality for Gaussian divisible ensembles: It proves that the local statistics of *Gaussian divisible ensembles* $H_0 + aH^G$ are the same as those of the Gaussian ensembles H^G as long as $a \geqslant N^{-1/2+\varepsilon}$, i.e., already for very small a.

Step 3. Approximation by a Gaussian divisible ensemble: It is a type of "density argument" that extends the local spectral universality from Gaussian divisible ensembles to all Wigner ensembles.

The conceptually novel point is Step 2. The eigenvalue distributions of the Gaussian divisible ensembles, written in the form $e^{-t/2}H_0 + \sqrt{1-e^{-t}}H^{\rm G}$, are the same as that of the solution of a *matrix valued Ornstein-Uhlenbeck (OU) process* H_t

$$\mathrm{d}H_t = \frac{\mathrm{d}\mathbf{B}_t}{\sqrt{N}} - \frac{1}{2}H_t\mathrm{d}t, \qquad H_{t=0} = H_0, \tag{1.19}$$

for any time $t \geqslant 0$, where $\mathbf{B}_t$ is a matrix valued standard Brownian motion of the corresponding symmetry class (The OU process is preferable over its rescaled version $H_0 + aH^G$ since it keeps the variance constant). Dyson [29] observed half a century ago that the dynamics of the eigenvalues $\lambda_i = \lambda_i(t)$ of H_t is given by an interacting stochastic particle system, called the *Dyson Brownian motion (DBM)*, where the eigenvalues are the particles:

$$\mathrm{d}\lambda_i = \sqrt{\frac{\beta}{2}}\frac{1}{\sqrt{N}}\mathrm{d}B_i + \Big(-\frac{\lambda_i}{2} + \frac{1}{N}\sum_{j\neq i}\frac{1}{\lambda_i - \lambda_j}\Big)\mathrm{d}t, \qquad i = 1, 2, \ldots, N. \tag{1.20}$$

Here $\mathrm{d}B_i$ are independent white noises.

In addition, the invariant measure of this dynamics is exactly the eigenvalue distribution of GOE or GUE, i.e. (1.6) with $V(x) = \frac{1}{2}x^2$. This invariant measure is thus a Gibbs measure of point particles in one dimension interacting via a long range logarithmic potential. In fact, β can be any positive parameter, the corresponding DBM (1.20) may be studied even if there is no invariant matrix ensemble behind. Using a heuristic physical argument, Dyson remarked [29] 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. Step 2 gives a precise mathematical meaning of this vague idea. 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. 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 $e^{-t/2}H_0 + \sqrt{1-e^{-t}}H^{\rm G}$ with a suitably chosen H_0.

We note that in the Hermitian case, Step 2 can be circumvented by using the Harish-Chandra-Itzykson-Zuber formula. This approach was followed by Tao and Vu [73] who gave an alternative proof of universality for Wigner matrices in the Hermitian symmetry class as well as for the real symmetric class but only under a certain moment matching condition.

The three step strategy has been refined and streamlined in the last years. By now it has reached a stage when the content of Step 2 and Step 3 can be presented as a very general "black-box" result that is model independent assuming that Step 1, the local law, holds. The only model dependent ingredient is the local law. Hence to prove local spectral universality for a new ensemble, one needs to verify the local law. Thus in these lecture notes we will focus on the recent developments in the direction of the local laws.

We will discuss generalizations of the original Wigner ensemble to relax the basic conditions *"independent, identically distributed"*. First we drop the identical distribution and allow the variances $s_{ij} = \mathbb{E}|h_{ij}|^2$ to vary. The simplest class is the **generalized Wigner matrices**, defined in (1.12), which still leads to the Wigner semicircle law. The next level of generality is to allow arbitrary matrix of variances S. The density of states is not the semicircle any more and we need to solve a genuine **vector Dyson equation** to find the answer. The most general case discussed in these notes are correlated matrices, where different matrix elements have nontrivial correlation that leads to a **matrix Dyson equation**. In all cases we still keep the mean field assumption, i.e. the typical size of the matrix elements is $|h_{ij}| \sim N^{-1/2}$. Since Wigner's vision on the universality of local eigenvalue statistics predicts the same universal behavior for an even much larger class of hermitian random matrices (or operators), it is fundamentally important to extend the validity of the mathematical proofs as much as possible beyond the Wigner case.

We remark that there are several other directions to extend the Wigner ensemble that we will not discuss here in details, we just mention some of them with a few references, but we do not aim at completeness; apologies for any omissions. First, in these notes we will assume very high moment conditions on the matrix elements. These make the proofs easier and the tail probabilities of the estimates stronger. Several works have focused on *lowering the moment assumption* [2,51,57] and even considering *heavy tailed distributions* [16,18]. An important special case is the class of *sparse matrices* such as adjacency matrix of Erdős-Rényi random graphs and d-regular graphs [13,14,33,40,54]. Another direction is to remove the condition that the matrix elements are centered; this ensemble often goes under the name of *deformed Wigner matrices.* One typically separates the expectation and writes $H = A + W$, where A is a deterministic matrix and W is a Wigner matrix with centered entries. Diagonal deformations (A is diagonal) are easier to handle, this class was considered even for a large diagonal in [58,59,62,64]. The general A was considered in [52]. Finally, a very challenging direction is to depart from the mean field condition, i.e. allow some matrix elements to be much bigger than $N^{-1/2}$. The ultimate example is the *random band matrices* that goes towards the random Schrödinger operators [12,30–32,66,69–71].

1.2.5 User's guide

These lecture notes were intended to Ph.D students and postdocs with general interest in analysis and probability; we assume knowledge of these areas on a beginning Ph.D. level. The overall style is informal, the proof of many statements are only sketched or indicated. Several technicalities are swept under the rug – for the precise theorems the reader should consult with the original papers. We put emphasize on conveying the main ideas in a colloquial way.

In Section 2 we collected basic tools from analysis such as Stieltjes transform and resolvent. We also introduce the semicircle law. We outline the power method that was traditionally important in random matrices, but we will not rely on it in these notes, so this part can be skipped. In Section 3 we outline the main method to obtain local laws, the resolvent approach and we explain in an informal way its two constituents; the probabilistic and deterministic parts. In Section 4 we introduce four models of Wigner-like ensembles with increasing complexity and we informally explain the novelty and the additional complications for each model. Section 5 on the physical motivations to study these models is a detour. Readers interested only in the mathematical aspects may skip this section. Section 6 contains our main results on the local law formulated in a mathematically precise form. We did not aim at presenting the strongest results and the weakest possible conditions; the selection was guided to highlight some key phenomena. Some consequences of these local laws are also presented with sketchy proofs. Section 7 and 8 contain the main mathematical part of these notes, here we give a more detailed analysis of the vector and the matrix Dyson equation and their stability properties. In these sections we aim at rigorous presentation although not every proof contains all details. Finally, in Section 9 we present the main ideas of the proof of the local laws based upon the stability results on the Dyson equation.

These lecture notes are far from being a comprehensive text on random matrices. Many key issues are left out and even those we discuss will be presented in their simplest form. For more interested readers, we refer to the recent book [39] that focuses on the three step strategy and discusses all steps in details. For readers interested in other aspects of random matrix theory, in addition to the classical book of Mehta [63], several excellent works are available that present random matrices in a broader scope. The books by Anderson, Guionnet and Zeitouni [10] and Pastur and Shcherbina [65] contain extensive material starting from the basics. Tao's book [72] provides a different aspect to this subject and is self-contained as a graduate textbook. Forrester's monograph [46] is a handbook for any explicit formulas related to random matrices. Finally, [8] is an excellent comprehensive overview of diverse applications of random matrix theory in mathematics, physics, neural networks and engineering.

Notational conventions. In order to focus on the essentials, we will not follow the dependence 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 $A \leqslant CB$. If A and B are comparable in the sense that $A \lesssim B$ and $B \lesssim A$, then we write $A \sim B$. In informal explanations, we will often use $A \approx B$ which indicates closeness in a not precisely specified sense. We introduce the notation $[\![A, B]\!] := \mathbb{Z} \cap [A, B]$ for the set of integers between any two real numbers $A < B$. We will usually denote vectors in $\mathbb{C}^N$ by boldface letters; $\mathbf{x} = (x_1, x_2, \ldots, x_N)$.

Acknowledgement. A special thank goes to Torben Krüger for many discussions and suggestions on the presentation of this material as well as for his careful proofreading and invaluable comments.

2 Tools

2.1 Stieltjes transform

In this section we introduce our basic tool, the **Stieltjes transform** of a measure. We denote by

$$\mathbb{H} := \{z \in \mathbb{C} \,:\, \operatorname{Im} z > 0\}$$

the (open) upper half of the complex plane.

Definition 2.1. *Let μ be a Borel probability measure on $\mathbb{R}$. Its Stiltjes transform at a* **spectral parameter** *$z \in \mathbb{H}$ is defined by*

$$m_\mu(z) := \int_{\mathbb{R}} \frac{\mathrm{d}\mu(x)}{x - z}. \tag{2.1}$$

Exercise 2.2. *The following three properties are straightforward to check:*

i) The Stieltjes transform $m_\mu(z)$ is analytic on $\mathbb{H}$ and it maps $\mathbb{H}$ to $\mathbb{H}$, i.e. $\operatorname{Im} m_\mu(z) > 0$.

ii) We have $-i\eta m_\mu(i\eta) \to 1$ as $\eta \to \infty$.

iii) We have the bound

$$|m_\mu(z)| \leqslant \frac{1}{\operatorname{Im} z}.$$

In fact, properties i)-ii) characterize the Stieltjes transform in a sense that if a function $m : \mathbb{H} \to \mathbb{H}$ satisfies i)–ii), then there exists a probability measure μ such that $m = m_\mu$ (for the proof, see e.g. Appendix B of [75]; it is also called the Nevanlinna's representation theorem).

From the Stieltjes transform one may recover the measure:

Lemma 2.3 (Inverse Stieltjes transform)**.** *Suppose that μ is a probability measure on $\mathbb{R}$ and let m_μ be its Stieltjes transform. Then for any $a < b$ we have*

$$\lim_{\eta\to 0}\frac{1}{\pi}\int_a^b \operatorname{Im} m_\mu(E+i\eta)\mathrm{d}E = \mu(a,b) + \frac{1}{2}\big[\mu(\{a\})+\mu(\{b\})\big]$$

Furthermore, if μ is absolutely continuous wrt. the Lebesgue measure, i.e. $\mu(\mathrm{d}E) = \mu(E)\mathrm{d}E$ with some density function $\mu(E) \in L^1$, then

$$\frac{1}{\pi}\lim_{\eta\to 0+}\operatorname{Im} m_\mu(E+i\eta) \to \mu(E)$$

pointwise for almost every E.

In particular, Lemma 2.3 guarantees that $m_\mu = m_\nu$ if and only of $\mu = \nu$, i.e. the Stieltjes transform uniquely characterizes the measure. Furthermore, pointwise convergence of a sequence of Stieltjes transforms is equivalent to weak convergence of the measures. More precisely, we have

Lemma 2.4. *Let μ_N be a sequence of probability measures and let $m_N(z) = m_{\mu_N}(z)$ be their Stieltjes transforms. Suppose that*

$$\lim_{N\to\infty} m_N(z) =: m(z)$$

exists for any $z \in \mathbb{H}$ and $m(z)$ satisfies property ii), i.e. $-i\eta m(i\eta) \to 1$ as $\eta \to \infty$. Then there exists a probability measure μ such that $m = m_\mu$ and μ_N converges to μ in distribution.

The proof can be found e.g. in [49] and it relies on Lemma 2.3 and Montel's theorem. The converse of Lemma 2.4 is trivial: if the sequence μ_N converges in distribution to a probability measure μ, then clearly $m_N(z) \to m_\mu(z)$ pointwise, since the Stieltjes transform for any fixed $z \in \mathbb{H}$ is just the integral of the continuous bounded function $x \to (x-z)^{-1}$. Note that the additional condition ii) is a compactness (tightness) condition, it prevents that part of the measures μ_N escape to infinity in the limit.

All these results are very similar to the Fourier transform (characteristic function)

$$\phi_\mu(t) := \int_{\mathbb{R}} e^{-itx}\mu(\mathrm{d}x)$$

of a probability measure. In fact, there is a direct connection between them;

$$\int_0^\infty e^{-\eta t}e^{itE}\phi_\mu(t)\mathrm{d}t = i\int_{\mathbb{R}}\frac{\mathrm{d}\mu(x)}{x-E-i\eta} = im_\mu(E+i\eta)$$

for any $\eta > 0$ and $E \in \mathbb{R}$. In particular, due to the regularizing factor $e^{-t\eta}$, the large t behavior of the Fourier transform $\phi(t)$ is closely related to the small $\eta \sim 1/t$ behavior of the Stieltjes transform.

Especially important is the imaginary part of the Stieltjes transform since

$$\operatorname{Im} m_\mu(z) = \int_{\mathbb{R}}\frac{\eta}{|x-E|^2+\eta^2}\mu(\mathrm{d}x), \qquad z = E+i\eta,$$

which can also be viewed as the convolution of μ with the Cauchy kernel on scale η:

$$P_\eta(E) = \frac{\eta}{E^2+\eta^2},$$

indeed

$$\operatorname{Im} m_\mu(E+i\eta) = (P_\eta \star \mu)(E).$$

Up to a normalization $1/\pi$, the Cauchy kernel is an approximate delta function on scale η. Clearly

$$\int_{\mathbb{R}}\frac{1}{\pi}\,P_\eta(E)\mathrm{d}E = 1$$

and the overwhelming majority of its mass is supported on scale η:

$$\int_{|E|\geqslant K\eta} \frac{1}{\pi} P_\eta(E) \leqslant \frac{2}{K}$$

for any K. Due to standard properties of the convolution, the *moral of the story* is that $\operatorname{Im} m_\mu(E+i\eta)$ **resolves the measure μ on a scale η around an energy E.**

Notice that the small η regime is critical; it is the regime where the integral in the definition of the Stieltjes transform (2.1) becomes more singular, and properties of the integral more and more depend on the local smoothness properties of the measure. In general, the regularity of the measure μ on some scales $\eta > 0$ is directly related to the Stieltjes transform $m(z)$ with $\operatorname{Im} z \approx \eta$.

The Fourier transform $\phi_\mu(t)$ of μ for large t also characterizes the local behavior of the measure μ on scales $1/t$, We will nevertheless work with the Stieltjes transform since for hermitian matrices (or self-adjoint operators in general) it is directly related to the resolvent, it is relatively easy to handle and it has many convenient properties.

Exercise 2.5. *Prove Lemma 2.3 by using Fubini's theorem and Lebesgue density theorem.*

2.2 Resolvent

Let $H = H^*$ be a hermitian matrix, then its resolvent at spectral parameter $z \in \mathbb{H}$ is defined as

$$G = G(z) = \frac{1}{H-z}, \qquad z \in \mathbb{H}.$$

In these notes, the spectral parameter z will always be in the upper half plane, $z \in \mathbb{H}$. We usually follow the convention that $z = E + i\eta$, where $E = \operatorname{Re} z$ will often be referred as "energy" alluring to the quantum mechanical interpretation of E.

Let μ_N be the normalized *empirical measure of the eigenvalues* of H:

$$\mu_N(\mathrm{d}x) = \frac{1}{N}\sum_{i=1}^N \delta(\lambda_i - x).$$

Then clearly the normalized trace of the resolvent is

$$\frac{1}{N}\operatorname{Tr} G(z) = \frac{1}{N}\sum_{i=1}^N \frac{1}{\lambda_i - z} = \int_{\mathbb{R}} \frac{\mu_N(\mathrm{d}x)}{x-z} = m_{\mu_N}(z) =: m_N(z)$$

exactly the Stieltjes transform of the empirical measure. This relation justifies why we focus on the Stieltjes transform; based upon Lemma 2.4, if we could identify the (pointwise) limit of $m_N(z)$, then the asymptotic eigenvalue density ϱ would be given by the inverse Stieltjes transform of the limit.

Since μ_N is a discrete (atomic) measure on small $(1/N)$ scales, it may behave very badly (i.e. it is strongly fluctuating and may blow up) for η smaller than $1/N$, depending on whether there happens to be an eigenvalue in an η-vicinity of $E = \operatorname{Re} z$. Since the eigenvalue spacing is (typically) of order $1/N$, for $\eta \ll 1/N$ there is no approximately deterministic ("self-averaging") behavior of m_N. However, as long as $\eta \gg 1/N$, we may hope a *law of large number phenomenon*; this would be equivalent to the fact that the eigenvalue density does not have much fluctuation above its inter-particle scale $1/N$. The local law on m_N down to the smallest possible (optimal) scale $\eta \gg 1/N$ will confirm this hope.

In fact, the resolvent carries much more information than merely its trace. In general the resolvent of a hermitian matrix is a very rich object: it gives information on the eigenvalues and eigenvectors for energies near the real part of the spectral parameter. For example, by spectral decomposition we have

$$G(z) = \sum_i \frac{|\mathbf{u}_i\rangle\langle\mathbf{u}_i|}{\lambda_i - z}$$

where $\mathbf{u}_i$ are the (ℓ^2-normalized) eigenvectors associated with λ_i. For example, the diagonal matrix elements of the resolvent at z are closely related to the eigenvectors with eigenvalues near $E = \operatorname{Re} z$:

$$G_{xx} = \sum_i \frac{|\mathbf{u}_i(x)|^2}{\lambda_i - z}, \qquad \operatorname{Im} G_{xx} = \sum_i \frac{\eta}{|\lambda_i - E|^2 + \eta^2} |\mathbf{u}_i(x)|^2.$$

Notice that for very small η, the factor $\eta/(|\lambda_i - E|^2 + \eta^2)$ effectively reduces the sum from all $i = 1, 2, \ldots, N$ to those indices where λ_i is η-close to E; indeed this factor changes from the very large value $1/\eta$ to a very small value η as i moves away. Roughly speaking

$$\operatorname{Im} G_{xx} = \sum_i \frac{\eta}{|\lambda_i - E|^2 + \eta^2} |\mathbf{u}_i(x)|^2 \approx \sum_{i : |\lambda_i - E| \lesssim \eta} \frac{\eta}{|\lambda_i - E|^2 + \eta^2} |\mathbf{u}_i(x)|^2.$$

This idea can be made rigorous at least as an upper bound on each summand. A physically important consequence will be that one may directly obtain ℓ^∞ bounds on the eigenvectors: for any fixed $\eta > 0$ we have

$$\|\mathbf{u}_i\|_\infty^2 := \max_x |\mathbf{u}_i(x)|^2 \leqslant \eta \cdot \max_x \max_{E \in \mathbb{R}} \operatorname{Im} G_{xx}(E + i\eta). \tag{2.2}$$

In other words, if we can control diagonal elements of the resolvent on some scale $\eta = \operatorname{Im} z$, then we can prove an $\sqrt{\eta}$-sized bound on the max norm of the eigenvector. The strongest result is always the smallest possible scale. Since the local law will hold down to scales $\eta \gg 1/N$, in particular we will be able to establish that $\operatorname{Im} G_{xx}(E + i\eta)$ remains bounded as long as $\eta \gg 1/N$, thus we will prove the **complete delocalization** of the eigenvectors:

$$\|\mathbf{u}_i\|_\infty \leqslant \frac{N^\varepsilon}{\sqrt{N}} \tag{2.3}$$

for any $\varepsilon > 0$ fixed, independent of N. Note that the bound (2.3) is optimal (apart from the N^ε factor) since clearly

$$\|\mathbf{u}\|_\infty \geqslant \frac{\|\mathbf{u}\|_2}{\sqrt{N}}$$

for any $\mathbf{u} \in \mathbb{C}^N$.

We also note that if $\operatorname{Im} G_{xx}(E + i\eta)$ can be controlled only for energies in a fixed subinterval $I \subset \mathbb{R}$, e.g. the local law holds only for all $E \in I$, the we can conclude complete delocalization for those eigenvectors whose eigenvalues lie in I.

2.3 The semicircle law for Wigner matrices via the moment method

This section introduces the traditional moment method to identify the semicircle law. We included this material for historical relevance, but it will not be needed later hence it can be skipped at first reading.

For large z one can expand m_N as follows

$$m_N(z) = \frac{1}{N} \operatorname{Tr} \frac{1}{H - z} = -\frac{1}{Nz} \sum_{m=0}^{\infty} \operatorname{Tr} \Big(\frac{H}{z} \Big)^m, \tag{2.4}$$

so after taking the expectation, we need to compute traces of high moments of H:

$$\mathbb{E}\, m_N(z) = \sum_{k=0}^{\infty} z^{-(2k+1)} \frac{1}{N} \mathbb{E} \operatorname{Tr} H^{2k}. \tag{2.5}$$

Here we tacitly used that the contributions of odd powers are algebraically zero, which clearly holds at least if we assume that h_{ij} have symmetric distribution for simplicity. Indeed, in this case H^{2k+1} and $(-H)^{2k+1}$ have the same distribution, thus

$$\mathbb{E} \operatorname{Tr} H^{2k+1} = \mathbb{E} \operatorname{Tr}(-H)^{2k+1} = -\mathbb{E} \operatorname{Tr} H^{2k+1}.$$

The computation of even powers, $\mathbb{E}\,\mathrm{Tr}\,H^{2k}$, reduces to a combinatorial problem. Writing out

$$\mathbb{E}\,\mathrm{Tr}\,H^{2k} = \sum_{i_1,i_2,\ldots i_{2k}} \mathbb{E} h_{i_1 i_2} h_{i_2 i_3} \ldots h_{i_{2k} i_1},$$

one notices that, by $\mathbb{E}h_{ij} = 0$, all those terms are zero where at least one $h_{i_j i_{j+1}}$ stands alone, i.e. is not paired with itself or its conjugate. This restriction poses a severe constraint on the relevant index sequences $i_1, i_2, \ldots, i_{2k}$. For the terms where an exact pairing of all the $2k$ factors is available, we can use $\mathbb{E}|h_{ij}|^2 = N^{-1}$ to see that all these terms contribute by N^{-k}. There are terms where three or more h's coincide, giving rise to higher moments of h, but their combinatorics is of lower order. Following Wigner's classical calculation (called the **moment method**, see e.g. [10]), one needs to compute the number of relevant index sequences that give rise to a perfect pairing and one finds that the leading term is given by the Catalan numbers, i.e.

$$\frac{1}{N}\mathbb{E}\,\mathrm{Tr}\,H^{2k} = \frac{1}{k+1}\binom{2k}{k} + O_k\big(\frac{1}{N}\big). \tag{2.6}$$

Notice that the N-factors cancelled out in the leading term.

Thus, continuing (2.5) and neglecting the error terms, we get

$$\mathbb{E}\, m_N(z) \approx -\sum_{k=0}^{\infty} \frac{1}{k+1}\binom{2k}{k} z^{-(2k+1)}, \tag{2.7}$$

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 (2.4) is valid only for large z, given that the limit is an analytic function of z, one can extend the relation

$$\lim_{N\to\infty} \mathbb{E} m_N(z) = \frac{1}{2}(-z+\sqrt{z^2-4}) \tag{2.8}$$

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.,

$$m_{sc}(z) := \frac{1}{2}(-z+\sqrt{z^2-4}) = \int_{\mathbb{R}} \frac{\varrho_{sc}(x)\mathrm{d}x}{x-z}, \qquad \varrho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}. \tag{2.9}$$

The square root function is chosen with a branch cut in the segment $[-2,2]$ so that $\sqrt{z^2-4} \sim z$ at infinity. This guarantees that $\mathrm{Im}\, m_{sc}(z) > 0$ for $\mathrm{Im}\, z > 0$.

Exercise 2.6. *As a simple calculus exercise, verify* (2.9). *Either use integration by parts, or compute the moments of the semicircle law and verify that they are given by the Catalan numbers, i.e.*

$$\int_{\mathbb{R}} x^{2k} \varrho_{sc}(x)\mathrm{d}x = \frac{1}{k+1}\binom{2k}{k}. \tag{2.10}$$

Since the Stieltjes transform identifies the measure uniquely, and pointwise convergence of Stieltjes transforms implies weak convergence of measures, we obtain

$$\mathbb{E}\,\varrho_N(\mathrm{d}x) \rightharpoonup \varrho_{sc}(x)\mathrm{d}x. \tag{2.11}$$

The relation (2.8) actually holds with high probability, i.e., for any z with $\mathrm{Im}\, z > 0$,

$$\lim_{N\to\infty} m_N(z) = \frac{1}{2}(-z+\sqrt{z^2-4}), \tag{2.12}$$

in probability, implying a similar strengthening of the convergence in (2.11). 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* sketched above utilized the fact that the moments of the semicircle density are given by the Catalan numbers (2.10), which also emerged as the normalized traces of powers of H, see (2.6). The *resolvent method* relies on the fact that m_N approximately satisfies a self-consistent equation,

$$m_N(z) \approx -\frac{1}{z+m_N(z)}, \tag{2.13}$$

that is very close to the quadratic equation that m_{sc} from (2.9) exactly satisfies:

$$m_{sc}(z) = -\frac{1}{z+m_{sc}(z)}. \tag{2.14}$$

Comparing these two equations, one finds that $m_N(z) \approx m_{sc}(z)$. Taking inverse Stieltjes transform, one concludes the semicircle law. In the next section we give more details on (2.13).

In other words, in the resolvent method the semicircle density emerges via a specific relation for its Stieltjes transform. The key relation (2.14) is the simplest form of the **Dyson equation**, or a **self-consistent equation** for the trace of the resolvent: later we will see a Dyson equation for the entire resolvent. It turns out that the resolvent approach allows us to perform a much more precise analysis than the moment method, especially in the short scale regime, where $\operatorname{Im} z$ 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 $\operatorname{Im} z$ will yield a local version of the semicircle law.

3 The resolvent method

In this section we sketch the two basic steps of the resolvent method for the simplest Wigner case but we will already make remarks preparing for the more complicated setup. The first step concerns the derivation of the approximate equation (2.13). This is a probabilistic step since $m_N(z)$ is a random object and even in the best case (2.13) can hold only with high probability. In the second step we compare the approximate equation (2.13) with the exact equation (2.14) to conclude that m_N and m_{sc} are close. We will view (2.13) as a perturbation of (2.14), so this step is about a stability property of the exact equation and it is a deterministic problem.

3.1 Probabilistic step

There are essentially two ways to obtain (2.13); either by Schur complement formula or by cumulant expansion. Typically the Schur method gives more precise results since it can be easier turned into a full asymptotic expansion, but it heavily relies on the independence of the matrix elements and that the resolvent of H is essentially diagonal. We now discuss these methods separately.

3.1.1 Schur complement method

The basic input is the following well-known formula from linear algebra:

Lemma 3.1 (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*

$$D := \begin{pmatrix} A & B^* \\ B & C \end{pmatrix} \tag{3.1}$$

and $n \times n$ matrix $\widehat{D}$ as

$$\widehat{D} := A - B^* C^{-1} B. \tag{3.2}$$

Then $\widehat{D}$ is invertible if D is invertible and for any $1 \leqslant i, j \leqslant n$, we have

$$(D^{-1})_{ij} = (\widehat{D}^{-1})_{ij} \tag{3.3}$$

for the corresponding matrix elements. □

We will use this formula for the resolvent of H. Recall that $G_{ij} = G_{ij}(z)$ denotes the matrix element of the resolvent

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

Let $H^{[i]}$ denote the i-th minor of H, i.e. the $(N-1)\times(N-1)$ matrix obtained from H by removing the i-th row and column:

$$H^{[i]}_{ab} := h_{ab}, \qquad a, b \neq i.$$

Similarly, we set

$$G^{[i]}(z) := \frac{1}{H^{[i]} - z}$$

to be the resolvent of the minor. For $i = 1$ we have the following block-decomposition of H

$$H = \begin{pmatrix} h_{11} & [\mathbf{a}^1]^* \\ \mathbf{a}^1 & H^{[1]} \end{pmatrix},$$

where $\mathbf{a}^i \in \mathbb{C}^{N-1}$ is the i-th column of H without the i-th element.

Using Lemma 3.1 for $n = 1$, $m = N-1$ we have

$$G_{ii} = \frac{1}{h_{ii} - z - [\mathbf{a}^i]^* G^{[i]} \mathbf{a}^i}, \tag{3.4}$$

where

$$[\mathbf{a}^i]^* G^{[i]} \mathbf{a}^i = \sum_{k,l \neq i} h_{ik} G^{[i]}_{kl} h_{li}. \tag{3.5}$$

We use the convention that unspecified summations always run from 1 to N.

Now we use the fact that for Wigner matrices $\mathbf{a}^i$ and $H^{[i]}$ are independent. So in the quadratic form (3.5) we can condition on the i-th minor and momentarily consider only the randomness of the i-th column. Set $i = 1$ for notational simplicity. Then we have a quadratic form of the type

$$\mathbf{a}^* B \mathbf{a} = \sum_{k,l=2}^{N} \bar{a}_k B_{kl} a_l$$

where $B = G^{[1]}$ is considered as a fixed deterministic matrix and $\mathbf{a}$ is a random vector with centered i.i.d. components and $\mathbb{E}|a_k|^2 = 1/N$. We decompose it into its expectation w.r.t. $\mathbf{a}$, denoted by $\mathbb{E}_{\mathbf{a}}$, and the fluctuation:

$$\mathbf{a}^* B \mathbf{a} = \mathbb{E}_{\mathbf{a}} \mathbf{a}^* B \mathbf{a} + Z, \qquad Z := \mathbf{a}^* B \mathbf{a} - \mathbb{E}_{\mathbf{a}} \mathbf{a}^* B \mathbf{a}. \tag{3.6}$$

The expectation gives

$$\mathbb{E}_{\mathbf{a}} \mathbf{a}^* B \mathbf{a} = \mathbb{E}_{\mathbf{a}} \sum_{k,l=2}^{N} \bar{a}_k B_{kl} a_l = \frac{1}{N} \sum_{k=2}^{N} B_{kk} = \frac{1}{N} \operatorname{Tr} B,$$

where we used that a_k and a_l are independent, $\mathbb{E}_{\mathbf{a}} \bar{a}_k a_l = \delta_{kl} \cdot \frac{1}{N}$, so the double sum collapses to a single sum. Neglecting the fluctuation Z for a moment (see an argument later), we have from (3.4) that

$$G_{11} = -\frac{1}{z + \frac{1}{N} \operatorname{Tr} G^{[1]} + \text{error}}, \tag{3.7}$$

where we also included the small $h_{11} \sim N^{-1/2}$ into the error term. Furthermore, it is easy to see that $\frac{1}{N} \operatorname{Tr} G^{[1]}$ and $\frac{1}{N} \operatorname{Tr} G$ are close to each other, this follows from a basic fact from linear algebra that the eigenvalues of H and its minor $H^{[1]}$ **interlace** (see Exercise 3.2).

Similar formula holds for each i, not only for $i = 1$. Summing them up, we have

$$\frac{1}{N}\operatorname{Tr} G \approx -\frac{1}{z + \frac{1}{N}\operatorname{Tr} G},$$

which is exactly (2.13), modulo the argument that the fluctuation Z is small. Notice that we were aiming only at $\frac{1}{N}\operatorname{Tr} G$, but in fact the procedure gave us more. After approximately identifying $\frac{1}{N}\operatorname{Tr} G \approx \frac{1}{N}\operatorname{Tr} G^{[1]}$ with m_{sc}, we can feed this information back to (3.7) to obtain information for each diagonal matrix element of the resolvent:

$$G_{11} \approx -\frac{1}{z + m_{sc}} = m_{sc},$$

i.e. not only the trace of G are close to m_{sc}, but each diagonal matrix element.

What about the offidiagonals? It turns out that they are small. The simplest argument to indicate this is using the **Ward identity** that is valid for resolvents of any self-adjoint operator T:

$$\sum_j \Big|\big(\frac{1}{T-z}\big)_{ij}\Big|^2 = \frac{1}{\operatorname{Im} z}\operatorname{Im}\big(\frac{1}{T-z}\big)_{ii}. \tag{3.8}$$

We recall that the imaginary part of a matrix M is given by $\operatorname{Im} M = \frac{1}{2i}(M - M^*)$ and notice that $(\operatorname{Im} M)_{aa} = \operatorname{Im} M_{aa}$ so there is no ambiguity in the notation of its diagonal elements. Notice that the summation in (3.8) is removed at the expense of a factor $1/\operatorname{Im} z$. So if $\eta = \operatorname{Im} z \gg 1/N$ and diagonal elements are controlled, the Ward identity is a substantial improvement over the naive bound of estimating each of the N terms separately. In particular, applying (3.8) for G we get

$$\sum_j |G_{ij}|^2 = \frac{1}{\operatorname{Im} z}\operatorname{Im} G_{ii}.$$

Since the diagonal elements have already been shown to be close to m_{sc}, we get

$$\frac{1}{N}\sum_j |G_{ij}|^2 \approx \frac{\operatorname{Im} m_{sc}}{N \operatorname{Im} z},$$

i.e. on average we have

$$|G_{ij}| \lesssim \frac{1}{\sqrt{N\eta}}, \qquad i \neq j.$$

With a bit more argument, one can show that this relation holds for every $j \neq i$ and not just on average. We thus showed that the resolvent G of a Wigner matrix is close to the m_{sc} times the identity matrix$I$, very roughly

$$G(z) \approx m_{sc}(z)I. \tag{3.9}$$

Such relation must be treated with a certain care, since G is a large matrix and the sloppy formulation in (3.9) does not indicate in which sense the closeness $\approx$ is meant. It turns our that it holds in **normalized trace sense**:

$$\frac{1}{N}\operatorname{Tr} G \approx m_{sc},$$

in **entrywise sense**:

$$G_{ij} \approx m_{sc}\delta_{ij} \tag{3.10}$$

for every fixed i, j; and more generally in **isotropic sense**:

$$\langle \mathbf{x}, G\mathbf{y}\rangle \approx m_{sc}\langle \mathbf{x}, \mathbf{y}\rangle$$

for every fixed (deterministic) vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$. In all cases, these relations are meant with very high probability. But (3.9) does **not** hold in operator norm sense, e.g

$$\|G\| = \frac{1}{\eta}, \qquad \text{while} \quad \|m_{sc}I\| = |m_{sc}| \sim O(1)$$

even if $\eta \to 0$. One may not invert (3.9) either, since the relation

$$H - z \approx \frac{1}{m_{sc}} I \tag{3.11}$$

is very wrong, in fact

$$H - z \approx -z$$

if we disregard small off-diagonal elements as we did in (3.10). The point is that the cumulative effects of many small off diagonal matrix elements substantially changes the matrix. In fact, using (2.14), the relation (3.10) in the form

$$\Big(\frac{1}{H-z}\Big)_{ij} \approx \frac{1}{-z - m_{sc}(z)} \delta_{ij} \tag{3.12}$$

exactly shows how much the spectral parameter must be shifted compared to the naive (and wrong) approximation $(H-z)^{-1} \approx -1/z$. This amount is $m_{sc}(z)$ and it is often called **self-energy shift** in the physics literature. On the level of the resolvent (and in the senses described above), the effect of the random matrix H can be simply described by this shift.

Finally, we indicate the mechanism that makes the fluctuation term Z in (3.6) small. We compute only its variance, higher moment calculations are similar but more involved:

$$\mathbb{E}_{\mathbf{a}} |Z|^2 = \sum_{mn} \sum_{kl} \mathbb{E}_{\mathbf{a}} \Big[a_m \bar{B}_{mn} \bar{a}_n - \mathbb{E}_{\mathbf{a}} a_m \bar{B}_{mn} \bar{a}_n \Big] \Big[\bar{a}_k B_{kl} a_l - \mathbb{E}_{\mathbf{a}} \bar{a}_k B_{kl} a_l \Big].$$

The summations run for all indices from 2 to N. Since $\mathbb{E}_{\mathbf{a}} a_m = 0$, in the terms with nonzero contribution we need to pair every a_m to another $\bar{a}_m$. For simplicity, here we assume that we work with the complex symmetry class and $\mathbb{E} a_m^2 = 0$ (i.e. the real and imaginary parts of each matrix elements h_{ij} are independent and identically distributed). If a_m is paired with $\bar{a}_n$ in the above sum, i.e. $m = n$, then this pairing is cancelled by the $\mathbb{E}_{\mathbf{a}} a_m \bar{B}_{mn} \bar{a}_n$ term. So a_i must be paired with an a from the other bracket and since $\mathbb{E} a^2 = 0$, it has to be paired with $\bar{a}_k$, thus $m = k$. Similarly $n = l$ and we get

$$\mathbb{E}_{\mathbf{a}} |Z|^2 = \frac{1}{N^2} \sum_{m \neq n} |B_{mn}|^2 + \mathbb{E}_{\mathbf{a}} |a|^4 \sum_m |B_{mm}|^2, \tag{3.13}$$

where the last term comes from the case when $m = n = k = l$. Assuming that the matrix elements h_{ij} have fourth moments in a sense that $\mathbb{E}|\sqrt{N} h_{ij}|^4 \leqslant C$, we have $\mathbb{E}_{\mathbf{a}} |a|^4 = O(N^{-2})$ in this last term and it is negligible. The main term in (3.13) has a summation over N^2 elements, so *a priori* it looks order one, i.e. too large. But in our application, B will be the resolvent of the minor, $B = G^{[1]}$, and we can use the Ward identity (3.8).

In our concrete application with $B = G^{[1]}$ we get

$$\mathbb{E}_{\mathbf{a}} |Z|^2 = \frac{1}{N\eta} \frac{1}{N} \sum_m \operatorname{Im} B_{mm} + \frac{C}{N^2} \sum_m |B_{mm}|^2 \leqslant \frac{C}{N\eta} \frac{1}{N} \operatorname{Im} \operatorname{Tr} G^{[1]} \leqslant \frac{C}{N\eta} \operatorname{Im} m^{[1]} = O\Big(\frac{1}{N\eta}\Big)$$

which is small, assuming $N\eta \gg 1$. To estimate the second term here we used that for the resolvent of any hermitian matrix T we have

$$\sum_m \Big| \Big(\frac{1}{T-z}\Big)_{mm} \Big|^2 \leqslant \frac{1}{\eta} \operatorname{Im} \operatorname{Tr} \frac{1}{T-z} \tag{3.14}$$

by spectral calculus. We also used that the traces of G and $G^{[1]}$ are close:

Exercise 3.2. *Let H be any hermitian matrix and $H^{[1]}$ its minor. Prove that their eigenvalues interlace, i.e. they satisfy*

$$\lambda_1 \leqslant \mu_1 \leqslant \lambda_2 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_{N-1} \leqslant \lambda_N,$$

where the λ's and μ's are the eigenvalues of H and $H^{[1]}$, respectively. Conclude from this that

$$\Big| \operatorname{Tr} \frac{1}{H-z} - \operatorname{Tr} \frac{1}{H^{[1]} - z} \Big| \leqslant \frac{1}{\operatorname{Im} z}$$

Exercise 3.3. *Prove the Ward identity* (3.8) *and the estimate* (3.14) *by using the spectral decomposition of* $T = T^*$.

3.1.2 Cumulant expansion

Another way to prove (2.13) starts with the defining identity of the resolvent: $HG = I + zG$ and computes its expectation:

$$\mathbb{E}HG = I + z\mathbb{E}G. \tag{3.15}$$

Here H and G are not independent, but it has the structure that the basic random variable H multiplies a function of it viewing $G = G(H)$. In a single random variable h it looks like $\mathbb{E}hf(h)$. If h were a centered real Gaussian, then we could use the basic **integration by parts** identity of Gaussian variables:

$$\mathbb{E}hf(h) = \mathbb{E}h^2 \mathbb{E}f'(h). \tag{3.16}$$

In our concrete application, when f is the resolvent whose derivative is its square, in the Gaussian case we have the formula

$$\mathbb{E}HG = -\mathbb{E}\widetilde{\mathbb{E}}\big[\widetilde{H}G\widetilde{H}\big]G, \tag{3.17}$$

where tilde denotes an independent copy of H. We may define a linear map $\mathcal{S}$ on the space of $N \times N$ matrices by

$$\mathcal{S}[R] := \widetilde{\mathbb{E}}\big[\widetilde{H}R\widetilde{H}\big], \tag{3.18}$$

then we can write (3.17) as

$$\mathbb{E}HG = -\mathbb{E}\mathcal{S}[G]G.$$

This indicates to smuggle the $\mathbb{E}HG$ term into $HG = I + zG$ and write it as

$$D = I + \big(z + \mathcal{S}[G]\big)G, \qquad D := HG + \mathcal{S}[G]G. \tag{3.19}$$

With these notations, (3.17) means that $\mathbb{E}D = 0$. Notice that the term $\mathcal{S}[G]G$ acts as a counter-term to balance HG.

Suppose we can prove that D is small with high probability, i.e. not only $\mathbb{E}D = 0$ but also $\mathbb{E}|D_{ij}|^2$ is small for any i, j, then

$$I + \big(z + \mathcal{S}[G]\big)G \approx 0. \tag{3.20}$$

So it is not unreasonable to hope that the solution G will be, in some sense, close to the solution M of the deterministic equation

$$I + \big(z + \mathcal{S}[M]\big)M = 0 \tag{3.21}$$

with the side condition that $\operatorname{Im} M := \frac{1}{2i}(M - M^*) \geqslant 0$ (positivity in the sense of hermitian matrices). It turns out that this equation in its full generality will play a central role in our analysis for much larger class of random matrices, see Section 4.4 later. The operator $\mathcal{S}$ is called the **self-energy operator** following the analogy explained around (3.12).

To see how $\mathcal{S}$ looks like, in the real Gaussian Wigner case (GOE) we have

$$\mathcal{S}[R]_{ij} = \widetilde{\mathbb{E}}\big[\widetilde{H}R\widetilde{H}\big]_{ij} = \widetilde{\mathbb{E}}\sum_{ab}\widetilde{h}_{ia}R_{ab}\widetilde{h}_{bj} = \delta_{ij}\frac{1}{N}\operatorname{Tr} R + \frac{1}{N}R_{ji}.$$

Plugging this relation back into (3.20) with $R = G$ and neglecting the second term $\frac{1}{N}G_{ji}$ we have

$$0 \approx I + \big(z + \frac{1}{N}\operatorname{Tr} G\big)G.$$

Taking the normalized trace, we end up with

$$1 + (z + m_N)m_N \approx 0, \tag{3.22}$$

i.e. we proved (2.13).

Exercise 3.4. *Prove* (3.16) *by a simple integration by parts and then use* (3.16) *to prove* (3.17)*. Formulate and prove the complex versions of these formulas (assume that* $\operatorname{Re} h$ *and* $\operatorname{Im} h$ *are independent).*

Exercise 3.5. *Compute the variance* $\mathbb{E}|D|^2$ *and conclude that it is small in the regime where* $N\eta \gg 1$ *(essentially as* $(N\eta)^{-1/2}$*). Compute* $\mathbb{E}\big|\frac{1}{N}\operatorname{Tr} D\big|^2$ *as well and show that it is essentially of order* $(N\eta)^{-1}$.

This argument so far heavily used that H is Gaussian. However, the basic integration by parts formula (3.16) can be extended to non-Gaussian situation. For this, we recall the **cumulants** of random variables. We start with a single random variable h. As usual, its moments are defined by

$$m_k := \mathbb{E}h^k,$$

and they are generated by the *moment generating function*

$$\mathbb{E}e^{th} = \sum_{k=0}^{\infty} \frac{t^k}{k!} m_k$$

(here we assume that all moments exist and even the exponential moment exists at least for small t). The **cumulants** κ_k of h are the Taylor coefficients of the *logarithm* of the moment generating function, i.e. they are defined by the identity

$$\log \mathbb{E}e^{th} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \kappa_k.$$

The sequences of $\{m_k \;:\; k = 0,1,2\ldots\}$ and $\{\kappa_k \;:\; k = 0,1,2\ldots\}$ mutually determine each other; these relations can be obtained from formal power series manipulations. For example

$$\kappa_0 = m_0 = 1, \qquad \kappa_1 = m_1, \qquad \kappa_2 = m_2 - m_1^2, \qquad \kappa_3 = m_3 - 3m_2m_1 + 2m_1^3, \ldots$$

and

$$m_1 = \kappa_1, \qquad m_2 = \kappa_2 + \kappa_1^2, \qquad m_3 = \kappa_3 + 3\kappa_2\kappa_1 + 2\kappa_1^3, \ldots$$

The general relations are given by

$$m_k = \sum_{\pi\in\Pi_k} \prod_{B\in\pi} \kappa_{|B|}, \qquad \kappa_k = \sum_{\pi\in\Pi_k} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B\in\pi} m_{|B|}, \tag{3.23}$$

where Π_k is the set of all partitions of a k-element base set, say $\{1,2,\ldots,k\}$. Such partition π consists of a collection of nonempty, mutually disjoint sets $\pi = \{B_1, B_2, \ldots B_{|\pi|}\}$ such that $\cup B_i = \{1,2,\ldots,k\}$ and $B_i \cap B_j = \emptyset$, $i \neq j$.

For Gaussian variables, all but the first and second cumulants vanish, $\kappa_3 = \kappa_4 = \ldots = 0$, and this is the reason for the very simple form of the relation (3.16). For general non-Gaussian h we have

$$\mathbb{E}hf(h) = \sum_{k=0}^{\infty} \frac{\kappa_{k+1}}{k!} \mathbb{E}f^{(k)}(h). \tag{3.24}$$

Similarly to the Taylor expansion, one does not have to expand it up to infinity, there are versions of this formula containing only a finite number of cumulants plus a remainder term.

To see the formula (3.24), we use Fourier transform:

$$\hat{f}(t) = \int_{\mathbb{R}} e^{ith} f(h)\mathrm{d}h, \qquad \hat{\mu}(t) = \int_{\mathbb{R}} e^{ith}\mu(\mathrm{d}h) = \mathbb{E}e^{ith},$$

where μ is the distribution of h, then

$$\log \hat{\mu}(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \kappa_k.$$

By Parseval identity (neglecting 2π's and assuming f is real)

$$\mathbb{E}hf(h) = \int_{\mathbb{R}} hf(h)\mu(\mathrm{d}h) = i\int_{\mathbb{R}} \overline{\hat{f}'(t)}\hat{\mu}(t)\mathrm{d}t.$$

Integration by parts gives

$$i\int_{\mathbb{R}} \overline{\hat{f}'(t)}\hat{\mu}(t)\mathrm{d}t = -i\int_{\mathbb{R}} \overline{\hat{f}(t)}\hat{\mu}'(t)\mathrm{d}t = -i\int_{\mathbb{R}} \overline{\hat{f}(t)}\hat{\mu}(t)\big(\log\hat{\mu}(t)\big)'\mathrm{d}t$$

$$= \sum_{k=0}^{\infty}\frac{\kappa_{k+1}}{k!}\int_{\mathbb{R}}(it)^k\overline{\hat{f}(t)}\hat{\mu}(t)\mathrm{d}t = \sum_{k=0}^{\infty}\frac{\kappa_{k+1}}{k!}\mathbb{E}f^{(k)}(h)$$

by Parseval again.

So far we considered one random variable only, but joint cumulants can also be defined for any number of random variables. This becomes especially relevant beyond the independent case, e.g. when the entries of the random matrix have correlations. For the Wigner case, many of these formulas simplify, but it is useful to introduce joint cumulants in full generality.

If $\mathbf{h} = (h_1, h_2, \dots h_m)$ is a collection of random variables (with possible repetition), then

$$\kappa(\mathbf{h}) = \kappa(h_1, h_2, \dots h_m)$$

are the coefficients of the logarithm of the moment generating function:

$$\log\mathbb{E}e^{\mathbf{t}\cdot\mathbf{h}} = \sum_{\mathbf{k}=0}^{\infty}\frac{\mathbf{t}^{\mathbf{k}}}{\mathbf{k}!}\kappa_{\mathbf{k}}.$$

Here $\mathbf{t} = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$, and $\mathbf{k} = (k_1, k_2, \dots, k_n) \in \mathbb{N}^n$ is a multi index with n components and

$$\mathbf{t}^{\mathbf{k}} := \prod_{i=1}^{n} t_i^{k_i}, \qquad \mathbf{k}! = \prod_i k_i!, \qquad \kappa_{\mathbf{k}} = \kappa(h_1, h_1, \dots h_2, h_2, \dots),$$

where h_j appears k_j-times (order is irrelevant, the cumulants are fully symmetric functions in all their variables). The formulas (3.23) naturally generalize, see e.g. Appendix A of [42] for a good summary. The analogue of (3.24) is

$$\mathbb{E}h_1 f(\mathbf{h}) = \sum_{\mathbf{k}}\frac{\kappa_{\mathbf{k}+\mathbf{e}_1}}{\mathbf{k}!}\mathbb{E}f^{(\mathbf{k})}(\mathbf{h}), \qquad \mathbf{h} = (h_1, h_2, \dots, h_n), \tag{3.25}$$

where the summation is for all n-multiindices and

$$\mathbf{k}+\mathbf{e}_1 = (k_1+1, k_2, k_3, \dots, k_n)$$

and the proof is the same.

We use these cumulant expansion formulas to prove that D defined in (3.19) is small with high probability by computing $\mathbb{E}|D_{ij}|^{2p}$ with large p. Written as

$$\mathbb{E}|D_{ij}|^{2p} = \mathbb{E}\big(HG + \mathcal{S}[G]G\big)_{ij}D_{ij}^{p-1}\bar{D}_{ij}^{p},$$

we may use (3.25) to do an integration by parts in the first H factor, considering everything else as a function f. It turns out that the $\mathcal{S}[G]G$ term cancels the second order cumulant and naively the effect of higher order cumulants are negligible since a cumulant of order k is $N^{-k/2}$. However, the derivatives of f can act on the $D^{p-1}\bar{D}^p$ part of f, resulting in a complicated combinatorics and in fact many cumulants need to be tracked, see [42] for an extensive analysis.

3.2 Deterministic stability step

In this step we compare the approximate equation (2.13) satisfied by the empirical Stieltjes transform and the exact equation (2.14) for the limiting Stieltjes transform

$$m_N(z) \approx -\frac{1}{z+m_N(z)}, \qquad m_{sc}(z) = -\frac{1}{z+m_{sc}(z)}.$$

In fact, considering the format (3.20) and (3.22), sometimes it is better to relate the following two equations

$$1+(z+m_N)m_N \approx 0, \qquad 1+(z+m_{sc})m_{sc} = 0.$$

This distinction is irrelevant for Wigner matrices, where the basic object to investigate is m_N, a scalar quantity – multiplying an equation with it is a trivial operation. But already (3.19) indicates that there is an approximate equation for the entire resolvent G as well and not only for its trace and in general we are interested in resolvent matrix elements as well. Since inverting G is a nontrivial operation (see the discussion after (3.9)), the three possible versions of (3.19) are very different:

$$I+(z+\mathcal{S}[G])G \approx 0, \qquad G \approx -\frac{1}{z+\mathcal{S}[G]}, \qquad -\frac{1}{G} \approx z+\mathcal{S}[G]$$

In fact the last version is blatantly wrong, see (3.11). The first version is closer to the spirit of the cumulant expansion method, the second is closer to Schur formula method.

In both cases, we need to understand the stability of the equation

$$m_{sc}(z) = -\frac{1}{z+m_{sc}(z)} \quad \text{or} \quad 1+(z+m_{sc})m_{sc} = 0$$

against a small additive perturbation. For definiteness, we look at the second equation and compare m_{sc} with m_ε, where m_ε solves

$$1+(z+m_\varepsilon)m_\varepsilon = \varepsilon$$

for some small ε. Since these are quadratic equations, one may write up the solutions explicitly and compare them, but this approach will not work in the more complicated situations. Instead, we subtract these two equations and find that

$$(z+2m_{sc})(m_\varepsilon - m_{sc}) + (m_\varepsilon - m_{sc})^2 = \varepsilon$$

We may also eliminate z using the equation $1+(z+m_{sc})m_{sc} = 0$ and get

$$\frac{m_{sc}^2-1}{m_{sc}}(m_\varepsilon - m_{sc}) + (m_\varepsilon - m_{sc})^2 = \varepsilon. \tag{3.26}$$

This is a quadratic equation for the difference $m_\varepsilon - m_{sc}$ and its stability thus depends on the invertibility of the linear coefficient $(m_{sc}^2-1)/m_{sc}$, which is determined by the limiting equation only. If we knew that

$$|m_{sc}| \leqslant C, \qquad |m_{sc}^2 - 1| \geqslant c \tag{3.27}$$

with some positive constants c, C, then the linear coefficient would be invertible

$$\left| \Big[\frac{m_{sc}^2-1}{m_{sc}}\Big]^{-1} \right| \leqslant C/c \tag{3.28}$$

and (3.26) would imply that

$$|m_\varepsilon - m_{sc}| \leqslant C'\varepsilon$$

at least if we had an a priori information that $|m_\varepsilon - m_{sc}| \leqslant c/2C$. This a priori information can be obtained for large $\eta = \operatorname{Im} z$ easily since in this regime both m_{sc} and m_ε are of order η (we still remember that m_ε represents a Stieltjes transform). Then we can use a fairly standard continuity argument to reduce $\eta = \operatorname{Im} z$

and keeping $E = \operatorname{Re} z$ fixed to see that the bound $|m_\varepsilon - m_{sc}| \leqslant c/2C$ holds for small η as well, as long as the perturbation $\varepsilon = \varepsilon(\eta)$ is small.

Thus the key point of the stability analysis is to show that the **stability constant** (later: operator/matrix) given in (3.28) is bounded. As indicated in (3.27), the control of the stability constant typically will have two ingredients: we need

(i) an upper bound on m_{sc}, the solution of the deterministic Dyson equation (2.14);
(ii) an upper bound on the inverse of $1 - m_{sc}^2$.

In the Wigner case, when m_{sc} is explicitly given (2.9), both bounds are easy to obtain. In fact, m_{sc} remains bounded for any z, while $1 - m_{sc}^2$ remains separated away from zero except near two special values of the spectral parameter: $z = \pm 2$. These are exactly the edges of the semicircle law, where an instability arises since here $m_{sc} \approx \pm 1$ (the same instability can be seen from the explicit solution of the quadratic equation).

We will see that it is not a coincidence: the edges of the asymptotic density ϱ are always the critical points where the stability constant blows up. These regimes require more careful treatment which typically consists in exploiting the fact that the error term D is proportional with the local density, hence it is also smaller near the edge. This additional smallness of D competes with the deteriorating upper bound on the stability constant near the edge.

In these notes we will focus on the behavior in the bulk, i.e. we consider spectral parameters $z = E + i\eta$ where $\varrho(E) \geqslant c > 0$ with some fixed positive constants. This will simplify many estimates. The regimes where E is separated away from the support of ϱ are even easier and we will not consider them here. The edge analysis is more complicated and we refer the reader to the original papers.

4 Models of increasing complexity

In this section we introduce subsequent the generalizations of the original Wigner ensemble. We also mention the key features of their resolvent that will be proven later along the local laws. The $N \times N$ matrix

$$H = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1N} \\ h_{21} & h_{22} & \dots & h_{2N} \\ \vdots & \vdots & & \vdots \\ h_{N1} & h_{N2} & \dots & h_{NN} \end{pmatrix} \tag{4.1}$$

will always be hermitian, $H = H^*$ and centered, $\mathbb{E}H = 0$. The distinction between real symmetric and complex hermitian cases play no role here; both symmetry classes are allowed. Many quantities, such as the distribution of H, the matrix of variances S, naturally depend on N, but for notational simplicity we will often omit this dependence from the notation.

We will always assume that we are in the mean field regime, i.e. the typical size of the matrix elements is of order $N^{-1/2}$ in a high moment sense:

$$\max_{ij} \mathbb{E}\big|\sqrt{N} h_{ij}\big|^p \leqslant \mu_p \tag{4.2}$$

for any p with some sequence of constants μ_p. This strong moment condition can be substantially relaxed but we will not focus on this direction.

4.1 Wigner matrix

We assume that the matrix elements of H are independent (up to the hermitian symmetry) and identically distributed. We choose the normalization such that

$$\mathbb{E}|h_{ij}|^2 = \frac{1}{N},$$

see (1.4) for explanation. The asymptotic density of eigenvalues (also called the **asymptotic density of states**) is the semicircle law, $\varrho_{sc}(x)$ (1.9) and its Stieltjes transform $m_{sc}(z)$ is given explicitly in (2.9). The corresponding self-consistent (deterministic) equation (Dyson equation) is a **scalar equation**

$$1 + (z+m)m = 0, \qquad \operatorname{Im} m > 0,$$

that is solved by $m = m_{sc}$. The stability "operator" is just the constant

$$\frac{1}{1-m^2}, \qquad m = m_{sc}.$$

The resolvent $G(z) = (H-z)^{-1}$ is approximately constant diagonal in the **entrywise sense**, i.e.

$$G_{ij}(z) \approx \delta_{ij} m_{sc}(z). \tag{4.3}$$

In particular, the diagonal elements are approximately the same

$$G_{ii} \approx G_{jj} \approx m_{sc}(z).$$

This also implies that the normalized trace (Stieltjes transform of the empirical eigenvalue density) is close to m_{sc}

$$m_N(z) = \frac{1}{N} \operatorname{Tr} G(z) \approx m_{sc}(z), \tag{4.4}$$

which we often call an approximation in **average (or tracial) sense**.

Moreover, G is also diagonal in **isotropic sense**, i.e. for any vectors $\mathbf{x}, \mathbf{y}$ (more precisely, any sequence of vectors $\mathbf{x}^{(N)}, \mathbf{y}^{(N)} \in \mathbb{C}^N$) we have

$$G_{\mathbf{xy}} := \langle \mathbf{x}, G\mathbf{y} \rangle \approx m_{sc}(z) \langle \mathbf{x}, \mathbf{y} \rangle. \tag{4.5}$$

In Section 4.5 we will comment on the precise meaning of $\approx$ in this context, incorporating the fact that G is random.

If these relations hold for any fixed $\eta = \operatorname{Im} z$, independent of N, then we talk about **global law**. If they hold down to $\eta \geqslant N^{-1+\gamma}$ with some $\gamma \in (0,1)$, then we talk about **local law**. If $\gamma > 0$ can be chosen arbitrarily small (independent of N), than we talk about **local law on the optimal scale**.

4.2 Generalized Wigner matrix

We assume that the matrix elements of H are independent (up to the hermitian symmetry), but not necessarily identically distributed. We define the matrix of variances as

$$S := \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1N} \\ s_{21} & s_{22} & \dots & s_{2N} \\ \vdots & \vdots & & \vdots \\ s_{N1} & s_{N2} & \dots & s_{NN} \end{pmatrix}, \qquad s_{ij} := \mathbb{E}|h_{ij}|^2. \tag{4.6}$$

We assume that

$$\sum_{j=1}^N s_{ij} = 1, \qquad \text{for every} \quad i = 1, 2, \dots, N, \tag{4.7}$$

i.e., the deterministic $N \times N$ matrix of variances, $S = (s_{ij})$, is symmetric and doubly stochastic. The key point is that the row sums are all the same. The fact that the sum in (4.7) is exactly one is a chosen normalization. The original Wigner ensemble is a special case, $s_{ij} = \frac{1}{N}$.

Although generalized Wigner matrices form a bigger class than the Wigner matrices, the key results are exactly the same. The asymptotic density of states is still the semicircle law, G is constant diagonal in entrywise and isotropic sense:

$$G_{ij} \approx \delta_{ij} m_{sc} \qquad \text{and} \quad G_{\mathbf{xy}} = \langle \mathbf{x}, G\mathbf{y} \rangle \approx m_{sc} \langle \mathbf{x}, \mathbf{y} \rangle.$$

In particular, the diagonal elements are approximately the same

$$G_{ii} \approx G_{jj}$$

and we have the same averaged law

$$m_N(z) = \frac{1}{N} \operatorname{Tr} G(z) \approx m_{sc}(z).$$

However, within the proof some complications arise. Although eventually G_{ii} turns out to be essentially independent of i, there is no a-priori complete permutation symmetry among the indices. We will need to consider the equations for each G_{ii} as a coupled system of N equations. The corresponding Dyson equation is a genuine **vector equation** of the form

$$1 + (z + (S\mathbf{m})_i)m_i = 0, \qquad i = 1, 2, \dots N \tag{4.8}$$

for the unknown N-vector $\mathbf{m} = (m_1, m_2, \dots, m_N)$ with $m_j \in \mathbb{H}$ and we will see that $G_{jj} \approx m_j$. The matrix S may also be called **self-energy matrix** according to the analogy explained around (3.12). Owing to (4.7), the solution to (4.8) is still the constant vector $m_i = m_{sc}$, but the stability operator depends on S and it is given by the matrix

$$\frac{1}{1 - m_{sc}^2 S}.$$

4.3 Wigner type matrix

We still assume that the matrix elements are independent, but we impose no special algebraic condition on the variances S. For normalization purposes, we will assume that $\|S\|$ is bounded, independently of N, this guarantees that the spectrum of H also remains bounded. We only require an upper bound of the form

$$\max_{ij} s_{ij} \leqslant \frac{C}{N} \tag{4.9}$$

for some constant C. This is a typical mean field condition, it guarantees that no matrix element is too big. Notice that at this stage there is no requirement for a lower bound, i.e. some s_{ij} may vanish. However, the analysis becomes considerably harder if large blocks of S can become zero, so for pedagogical convenience later in these notes we will assume that $s_{ij} \geqslant c/N$ for some $c > 0$.

The corresponding Dyson equation is the same **vector Dyson equation** as (4.8):

$$1 + (z + (S\mathbf{m})_i)m_i = 0, \qquad i = 1, 2, \dots N \tag{4.10}$$

but the solution is not constant any more. We will see that the system of equations (4.10) still has a unique solution $\mathbf{m} = (m_1, m_2, \dots, m_N)$ under the side condition $m_j \in \mathbb{H}$, but the components of $\mathbf{m}$ may differ and they are not given by m_{sc} any more.

The components m_i approximate the diagonal elements of the resolvent G_{ii}. Correspondingly, their average

$$\langle \mathbf{m} \rangle := \frac{1}{N} \sum_i m_i, \tag{4.11}$$

is the Stieltjes transform of a measure ϱ that approximates the empirical density of states. We will call this measure the **self-consistent density of states** since it is obtained from the self-consistent Dyson equation. In general there is no explicit formula for ϱ, it has to be computed by taking the inverse Stieltjes transform of $\langle \mathbf{m}(z) \rangle$:

$$\varrho(\mathrm{d}\tau) = \lim_{\eta \to 0+} \frac{1}{\pi} \operatorname{Im} \langle \mathbf{m}(\tau + i\eta) \rangle \mathrm{d}\tau. \tag{4.12}$$

No simple closed equation is known for the scalar quantity $\langle \mathbf{m}(z) \rangle$, even if one is interested only in the self-consistent density of states or its Stieltjes transform, the only known way to compute it is to solve (4.10) first and then take the average of the solution vector. Under some further conditions on S, the density of states is supported on finitely many intervals, it is real analytic away from the edges of these intervals and it has a specific singularity structure at the edges, namely it can have either square root singularity or cubic root cusp, see Section 6.1 later.

The resolvent is still approximately diagonal and it is given by the i-th component of $\mathbf{m}$:

$$G_{ij}(z) \approx \delta_{ij} m_i(z),$$

but in general

$$G_{ii} \not\approx G_{jj}, \qquad i \neq j.$$

Accordingly, the isotropic law takes the form

$$G_{\mathbf{xy}} = \langle \mathbf{x}, G\mathbf{y} \rangle \approx \langle \bar{\mathbf{x}} \mathbf{m} \mathbf{y} \rangle$$

and the averaged law

$$\frac{1}{N} \operatorname{Tr} G \approx \langle \mathbf{m} \rangle.$$

Recalling the notation of entrywise product of vectors, we have $\langle \mathbf{xmy} \rangle = \frac{1}{N} \sum_i x_i m_i y_i$.

The stability operator is

$$\frac{1}{1 - \mathbf{m}^2 S}, \tag{4.13}$$

where $\mathbf{m}^2$ is understood as an entrywise multiplication, so the linear operator $\mathbf{m}^2 S$ acts on any vector $\mathbf{x} \in \mathbb{C}^N$ as

$$[(\mathbf{m}^2 S)\mathbf{x}]_i := m_i^2 \sum_j s_{ij} x_j.$$

Notational convention. Sometimes we write the equation (4.10) in the concise vector form as

$$-\frac{1}{\mathbf{m}} = z + S\mathbf{m}.$$

Here we introduce the convention that for any vector $\mathbf{m} \in \mathbb{C}^N$ and for any function $f : \mathbb{C} \to \mathbb{C}$, the symbol $f(\mathbf{m})$ denotes the N-vector with components $f(m_j)$, i.e.

$$f(\mathbf{m}) := \big(f(m_1), f(m_2), \ldots, f(m_N)\big), \qquad \text{for any } \mathbf{m} = (m_1, m_2, \ldots, m_N).$$

In particular, $1/\mathbf{m}$ is the vector of the reciprocals $1/m_i$. Similarly, the entrywise product of two N-vectors $\mathbf{x}, \mathbf{y}$ is denoted by $\mathbf{xy}$; this is the N-vector with components

$$(\mathbf{xy})_i := x_i y_i$$

and similarly for products of more than two factors. Finally $\mathbf{x} \leqslant \mathbf{y}$ for real vectors means $x_i \leqslant y_i$ for all i.

4.3.1 A remark on the density of states

The Wigner type matrix is the first ensemble where the various concepts of density of states truly differ. The wording ”density of states” has been used slightly differently by various authors in random matrix theory; here we use the opportunity to clarify this point. Typically, in the physics literature the **density of states** means the statistical average of the **empirical density of states** μ_N defined in (1.8), i.e.

$$\mathbb{E}\mu_N(\mathrm{d}\tau) = \mathbb{E}\frac{1}{N}\sum_{i=1}^{N} \delta(\lambda_i - \tau).$$

This object depends on N, but very often it has a limit (in a weak sense) as N, the system size, goes to infinity. The limit, if exists, is often called the **limiting density of states**.

In general it is not easy to find μ_N or its expectation; the vector Dyson equation is essentially the only way to proceed. However, the quantity computed in (4.12), called the **self-consistent density of states**, is not exactly the density of states, it is only a good approximation. The local law states that the empirical (random) eigenvalue density μ_N can be very well approximated by the self-consistent density of states, computed from the Dyson equation and (4.12). Here "very well" means in high probability and with an explicit error bound of size $1/N\eta$, i.e. on larger scales we have more precise bound, but we still have closeness even down to scales $\eta \geqslant N^{-1+\gamma}$. High probability bounds imply that also the density of states $\mathbb{E}\mu_N$ is close to the self-consistent density of states ϱ, but in general they are not the same. Note that the significance of the local law is to approximate a random quantity with a deterministic one if N is large; there is no direct statement about any $N \to \infty$ limit. The variance matrix S depends on N and a-priori there is no relation between S-matrices for different N's.

In some cases a limiting version of these objects also exists. For example, if the variances s_{ij} arise from a deterministic nonnegative profile function $S(x,y)$ on $[0,1]^2$ with some regularity, i.e.

$$s_{ij} = \frac{1}{N} S\Big(\frac{i}{N}, \frac{j}{N}\Big),$$

then the sequence of the approximating density of states $\varrho^{(N)}$ have a limit. If the global law holds, then this limit must be the limiting density of states, defined as the limit of $\mathbb{E}\mu_N$. This is the case for Wigner matrices in a trivial way: the self-consistent density of states is always the semicircle for any N. However, the density of states for finite N is not the semicircle law; it depends on the actual distribution of the matrix elements, but decreasingly as N increases.

In these notes we will focus on computing the self-consistent density of states and proving local laws for fixed N; we will not consider the possible large N limits of these objects.

4.4 Correlated random matrix

For this class we drop the independence condition, so the matrix elements of H may have nontrivial correlations in addition to the one required by the hermitian symmetry $h_{ij} = \bar{h}_{ji}$. The Dyson equation is still determined by the second moments of H, but the covariance structure of all matrix elements is not described by a matrix; but by a four-tensor. We already introduced in (3.18) the necessary "super operator"

$$\mathcal{S}[R] := \mathbb{E}HRH$$

acting linearly on the space of $N \times N$ matrices R. Explicitly

$$\mathcal{S}[R]_{ij} = \mathbb{E}\sum_{ab} h_{ia}R_{ab}h_{bj} = \sum_{ab} \big[\mathbb{E}h_{ia}h_{bj}\big]R_{ab}.$$

The analogue of the upper bound (4.9) is

$$\mathcal{S}[R] \leqslant C\langle R\rangle$$

for any positive definite matrix $R \geqslant 0$, where we introduced the notation

$$\langle R\rangle := \frac{1}{N}\operatorname{Tr} R.$$

In the actual proofs we will need a lower bound of the form $\mathcal{S}[R] \geqslant c\langle R\rangle$ and further conditions on the decay of correlations among the matrix elements of H.

The corresponding Dyson equation becomes a **matrix equation**

$$I + (z + \mathcal{S}[M])M = 0 \tag{4.14}$$

for the unknown matrix $M = M(z) \in \mathbb{C}^{N\times N}$ under the constraint that $\operatorname{Im} M \geqslant 0$. Recall that the imaginary part of any matrix is a hermitian matrix defined by

$$\operatorname{Im} M = \frac{1}{2i}(M - M^*).$$

In fact, one may add a hermitian **external source matrix** $A = A^*$ and consider the more general equation

$$I + (z + A + \mathcal{S}[M])M = 0. \tag{4.15}$$

In random matrix applications, A plays the role of the matrix of expectations, $A = \mathbb{E}H$. We will call (4.15) and (4.14) the **matrix Dyson equation** with or without external source. The equation (4.15) has a unique solution and in general it is a non-diagonal matrix even if A is diagonal. Notice that the Dyson equation contains only the second moments of the elements of H via the operator $\mathcal{S}$; no higher order correlations appear, although in the proofs of the local laws further conditions on the correlation decay are necessary.

The Stieltjes transform of the density of states is given by

$$\langle M(z)\rangle = \frac{1}{N}\operatorname{Tr} M(z).$$

The matrix $M = M(z)$ approximates the resolvent in the usual senses, i.e. we have

$$G_{ij}(z) \approx M_{ij}(z),$$

$$\langle \mathbf{x}, G\mathbf{y}\rangle \approx \langle \mathbf{x}, M\mathbf{y}\rangle,$$

and

$$\frac{1}{N}\operatorname{Tr} G \approx \langle M\rangle.$$

Since in general M is not diagonal, the resolvent G is not approximately diagonal any more. We will call M, the solution to the matrix Dyson equation (4.15), the **self-consistent Green function** or **self-consistent resolvent**.

The stability operator is of the form

$$\frac{1}{I - \mathcal{C}_M\mathcal{S}},$$

where $\mathcal{C}_M$ is the linear map acting on the space of matrices as $\mathcal{C}_M[R] := MRM$. In other words, the stability operator is the inverse of the linear map $R \to R - M\mathcal{S}[R]M$ on the space of matrices.

The independent case (Wigner type matrix) is a special case of the correlated ensemble and it is interesting to exhibit their relation. In this case the super-operator $\mathcal{S}$ maps diagonal matrices to diagonal matrix. For any vector $\mathbf{v} \in \mathbb{C}^N$ we denote by $\operatorname{diag}(\mathbf{v})$ the $N \times N$ diagonal matrix with $(\operatorname{diag}(\mathbf{v}))_{ii} = v_i$ in the diagonal. Then we have, for the independent case with $s_{ab} := \mathbb{E}|h_{ab}|^2$ as before,

$$\big(\mathcal{S}[\operatorname{diag}(\mathbf{v})]\big)_{ij} = \sum_a \big[\mathbb{E}\bar{h}_{ai}h_{aj}\big]v_a = \delta_{ij}(S\mathbf{v})_i,$$

thus

$$\mathcal{S}[\operatorname{diag}(\mathbf{v})] = \operatorname{diag}(S\mathbf{v}).$$

Exercise 4.1. *Check that in the independent case, the solution M to* (4.14) *is diagonal, $M = diag(\mathbf{m})$, where $\mathbf{m}$ solves the vector Dyson equation* (4.10). *Verify that the statements of the local laws formulated in the general correlated language reduce to those for the Wigner type problem. Check that the stability operator $(I - \mathcal{C}_M\mathcal{S})^{-1}$ restricted to diagonal matrices is equivalent to the stability operator* (4.13).

The following table summarizes the four classes of ensembles we discussed.

Name	Dyson Equation	For	Stability op	Feature
Wigner $\mathbb{E}\lvert h_{ij}\rvert^2 = s_{ij} = \frac{1}{N}$	$1+(z+m)m=0$	$m \approx \frac{1}{N}\operatorname{Tr} G$	$\frac{1}{1-m^2}$	Scalar Dyson equation, $m = m_{sc}$ is explicit
Generalized Wigner $\sum_j s_{ij} = 1$	$1+(z+S\mathbf{m})\mathbf{m}=0$	$m_i \approx \frac{1}{N}\operatorname{Tr} G$	$\frac{1}{1-m^2 S}$	Vector Dyson equation, Split S as $S^{\perp} + \lvert\mathbf{e}\rangle\langle\mathbf{e}\rvert$
Wigner-type s_{ij} arbitrary	$1+(z+S\mathbf{m})\mathbf{m}=0$	$m_i \approx G_{ii}$	$\frac{1}{1-\mathbf{m}^2 S}$	Vector Dyson equation, $\mathbf{m}$ to be determined
Correlated matrix $\mathbb{E}h_{xy}h_{uw} \not\sim \delta_{xw}\delta_{yu}$	$I+(z+\mathcal{S}[M])M=0$	$M_{ij} \approx G_{ij}$	$\frac{1}{1-M\mathcal{S}[\cdot]M}$	Matrix Dyson equation Super-operator

We remark that in principle the averaged law (density of states) for generalized Wigner ensemble could be studied via a scalar equation only since the answer is given by the scalar Dyson equation, but in practice a vector equation is studied in order to obtain entrywise and isotropic information. However, Wigner-type matrices need a vector Dyson equation even to identify the density of states. Correlated matrices need a full scale matrix equation since the answer M is typically a non-diagonal matrix.

4.5 The precise meaning of the approximations

In the previous sections we used the sloppy notation $\approx$ to indicate that the (random) resolvent G in various senses is close to a deterministic object. We now explain what we mean by that. Consider first (4.3), the entrywise statement for the Wigner case:

$$G_{ij}(z) \approx \delta_{ij} m_{sc}(z).$$

More precisely, we will see that

$$\big|G_{ij}(z) - \delta_{ij} m_{sc}(z)\big| \lesssim \frac{1}{\sqrt{N\eta}}, \qquad \eta = \operatorname{Im} z \tag{4.16}$$

holds. Here the somewhat sloppy notation $\lesssim$ indicates that the statement holds with very high probability and with an additional factor N^{ε}. The very precise form of (4.16) is the following: for any $\varepsilon, D > 0$ we have

$$\max_{ij} \mathbb{P}\Big(\big|G_{ij}(z) - \delta_{ij} m_{sc}(z)\big| \geqslant \frac{N^{\varepsilon}}{\sqrt{N\eta}}\Big) \leqslant \frac{C_{D,\varepsilon}}{N^D} \tag{4.17}$$

with some constant $C_{D,\varepsilon}$ independent of N, but depending on D, ε and the sequence μ_p bounding the moments in (4.2). We typically consider only spectral parameters with

$$|z| \leqslant C, \qquad \eta \geqslant N^{-1+\gamma} \tag{4.18}$$

for any fixed positive constants C and γ, and we encourage the reader to think of z satisfying these constraints, although our results are eventually valid for a larger set as well (the restriction $|z| \leqslant C$ can be replaced with $|z| \leqslant N^C$ and the lower bound on η is not necessary if $E = \operatorname{Re} z$ is away from the support of the density of states).

Notice that (4.17) is formulated for any fixed z, but the probability control is very strong, so one can extend the same bound to hold simultaneously for any z satisfying (4.18), i.e.

$$\mathbb{P}\Big(\exists z \in \mathbb{C}\ ;\ |z| \leqslant C, \operatorname{Im} z \geqslant N^{-1+\gamma} \text{ such that } \max_{ij} \big|G_{ij}(z) - \delta_{ij} m_{sc}(z)\big| \geqslant \frac{N^{\varepsilon}}{\sqrt{N\eta}}\Big) \leqslant \frac{C_{D,\varepsilon}}{N^D}. \tag{4.19}$$

Bringing the maximum over i, j inside the probability follows from a simple union bound. The same trick does not work directly for bringing the maximum over all z inside since there are uncountable many of them. But notice that the function

$$z \to G_{ij}(z) - \delta_{ij} m_{sc}(z)$$

is Lipschitz continuous with a Lipschitz constant C/η^2 which is bounded by CN^2 in the domain (4.18). Therefore, we can first choose a very dense, say N^{-3}-grid of z values, apply the union bound to them and then argue with Lipschitz continuity for all other z values.

Exercise 4.2. *Make this argument precise, i.e. show that* (4.19) *follows from* (4.17).

Similar argument does not quite work for the isotropic formulation. While (4.5) holds for any fixed (sequences of) ℓ^2-normalized vectors $\mathbf{x}$ and $\mathbf{y}$, i.e. in its precise formulation we have

$$\mathbb{P}\Big(\big| \langle \mathbf{x}, G(z)\mathbf{y}\rangle - m_{sc}(z)\langle \mathbf{x}, \mathbf{y}\rangle \big| \geqslant \frac{N^{\varepsilon}}{\sqrt{N\eta}} \Big) \leqslant \frac{C_{D,\varepsilon}}{N^D} \tag{4.20}$$

for any fixed $\mathbf{x}, \mathbf{y}$ with $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, we cannot bring the supremum over all $\mathbf{x}, \mathbf{y}$ inside the probability. Clearly $\max_{\mathbf{x},\mathbf{y}} \langle \mathbf{x}, G(z)\mathbf{y}\rangle$ would give the norm of G which is $1/\eta$.

Furthermore, a common feature of all our estimates is that the local law in averaged sense is one order more precise than the entrywise or isotropic laws, e.g. for the precise form of (4.4) we have

$$\mathbb{P}\Big(\big| \frac{1}{N} \operatorname{Tr} G(z) - m_{sc}(z) \big| \geqslant \frac{N^{\varepsilon}}{N\eta} \Big) \leqslant \frac{C_{D,\varepsilon}}{N^D}. \tag{4.21}$$

5 Physical motivations

The primary motivation to study local spectral statistics of large random matrices comes from nuclear and condensed matter physics. The random matrix models a quantum Hamiltonian and its eigenvalues correspond to the energy levels.

Other applications concern statistics (especially largest eigenvalues of sample covariance matrices of the form XX^* where X has independent entries), wireless communication and neural networks, but here we focus only on physical motivations.

5.1 Basics of quantum mechanics

We start with summarizing the basic setup of quantum mechanics. A quantum system is described by a *configuration space* Σ, e.g. $\Sigma = \{\uparrow, \downarrow\}$ for a single spin, or $\Sigma = \mathbb{Z}^3$ for an electron hopping on an ionic lattice or $\Sigma = \mathbb{R}^3$ for an electron in vacuum. Its elements $x \in \Sigma$ are called configurations and it is equipped with a natural measure (e.g. the counting measure for discrete Σ or the Lebesgue measure for $\Sigma = \mathbb{R}^3$). The *state space* is a complex Hilbert space, typically the natural L^2-space of Σ, i.e. $\ell^2(\Sigma) = \mathbb{C}^2$ in case of a single spin or $\ell^2(\mathbb{Z}^3)$ for an electron in a lattice. Its elements are called wave functions, these are normalized functions $\psi \in \ell^2(\Sigma)$, with $\|\psi\|_2 = 1$. The quantum wave function entirely describes the quantum state. In fact its overall phase does not carry measurable physical information; wave functions ψ and $e^{ic}\psi$ are indistinguishable for any real constant c. This is because only quadratic forms of ψ are measurable, i.e. only quantities of the form $\langle \psi, O\psi \rangle$ where O is a self-adjoint operator. The probability density $|\psi(x)|^2$ on the configuration space describes the probability to find the quantum particle at configuration x.

The dynamics of the quantum system, i.e. the process how ψ changes in time, is described by the *Hamilton operator*, which is a self-adjoint operator acting on the state space $\ell^2(\Sigma)$. If Σ is finite, then it is an $\Sigma \times \Sigma$ hermitian matrix. The matrix elements $H_{xx'}$ describe the quantum transition rates from configuration x to x'. The dynamics of ψ is described by the Schrödinger equation

$$i\partial_t \psi_t = H\psi_t$$

with some given initial condition $\psi_{t=0} := \psi_0$. The solution is given by $\psi_t = e^{-itH}\psi_0$. This simple formula is however, quite hard to compute or analyze, especially for large times. Typically one writes up the spectral decomposition of H as $H = \sum_n \lambda_n |\mathbf{v}_n\rangle\langle\mathbf{v}_n|$, where λ_n and $\mathbf{v}_n$ are the eigenvalues and eigenvectors of H, i.e. $H\mathbf{v}_n = \lambda_n\mathbf{v}_n$. Then

$$e^{-itH}\psi_0 = \sum_n e^{-it\lambda_n}\langle\mathbf{v}_n, \psi_0\rangle\mathbf{v}_n =: \sum_n e^{-it\lambda_n} c_n\mathbf{v}_n.$$

If ψ_0 coincides with one of the eigenvectors, $\psi_0 = \mathbf{v}_n$, then the sum above collapses and

$$\psi_t = e^{-itH}\psi_0 = e^{-it\lambda_n}\mathbf{v}_n.$$

Since the physics encoded in the wave function is insensitive to an overall phase, we see that eigenvectors remain unchanged along the quantum evolution.

Once ψ_0 is a genuine linear combination of several eigenvectors, quadratic forms of ψ_t become complicated:

$$\langle\psi_t, O\psi_t\rangle = \sum_{nm} e^{it(\lambda_m - \lambda_n)}\bar{c}_m c_n\langle\mathbf{v}_m, O\mathbf{v}_n, \rangle.$$

This double sum is highly oscillatory and subject to possible periodic and quasi-periodic behavior depending on the commensurability of the eigenvalue differences $\lambda_m - \lambda_n$. Thus the statistics of the eigenvalues carry important physical information on the quantum evolution.

The Hamiltonian H itself can be considered as an observable, and the quadratic form $\langle\psi, H\psi\rangle$ describes the *energy* of the system in the state ψ. Clearly the energy is a conserved quantity

$$\langle\psi_t, H\psi_t\rangle = \langle e^{-itH}\psi_0, He^{-itH}\psi_t\rangle = \langle\psi_0, H\psi_t\rangle.$$

The eigenvalues of H are called **energy levels** of the system.

Disordered quantum systems are described by random Hamiltonians, here the randomness comes from an external source and is often described phenomenologically. For example, it can represent impurities in the state space (e.g. the ionic lattice is not perfect) that we do not wish to (or cannot) describe with a deterministic precision, only their statistical properties are known.

5.2 The "grand" universality conjecture for disordered quantum systems

The general belief is that disordered quantum systems with "sufficient" complexity are subject to a strong dichotomy. They exhibit one of the following two behaviors: they are either in the *insulating* or in the *conducting* phase. These two phases are also called **localization** and **delocalization** regime. The behavior may depend on the energy range: the same quantum system can be simultaneously in both phases but at different energies.

The **insulator (or localized regime)** is characterized by the following properties:

1) Eigenvectors are **spatially localized**, i.e. the overwhelming mass of the probability density $|\psi(x)|^2 \mathrm{d}x$ is supported in a small subset of Σ. More precisely, there exists an $\Sigma' \subset \Sigma$, with $|\Sigma'| \ll |\Sigma|$ such that

$$\int_{\Sigma\setminus\Sigma'} |\psi(x)|^2 \mathrm{d}x \ll 1$$

2) **Lack of transport**: if the state ψ_0 is initially localized, then it remains so (maybe on a larger domain) for all times. Transport is usually measured with the mean square displacement if Σ has a metric. For example, for $\Sigma = \mathbb{Z}^d$ we consider

$$\langle x^2\rangle_t := \sum_{x\in\mathbb{Z}^3} x^2|\psi_t(x)|^2, \tag{5.1}$$

then localization means that

$$\sup_{t\geqslant 0}\langle x^2\rangle_t \leqslant C$$

assuming that at time $t = 0$ we had $\langle x^2\rangle_{t=0} < \infty$. Strictly speaking this concept makes sense only if Σ is infinite, but one can require that the constant C does not depend on some relevant size parameter of the model.

3) Green functions have a **finite localization length** ℓ, i.e. the off diagonal matrix elements of the resolvent decays exponentially (again for $\Sigma = \mathbb{Z}^d$ for simplicity)

$$|G_{xx'}| \leqslant Ce^{-|x-x'|/\ell}.$$

4) **Poisson local eigenvalue statistics**: Nearby eigenvalues are statistically independent, i.e. they approximately form a Poisson point process after appropriate rescaling.

The **conducting (or delocalized) regime** is characterized by the opposite features:

1) Eigenvectors are **spatially delocalized**, i.e. the mass of the probability density $|\psi(x)|^2$ is not concentrated on a much smaller subset of Σ.

2) **Transport via diffusion**: The mean square displacement (5.1) grows diffusively, e.g. for $\Sigma = \mathbb{Z}^d$

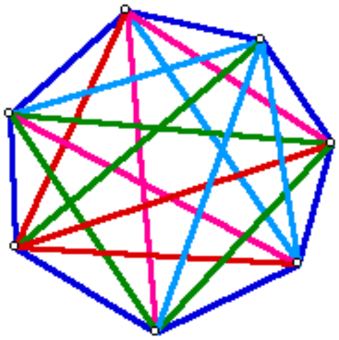
$$\langle x^2 \rangle_t \approx Dt$$

with some nonzero constant D (diffusion constant) for large times. If Σ is a finite part of $\mathbb{Z}^d$, e.g. $\Sigma = [1, L]^d \cap \mathbb{Z}^d$, then this relation should be modified so that the growth of $\langle x^2 \rangle_t$ with time can last only until the whole Σ is exhausted.

3) The Green function does not decay exponentially, the localization length $\ell = \infty$.

4) **Random matrix local eigenvalue statistics**: Nearby eigenvalues are statistically strongly dependent, in particular there is a level repulsion. They approximately form a GUE or GOE eigenvalue point process after appropriate rescaling. The symmetry type of the approximation is the same as the symmetry type of the original model (time reversal symmetry gives GUE).

The most prominent simple example for the conducting regime is the Wigner matrices or more generally Wigner-type matrices. They represent a quantum system where hopping from any site $x \in \Sigma$ to any other site $x' \in \Sigma$ is statistically equally likely (Wigner ensemble) or at least comparably likely (Wigner type ensemble), so it can be represented by a complete graph. The edges correspond to the matrix elements $h_{xx'}$ and they are independent. For Wigner matrices there is no specific spatial structure present, the system is completely homogeneous. Wigner type ensembles model a system with an inhomogeneous spatial structure, but it is still a **mean field** model since most transition rates are comparable. However, some results on Wigner type matrices allow zeros in the matrix of variances S, i.e. certain jumps are explicitly forbidden. The picture (5.2) schematically indicates the configuration space of $N = |\Sigma| = 7$ states with random quantum transition rates.



$$H = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1N} \\ h_{21} & h_{22} & \dots & h_{2N} \\ \vdots & \vdots & & \vdots \\ h_{N1} & h_{N2} & \dots & h_{NN} \end{pmatrix}. \tag{5.2}$$

The delocalization of the eigenvectors (item 1) was presented in (2.2), while item 4) is the WDM universality. The diffusive feature (item 2) is trivial since due to the mean field character, the maximal displacement is already achieved after $t \sim O(1)$. Thus the **Wigner matrix is in the delocalized regime.**

It is not so easy to present a non-trivial example for the insulator regime. A trivial example is if H is a diagonal matrix in the basis given by Σ, with i.i.d. entries in the diagonal, then items 1)–4) of the insulator regime clearly hold. Beyond the diagonal, even a short range hopping can become delocalized, for example the lattice Laplacian on $\mathbb{Z}^d$ has delocalized eigenvectors (plane waves). However, if the Laplacian is perturbed by a random diagonal, then localization may occur – this is the celebrated *Anderson metal-insulator transition* [11], which we discuss next.

5.3 Anderson model

The prototype of the random Schrödinger operators is the Anderson model on the d-dimensional square lattice $\mathbb{Z}^d$. It consists of a Laplacian (hopping term to the neighbors) and a random potential:

$$H = \Delta + \lambda V \tag{5.3}$$

acting on $\ell^2(\mathbb{Z}^d)$. The matrix elements of the Laplacian are given by

$$\Delta_{xy} = \mathbf{1}(|x-y| = 1)$$

and the potential is diagonal, i.e.

$$V_{xy} = \delta_{xy} v_x,$$

where $\{v_x \;:\; i \in \mathbb{Z}^d\}$ is a collection of real i.i.d. random variables sitting on the lattice sites. For definiteness we assume that

$$\mathbb{E} v_x = 0, \qquad \mathbb{E} v_x^2 = 1$$

and λ is a coupling parameter. Notice that Δ is self-adjoint and bounded, while the potential at every site is bounded almost surely. For simplicity we may assume that the common distribution of v has bounded support, i.e. V, hence H are bounded operators. This eliminates some technical complications related to the proper definition of the self-adjoint extensions.

5.3.1 The free Laplacian

For $\lambda = 0$, the spectrum is well known, the eigenvector equation $\Delta f = \mu f$, i.e.

$$\sum_{|y-x|=1} f_y = \mu f_x, \qquad \forall x \in \mathbb{Z}^d,$$

has plane waves as eigenfunctions parametrized by the d-torus, $k = (k_1, k_2, \ldots, k_d) \in [-\pi, \pi]^d$:

$$f_x = e^{ik\cdot x}, \qquad \mu = 2\sum_{i=1}^{d} \cos k_i.$$

Although these plane waves are not ℓ^2-normalizable, they still form a complete system of generalized eigenvectors for the bounded self-adjoint operator Δ. The spectrum is the interval $[-2d, 2d]$ and it is a purely absolutely continuous spectrum (we will not need its precise definition if you are unfamiliar with it). Readers uncomfortable with unbounded domains can take a large torus $[-L, L]^d$, $L \in \mathbb{N}$, instead of $\mathbb{Z}^d$ as the configuration space. Then everything is finite dimensional, and the wave-numbers k are restricted to a finite lattice within the torus $[-\pi, \pi]^d$. Notice that the eigenvectors are still plane waves, in particular they are completely delocalized. One may also study the time evolution $e^{it\Delta}$ (basically by Fourier transform) and one finds **ballistic behavior**, i.e. for the mean square displacement (5.1) one finds

$$\langle x^2 \rangle_t = \sum_{x\in\mathbb{Z}^d} x^2 |\psi_t(x)|^2 \sim Ct^2, \qquad \psi_t = e^{it\Delta}\psi_0$$

for large t. Thus for $\lambda = 0$ the system in many aspects is in the delocalized regime. Since randomness is completely lacking, it is not expected that other features of the delocalized regime hold, e.g. the local spectral statistics is not the one from random matrices – it is rather related to a lattice point counting problem. Furthermore, the eigenvalues have degeneracies, i.e. level repulsion, a main characteristics for random matrices, does not hold.

5.3.2 Turning on the randomness

Now we turn on the randomness by taking some $\lambda \neq 0$. This changes the behavior of the system drastically in certain regimes. More precisely:

- In $d = 1$ **dimension** the system is in the localized regime as soon as $\lambda \neq 0$, see [50]
- In $d = 2$ **dimensions** On physical grounds it is conjectured that the system is localized for any $\lambda \neq 0$ [74]. No mathematical proof exists.
- In the most important physical $d = 3$ **dimensions** we expect a phase transition: The system is localized for large disorder, $|\lambda| \geqslant \lambda_0(d)$ or at the spectral edges [3, 47]. For small disorder and away from the spectral edges delocalization is expected but there is no rigorous proof. This is the celebrated **extended states or delocalization conjecture**, one of the few central holy grails of mathematical physics.

Comparing random Schrödinger with random matrices, we may write up the matrix of the $d = 1$ dimensional operator H (5.3) in the basis given by $\Sigma = [\![1, L]\!]$:

$$H = \Delta + \sum_{x=1}^{L} v_x = \begin{pmatrix} v_1 & 1 & & & & \\ 1 & v_2 & 1 & & & \\ & 1 & \ddots & & & \\ & & & \ddots & 1 & \\ & & & 1 & v_{L-1} & 1 \\ & & & & 1 & v_L \end{pmatrix}.$$

It is tridiagonal matrix with i.i.d. random variables in the diagonal and all ones in the minor diagonal. It is a **short range model** as immediate quantum transitions (jumps) are allowed only to the nearest neighbors. Structurally this H is very different from the typical Wigner matrix (5.2) where all matrix elements are roughly comparable (**mean field model**).

5.4 Random band matrices

Random band matrices naturally interpolate between the mean field Wigner ensemble and the short range random Schrödinger operators. Let the state space be

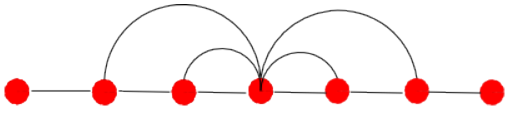
$$\Sigma := [1, L]^d \cap \mathbb{Z}^d$$

a lattice box of linear size L in d dimensions. The total dimension of the state space is $N = |\Sigma| = L^d$. The entries of $H = H^*$ are centered, independent but not identically distributed – it is like the Wigner type ensemble, but without the mean field condition $s_{xy} = \mathbb{E}|h_{xy}|^2 \leqslant C/N$. Instead, we introduce a new parameter, $1 \leqslant W \leqslant L$ the **bandwidth** or the interaction range. We assume that the variances behave as

$$\mathbb{E}|h_{xy}|^2 = \frac{1}{W^d} f\Big(\frac{|x-y|}{W}\Big).$$

In $d = 1$ physical dimension the corresponding matrix is an $L \times L$ matrix with a nonzero band of width W around the diagonal. From any site a direct hopping of size W is possible, see the figure below with $L = 7$, $W = 4$:

$$H = \begin{pmatrix} * & * & * & 0 & 0 & 0 & 0 \\ * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 \\ 0 & * & * & * & * & * & 0 \\ 0 & 0 & * & * & * & * & * \\ 0 & 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * \end{pmatrix}$$



Clearly $W = L$ corresponds to the Wigner ensemble, while $W = 1$ is very similar to the random Schrödinger with its short range hopping. The former is delocalized, the latter is localized, hence there is a transition with can be probed by changing W from 1 to L. The following table summarizes "facts" from physics literature on the transition threshold:

Anderson metal-insulator transition occurs at the following thresholds:

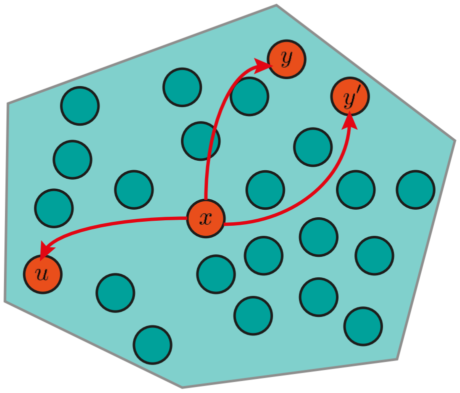
$W \sim L^{1/2}$	$(d = 1)$	Supersymmetry [48]
$W \sim \sqrt{\log L}$	$(d = 2)$	Renormalization group scaling [1]
$W \sim W_0(d)$	$(d \geqslant 3)$	extended states conjecture [11]

All these conjectures are mathematically open, the most progress has been done in $d = 1$. It is known that we have localization in the regime $W \ll L^{1/8}$ [66] and delocalization for $W \gg L^{4/5}$ [41]. The two point correlation function of the characteristic polynomial was shown to be given by the Dyson sine kernel up to the threshold $W \gg L^{1/2}$ in [68].

In these lectures we restrict our attention to mean field models, i.e. band matrices will not be discussed. We nevertheless mentioned them because they are expected to be easier than the short range random Schrödinger operators and they still exhibit the Anderson transition in a highly nontrivial way.

5.5 Mean field quantum Hamiltonian with correlation

Finally we explain how correlated random matrices with a certain correlation decay are motivated. We again equip the state space Σ with a metric to be able to talk about "nearby" states. It is then reasonable to assume that h_{xy} and $h_{xy'}$ are correlated if y and y' are close with a decaying correlation as $\mathrm{dist}(y, y')$ increases. For example, in the figure h_{xy} and $h_{xy'}$ are strongly correlated but h_{xy} and h_{xu} are not (or only very weakly) correlated. We can combine this feature with an inhomogeneous spatial structure as in the Wigner-type ensembles.



6 Results

Here we list a few representative results with precise conditions. The results can be divided roughly into three categories:

- Properties of the solution of the Dyson equation, especially the singularity structure of the density of states and the boundedness of the stability operator. This part of the analysis is deterministic.

- Local laws, i.e. approximation of the (random) resolvent G by the solution of the corresponding Dyson equation with very high probability down to the optimal scale $\eta \gg 1/N$.

- Bulk universality of the local eigenvalue statistics on scale $1/N$.

6.1 Properties of the solution to the Dyson equations

6.1.1 Vector Dyson equation

First we focus on the vector Dyson equation (4.10) with a general symmetric variance matrix S motivated by Wigner type matrices:

$$-\frac{1}{\mathbf{m}} = z + S\mathbf{m}, \qquad \mathbf{m} \in \mathbb{H}^N, \quad z \in \mathbb{H} \tag{6.1}$$

(recall that the inverse of a vector $1/\mathbf{m}$ is understood component wise, i.e. $1/\mathbf{m}$ is an N vector with components $(1/\mathbf{m})_i = 1/m_i$). We may add an external source which is real vector $\mathbf{a} \in \mathbb{R}^N$ and the equation is modified to

$$-\frac{1}{\mathbf{m}} = z + \mathbf{a} + S\mathbf{m}, \qquad \mathbf{m} \in \mathbb{H}^N, \quad z \in \mathbb{H}, \tag{6.2}$$

but we will consider the $\mathbf{a} = 0$ case for simplicity. We equip the space $\mathbb{C}^N$ with the maximum norm,

$$\|\mathbf{m}\| := \max_i |m_i|,$$

and we let $\|S\|$ be the matrix norm induced by the maximum norm of vectors. We start with the existence and uniqueness result for (6.1), see e.g. Proposition 2.1 in [4]:

Theorem 6.1. *The equation* (6.1) *has a unique solution* $\mathbf{m} = \mathbf{m}(z)$ *for any* $z \in \mathbb{H}$. *For each* $i \in [\![1, N]\!]$ *there is a probability measure* $\nu_i(\mathrm{d}x)$ *on* $\mathbb{R}$ *(called* **generating measure***) such that* m_i *is the Stieltjes transform of* v_i*:*

$$m_i(z) = \int_{\mathbb{R}} \frac{\nu_i(\mathrm{d}\tau)}{\tau - z}, \tag{6.3}$$

and the support of all ν_i *lie in the interval* $[-2\|S\|^{1/2}, 2\|S\|^{1/2}]$. *In particular we have the trivial upper bound*

$$\|\mathbf{m}(z)\| \leqslant \frac{1}{\eta}, \qquad \eta = \operatorname{Im} z. \tag{6.4}$$

Recalling that the **self-consistent density of states** was defined in (4.11) via the inverse Stieltjes transform of $\langle \mathbf{m} \rangle = \frac{1}{N}\sum m_i$, we see that

$$\varrho = \langle \boldsymbol{\nu} \rangle = \frac{1}{N}\sum_i \nu_i.$$

We now list two assumptions on S, although for some results we will need only one of them:

- **Boundedness:** We assume that there exists two positive constants c, C such that

$$\frac{c}{N} \leqslant s_{ij} \leqslant \frac{C}{N} \tag{6.5}$$

- **Hölder regularity:**

$$|s_{ij} - s_{i'j'}| \leqslant C\Big[\frac{|i-i'| + |j-j'|}{N}\Big]^{1/2} \tag{6.6}$$

We remark that the lower bound in (6.5) can be substantially weakened, in particular large zero blocks are allowed. For example, we may assume only that S has a substantial diagonal, i.e. $s_{ij} \geqslant c \cdot \mathbf{1}(|i-j| \leqslant \varepsilon N)$ with some fixed positive c, ε, but for simplicity of the presentation we follow (6.5).

The Hölder regularity (6.6) expresses a regularity on the order N scale in the matrix. It can be understood in the easiest way if we imagine that the matrix elements s_{ij} come from a macroscopic profile function $S(x, y)$ on $[0,1] \times [0,1]$ as

$$s_{ij} = \frac{1}{N} S\big(\frac{i}{N}, \frac{j}{N}\big) \tag{6.7}$$

It is easy to check that if $S : [0,1]^2 \to \mathbb{R}_+$ is Hölder continuous with a Hölder exponent $1/2$, then (6.6) holds. In fact, the Hölder regularity condition can also be weakened to **piecewise 1/2-Hölder regularity** (with finitely many pieces), in that case we assume that s_{ij} is of the form (6.7) with a profile function $S(x,y)$ that is piecewise Hölder continuous with exponent 1/2, i.e. there exists a fixed (N-independent) partition $I_1 \cup I_2 \cup \ldots \cup I_n = [0,1]$ of the unit interval into smaller intervals such that

$$\max_{ab} \sup_{x,x' \in I_a} \sup_{y,y' \in I_b} \frac{|S(x,y) - S(x',y')|}{|x-x'|^{1/2} + |y-y'|^{1/2}} \leqslant C. \tag{6.8}$$

The main theorems summarizing the properties of the solution to (6.1) are the following. The first theorem assumes only (6.5) and it is relevant in the bulk. We will prove it later in Section 6.2.

Theorem 6.2. *Suppose that S satisfies* (6.5)*. Then we have the following bounds:*

$$|\mathbf{m}(z)| \lesssim \frac{1}{\varrho(z) + dist(z, supp \varrho)}, \qquad \varrho(z) \lesssim \operatorname{Im} \mathbf{m}(z) \lesssim (1+|z|^2)\|\mathbf{m}(z)\|^2 \varrho(z). \tag{6.9}$$

The second theorem additionally assumes (6.6), but the result is much more precise, in particular a complete analysis of singularities is possible.

Theorem 6.3. *[Theorem 2.6 in [4]] Suppose that S satisfies* (6.5) *and it is Hölder continuous* (6.6) *or piecewise Hölder continuous* (6.8)*. Then we have the following:*

(i) The generating measures have Lebesgue density, $\nu_i(\mathrm{d}\tau) = \nu_i(\tau)\mathrm{d}\tau$ and the **generating densities** *ν_i are uniformly 1/3-Hölder continuous, i.e.*

$$\max_i \sup_{\tau \neq \tau'} \frac{|\nu_i(\tau) - \nu_i(\tau')|}{|\tau - \tau'|^{1/3}} \leqslant C'. \tag{6.10}$$

(ii) The set on which ν_i is positive is independent of i:

$$\mathfrak{S} := \{\tau \in \mathbb{R} \, : \, \nu_i(\tau) > 0\}$$

and it is a union of finitely many open intervals. If S is Hölder continuous in the sense of (6.6)*, then $\mathfrak{S}$ consist of a single interval.*

(iii) The restriction of $\boldsymbol{\nu}(\tau)$ onto $\mathbb{R} \setminus \partial\mathfrak{S}$ is analytic in τ (as a vector-valued function).

(iv) At the (finitely many) points $\tau_0 \in \partial\mathfrak{S}$ the generating density has one of the following two behaviors:

CUSP: If τ_0 is at the intersection of the closure of two connected components of $\mathfrak{S}$, then $\boldsymbol{\nu}$ has a cubic root singularity, i.e.

$$\nu_i(\tau_0 + \omega) = c_i|\omega|^{1/3} + O(|\omega|^{2/3}) \tag{6.11}$$

with some positive constants c_i that are uniformly bounded from above and below, i.e. $c' \leqslant c_i \leqslant C'$ holds.

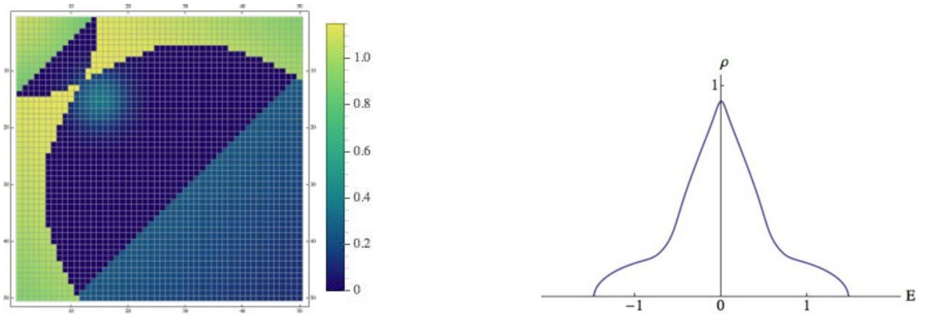
EDGE: If τ_0 is not a cusp, then it is the right or left endpoint of a connected component of $\mathfrak{S}$ and $\boldsymbol{\nu}$ has a square root singularity at τ_0:

$$\nu_i(\tau_0 \pm \omega) = c_i \omega^{1/2} + O(\omega), \qquad \omega \geqslant 0 \tag{6.12}$$

The constants c_i are uniformly bounded from above and below as in the cusp.

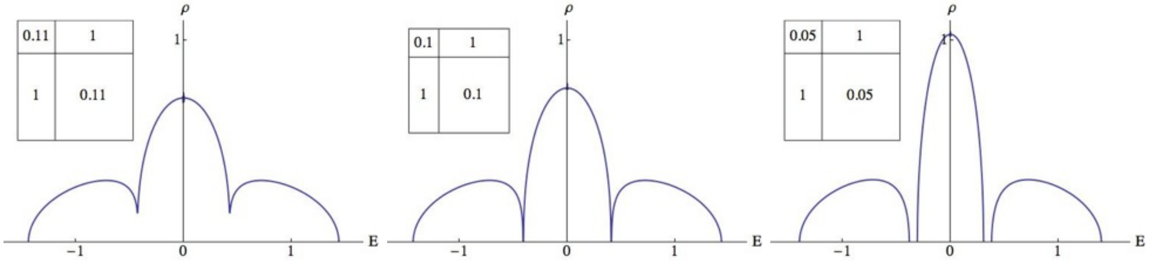
The positive constants c', C' in these statements depend only on the constants in the conditions (6.5) *and* (6.6)*, in particular they are independent of N.*

Some of these statements will be proved in Section 7. We now illustrate this theorem by a few pictures. The first picture indicates a nontrivial S-profile (different shades indicate different values in the matrix) and the corresponding self-consistent density of states.

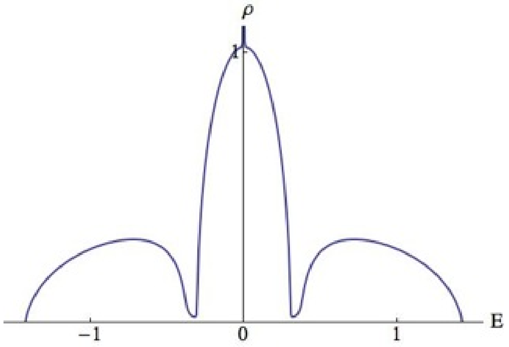


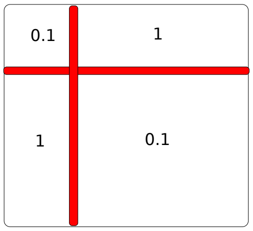
$\sum_j s_{ij} \neq \text{const}$ $\Longrightarrow$ Self-consistent density of states

The next picture shows how the support of the self-consistent density of states splits via cusps as the value of s_{ij} slowly changes. The matrices in the pictures represent the variance matrix S, or rather its profile function $S(x, y)$. Notice that the profile function $S(x, y)$ is only piecewise Hölder.

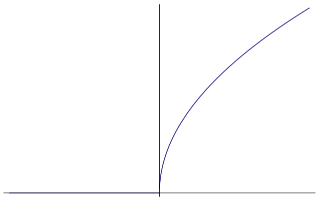


Cusps and splitting of the support are possible only if there is a discontinuity in the profile. If the above profile is smoothed out (indicated by the narrow shaded region in the picture of the matrix below), then the support becomes a single interval, with a specific smoothed out "almost cusp".

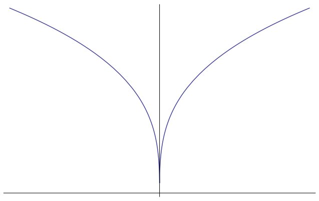




Finally we show the universal shape of the singularities and near singularities in the self-consistent density of states. The first two pictures are the edges and cusps, below them the approximate form of the density near the singularity in terms of the parameter $\omega = \tau - \tau_0$, compare with (6.12) and (6.11):

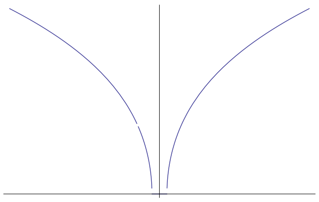


Edge, $\sqrt{\omega}$ singularity



Cusp, $|\omega|^{1/3}$ singularity

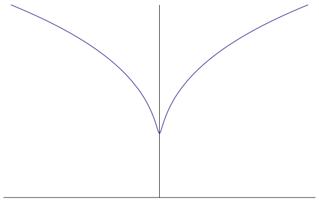
The next two pictures show the asymptotic form of the density right before and after the cusp formation. The relevant parameter t is an appropriate rescaling of ω; the size of the gap (after the cusp formation) and the minimum value of the density (before the cusp formation) set the relevant length scales on which the universal shape emerges:



Small-gap

$$\frac{(2+t)t}{1+(1+t+\sqrt{(2+t)t})^{2/3}+(1+t-\sqrt{(2+t)t})^{2/3}}$$

$t := \frac{|\omega|}{\text{gap}}$,



Smoothed cusp

$$\frac{\sqrt{1+t^2}}{(\sqrt{1+t^2}+t)^{2/3}+(\sqrt{1+t^2}-t)^{2/3}-1} - 1$$

$t := \frac{|\omega|}{(\text{minimum of } \varrho\ )^{1/3}}$

We formulated the vector Dyson equation in a discrete setup for N unknowns. We remark that it can be considered in a more abstract setup as follows. For a measurable space $\mathfrak{A}$ and a subset $\mathbb{D} \subseteq \mathbb{C}$ of the complex numbers, we denote by $\mathcal{B}(\mathfrak{A}, \mathbb{D})$ the space of bounded measurable functions on $\mathfrak{A}$ with values in $\mathbb{D}$. Let $(\mathfrak{X}, \pi(\mathrm{d}x))$ be a measure space with bounded positive (non-zero) measure π. Suppose we are given a real valued $a \in \mathcal{B}(\mathfrak{X}, \mathbb{R})$ and a non-negative, symmetric, $s_{xy} = s_{yx}$, function $s \in \mathcal{B}(\mathfrak{X}^2, \mathbb{R}_0^+)$. Then we consider the *quadratic vector equation (QVE)*,

$$-\frac{1}{m(z)} = z + a + Sm(z)\,, \qquad z \in \mathbb{H}\,, \tag{6.13}$$

for a function $m : \mathbb{H} \to \mathcal{B}(\mathfrak{X}, \mathbb{H})$, $z \mapsto m(z)$, where $S : \mathcal{B}(\mathfrak{X}, \mathbb{C}) \to \mathcal{B}(\mathfrak{X}, \mathbb{C})$ is the integral operator with kernel s,

$$(Sw)_x := \int s_{xy} w_y \pi(\mathrm{d}y)\,, \qquad x \in \mathfrak{X}\,,\ w \in \mathcal{B}(\mathfrak{X}, \mathbb{C})\,.$$

We equip the space $\mathcal{B}(\mathfrak{X}, \mathbb{C})$ with its natural norm,

$$\|w\| := \sup_{x \in \mathfrak{X}} |w_x|\,, \qquad w \in \mathcal{B}(\mathfrak{X}, \mathbb{C})\,.$$

With this norm $\mathcal{B}(\mathfrak{X}, \mathbb{C})$ is a Banach space. All results formulated in Theorem 6.3 are valid in this more general setup, for details, see [4]. The special case we discussed above corresponds to

$$\mathfrak{X} := \Big\{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\Big\}, \qquad \pi(\mathrm{d}x) = \frac{1}{N}\sum_{i=1}^{N} \delta\big(\frac{i}{N} - x\big).$$

The scaling here differs from (6.7) by a factor of N, since now $s_{xy} = S\big(x, y\big)$, $x, y \in \mathfrak{X}$ in which case there is an infinite dimensional limiting equation with $\mathfrak{X} = [0, 1]$ and $\pi(\mathrm{d}x)$ being the Lebesgue measure. If s_{ij} comes from a continuous profile, (6.7), then in the $N \to \infty$ limit, the vector Dyson equation becomes

$$-\frac{1}{m_x(z)} = z + \int_0^1 S(x, y) m_y(z)\mathrm{d}y, \qquad x \in [0, 1], \quad z \in \mathbb{H}.$$

6.1.2 Matrix Dyson equation

The matrix version of the Dyson equation naturally arises in the study of correlated random matrices, see Section 3.1.2 and Section 4.4. It takes the form

$$I + (z + \mathcal{S}[M])M = 0, \qquad \operatorname{Im} M > 0, \quad \operatorname{Im} z > 0, \qquad (MDE) \tag{6.14}$$

where we assume that $\mathcal{S} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is a linear operator that is

1) symmetric with respect to the Hilbert-Schmidt scalar product, i.e. $\operatorname{Tr} R^*\mathcal{S}[T] = \operatorname{Tr} \mathcal{S}[R]^*T$ for any matrices $R, T \in \mathbb{C}^{N\times N}$;

2) positivity preserving, i.e. $\mathcal{S}[R] \geqslant 0$ for any $R \geqslant 0$.

Somewhat informally we will refer to linear maps on the space of matrices as **superoperators** to distinguish them from usual matrices.

Originally, $\mathcal{S}$ is defined in (3.18) as a covariance operator of a hermitian random matrix H, but it turns out that (6.14) can be fully analyzed solely under these two conditions 1)–2). It is straightforward to check that $\mathcal{S}$ defined in (3.18) satisfies the conditions 1) and 2). Similarly to the vector Dyson equation (6.2) one may add an external source $A = A^* \in \mathbb{C}^{N\times N}$ and consider

$$I + (z + A + \mathcal{S}[M])M = 0, \qquad \operatorname{Im} M > 0 \tag{6.15}$$

but these notes will be restricted to $A = 0$.

Similarly to Theorem 6.1, we have an existence and uniqueness result for the solution (see [53]) moreover, we have a Stieltjes transform representation (Proposition 2.1 [7]):

Theorem 6.4. *For any $z \in \mathbb{H}$, the MDE (6.14) with the side condition $\operatorname{Im} M > 0$ has a unique solution $M = M(z)$ that is analytic in the upper half plane. The solution admits a Stieltjes transform representation*

$$M(z) = \int_{\mathbb{R}} \frac{V(\mathrm{d}\tau)}{\tau - z} \tag{6.16}$$

where $V(\mathrm{d}\tau)$ is a positive semidefinite matrix valued measure on $\mathbb{R}$ with normalization $V(\mathbb{R}) = I$. In particular

$$\|M(z)\| \leqslant \frac{1}{\operatorname{Im} z}. \tag{6.17}$$

The support of this measure lies in $[-2\|\mathcal{S}\|^{1/2}, 2\|\mathcal{S}\|^{1/2}]$, where $\|\mathcal{S}\|$ is the norm induced by the usual operator norm on $\mathbb{C}^{N\times N}$.

The solution M is called the **self-consistent Green function** or **self-consistent resolvent** since it will be used as a computable deterministic approximation to the random Green function G.

From now on we assume the following *flatness* condition on $\mathcal{S}$ that is the matrix analogue of the boundedness condition (6.5):

Flatness condition: The operator $\mathcal{S}$ is called **flat** if there exists two positive constants, c, C, independent of N, such that

$$c\langle R\rangle \leqslant \mathcal{S}[R] \leqslant C\langle R\rangle, \qquad \text{where} \quad \langle R\rangle := \frac{1}{N}\operatorname{Tr} R \tag{6.18}$$

holds for any positive matrix $R \geqslant 0$.

Under this condition we have the following quantitative results on the solution M (Proposition 2.2 and Proposition 4.2 of [7]):

Theorem 6.5. *Assume that $\mathcal{S}$ is flat, then the holomorphic function $\langle M\rangle : \mathbb{H} \to \mathbb{H}$ is the Stieltjes transform of a Hölder continuous probability density ϱ w.r.t. the Lebesgue measure:*

$$\langle V(\mathrm{d}\tau)\rangle = \varrho(\tau)\mathrm{d}\tau$$

i.e.

$$|\varrho(\tau_1) - \varrho(\tau_2)| \leqslant C|\tau_1 - \tau_2|^\delta \tag{6.19}$$

with some Hölder regularity exponent δ, independent of N ($\delta = 1/100$ would do). The density ϱ is called the **self-consistent density of states.** *Furthermore, ϱ is real analytic on the open set $\mathfrak{S} := \{\tau \in \mathbb{R} \; ; \; \varrho(\tau) > 0\}$ which is called the* **self-consistent bulk spectrum**. *For the solution itself we also have*

$$\|M(z)\| \leqslant \frac{C}{\varrho(z) + dist(z, \mathfrak{S})} \tag{6.20}$$

and

$$c\varrho(z) \leqslant \operatorname{Im} M(z) \leqslant C\|M(z)\|^2\varrho(z),$$

where $\varrho(z)$ is the harmonic extension of $\varrho(\tau)$ to the upper half plane. In particular, in the bulk regime of spectral parameters, where $\varrho(\operatorname{Re} z) \geqslant \delta$ for some fixed $\delta > 0$, we see that $M(z)$ is bounded and $\operatorname{Im} M(z)$ is comparable (as a positive definite matrix) with $\varrho(z)$.

Notice that unlike in the analogous Theorem 6.3 for the vector Dyson equation, here we do not assume any regularity on $\mathcal{S}$, but the conclusion is weaker. We do not get Hölder exponent 1/3 for the self-consistent density of states ϱ. Furthermore, cusp and edge analysis would also require further conditions on $\mathcal{S}$. Since in the correlated case we focus on the bulk spectrum, i.e. on spectral parameters z with $\operatorname{Re} z \in \mathfrak{S}$, we will not need detailed information about the density near the spectral edges.

6.2 Local laws for Wigner-type and correlated random matrices

We now state the precise form of the local laws.

Theorem 6.6 (Bulk local law for Wigner type matrices, Corollary 1.8 from [6])**.** *Let H be a centered Wigner type matrix with bounded variances $s_{ij} = \mathbb{E}|h_{ij}|^2$ i.e.* (6.5) *holds. Let $\mathbf{m}(z)$ be the solution to the vector Dyson equation* (6.2)*. We assume the uniform moment condition* (4.2) *for the matrix elements. Then the local law in the bulk holds, i.e. we fix positive constants $\delta, \gamma, \varepsilon$ and D, then for any spectral parameter $z = \tau + i\eta$ with*

$$\varrho(\tau) \geqslant \delta, \qquad \eta \geqslant N^{-1+\gamma} \tag{6.21}$$

we have the entrywise local law

$$\max_{ij} \mathbb{P}\Big(\big|G_{ij}(z) - \delta_{ij}m_i(z)\big| \geqslant \frac{N^\varepsilon}{\sqrt{N\eta}}\Big) \leqslant \frac{C}{N^D} \tag{6.22}$$

and, more generally, the isotropic law, i.e. for any non-random normalized vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$

$$\max_{ij} \mathbb{P}\Big(\big|\langle \mathbf{x}, G(z)\mathbf{y}\rangle - \langle \mathbf{x}, \mathbf{m}(z)\mathbf{y}\rangle\big| \geqslant \frac{N^\varepsilon}{\sqrt{N\eta}}\Big) \leqslant \frac{C}{N^D}. \tag{6.23}$$

Moreover for any non-random vector $\mathbf{w} = (w_1, w_2, \ldots) \in \mathbb{C}^N$ with $\max_i |w_i| \leqslant 1$ we have the averaged local law

$$\mathbb{P}\Big(\big|\frac{1}{N}\sum_i w_i\big[G_{ii}(z) - m_i(z)\big]\big| \geqslant \frac{N^\varepsilon}{N\eta}\Big) \leqslant \frac{C}{N^D}, \tag{6.24}$$

in particular (with $w_i = 1$) we have

$$\mathbb{P}\Big(\big|\frac{1}{N}\operatorname{Tr} G(z) - \langle \mathbf{m}(z)\rangle\big| \geqslant \frac{N^\varepsilon}{N\eta}\Big) \leqslant \frac{C}{N^D}. \tag{6.25}$$

The constant C in (6.22)*–*(6.25) *is independent of N and the choice of w_i, but it depends on $\delta, \gamma, \varepsilon, D$, the constants in* (6.5) *and the sequence μ_p bounding the moments in* (4.2).

As we explained around (4.19), in the entrywise local law (6.22) one may bring both superma on i, j and on the spectral parameter z inside the probability, i.e. one can guarantee that $G_{ij}(z)$ is close to $m_i(z)\delta_{ij}$ simultaneously for all indices and spectral parameters in the regime (6.21). Similarly, z can be brought inside the probability in (6.23) and (6.24), but the isotropic law (6.23) cannot hold simultaneously for all $\mathbf{x}, \mathbf{y}$ and similarly, the averaged law (6.24) cannot simultaneously hold for all $\mathbf{w}$.

We formulated the local law only under the boundedness condition (6.5) but only in the bulk of the spectrum for simplicity. Local laws near the edges and cusps require much more delicate analysis and some type of regularity on s_{ij}, e.g. the 1/2-Hölder regularity introduced in (6.6) would suffice. Much easier is the regime outside of the spectrum. The precise statement is found in Theorem 1.6 of [6].

For the correlated matrix we have the following local law from [7]:

Theorem 6.7 (Bulk local law for correlated matrices)**.** *Consider a random hermitian matrix $H \in \mathbb{C}^{N\times N}$ with correlated entries. Define the self-energy super operator $\mathcal{S}$ as*

$$\mathcal{S}[R] = \mathbb{E}\big[HRH\big] \tag{6.26}$$

acting on any matrix $R \in \mathbb{C}^{N\times N}$. Assume that the flatness condition (6.18) and the moment condition (4.2) hold. We also assume an exponential decay of correlations in the form

$$Cov\Big(\phi(W_A);\psi(W_B)\Big) \leqslant C(\phi,\psi)\ \ e^{-d(A,B)}. \tag{6.27}$$

Here $W = \sqrt{N}H$ is the rescaled random matrix, A, B are two subsets of the index set $[\![1,N]\!]\times[\![1,N]\!]$ and the distance d is the usual Euclidean distance between the sets $A \cup A^t$ and $B \cup B^t$, see figure below. Let M be the self-consistent Green function, i.e. the solution of the matrix Dyson equation (6.14) with $\mathcal{S}$ given in (6.26), and consider a spectral parameter in the bulk, i.e. $z = \tau + i\eta$ with

$$\varrho(\tau) \geqslant \delta, \qquad \eta \geqslant N^{-1+\gamma} \tag{6.28}$$

Then for any non-random normalized vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ we have the isotropic local law

$$\mathbb{P}\Big(\big|\langle \mathbf{x}, G(z)\mathbf{y}\rangle - \langle \mathbf{x}, M(z)\mathbf{y}\rangle\big| \geqslant \frac{N^\varepsilon}{\sqrt{N\eta}}\Big) \leqslant \frac{C}{N^D}, \tag{6.29}$$

in particular we have the entrywise law

$$\mathbb{P}\Big(\big|G_{ij}(z) - M_{ij}(z)\big| \geqslant \frac{N^\varepsilon}{\sqrt{N\eta}}\Big) \leqslant \frac{C}{N^D}, \tag{6.30}$$

for any i, j. Moreover for any fixed (deterministic) matrix T with $\|T\| \leqslant 1$, we have the averaged local law

$$\mathbb{P}\Big(\Big|\frac{1}{N}\operatorname{Tr} T\big[G(z) - M(z)\big]\Big| \geqslant \frac{N^\varepsilon}{N\eta}\Big) \leqslant \frac{C}{N^D}. \tag{6.31}$$

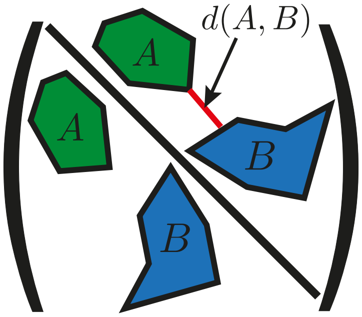
The constant C is independent of N and the choice of $\mathbf{x}, \mathbf{y}$, but it depends on $\delta, \gamma, \varepsilon, D$, the constants in (6.18) and the sequence μ_p bounding the moments in (4.2).

In our recent paper [42] we substantially relaxed the condition on the correlation decay (6.27) to the form

$$\operatorname{Cov}\Big(\phi(W_A);\psi(W_B)\Big) \leqslant \frac{C(\phi,\psi)}{1+d^2}e^{-d/N^{1/4}}, \qquad d = d(A,B),$$

and a similar condition on higher order cumulants, see [42] for the precise conditions.

In Theorem 6.7 we again formulated the result only in the bulk, but similar (even stronger) local law is available for energies τ that are separated away from the support of ϱ.



In these notes we will always assume that H is centered, $\mathbb{E}H = 0$ for simplicity, but our result holds in the general case as well. In that case $\mathcal{S}$ is given by

$$\mathcal{S}[R] = \mathbb{E}\big[(H - \mathbb{E}H)R(H - \mathbb{E}H)\big]$$

and M solves the MDE with external source $A := \mathbb{E}H$, see (6.15).

6.3 Bulk universality and other consequences of the local law

In this section we give precise theorems of three important consequences of the local law. We will formulate the results in the simplest case, in the bulk. We give some sketches of the proofs. Complete arguments that can be found in [6] and [7, 42].

6.3.1 Delocalization

The simplest consequence of the entrywise local law is the delocalization of the eigenvectors as explained in Section 2.2. The precise formulation goes as follows:

Theorem 6.8 (Delocalization of bulk eigenvectors)**.** *Let H be a Wigner type or, more generally, a correlated random matrix, satisfying the conditions of Theorem 6.6 or Theorem 6.7, respectively. Let ϱ be the self-consistent density of states obtained from solving the corresponding Dyson equation. Then for any $\delta, \gamma > 0$ and $D > 0$ we have*

$$\mathbb{P}\Big(\exists \mathbf{u}, \lambda,\ H\mathbf{u} = \lambda\mathbf{u},\ \ \|\mathbf{u}\|_2 = 1\ \ \varrho(\lambda) \geqslant \delta,\ \ \|\mathbf{u}\|_\infty \geqslant N^{-\frac{1}{2}+\gamma}\Big) \leqslant \frac{C}{N^D}.$$

Sketch of the proof. The proof was basically given in (2.2). The local laws guarantee that $\operatorname{Im} G_{jj}(z)$ is close to its deterministic approximant, $m_i(z)\delta_{ij}$ or $M_{ij}(z)$, these statements hold for any $E = \operatorname{Re} z$ in the bulk and for $\eta \geqslant N^{-1+\gamma}$. Moreover, (6.9) and (6.20) show that in the bulk regime both $|\mathbf{m}|$ and $\|M\|$ are bounded. From these two information we conclude that $\operatorname{Im} G_{jj}(z)$ is bounded with very high probability. □

6.3.2 Rigidity

The next standard consequence of the local law is the **rigidity of eigenvalues**. It states that with very high probability the eigenvalues in the bulk are at most $N^{-1+\gamma}$-distance away from their **classical locations** predicted by the corresponding quantiles of the self-consistent density of states, for any $\gamma > 0$. This error bar $N^{-1+\gamma}$ reflects that typically the eigenvalues are almost as close to their deterministically prescribed locations as the typical level spacing N^{-1}. This is actually an indication of a very strong correlation; e.g. if the eigenvalues were completely uncorrelated, i.e. given by a Poisson point process with intensity N, then the typical fluctuation of the location of the points would be $N^{-1/2}$.

Since local laws at spectral parameter $z = E + i\eta$ determine the local eigenvalue density on scale η, it is very natural that a local law on scale $\eta = \operatorname{Im} z$ locates individual eigenvalues with η-precision. Near the edges and cusps the local spacing is different ($N^{-2/3}$ and $N^{-3/4}$, respectively), and the corresponding rigidity result must respect this. For simplicity, here we state only the bulk result, as we did for the local law as well; for results at the edge and cusp, see [6].

Given the self-consistent density ϱ, for any energy E, define

$$i(E) := \Big\lceil N\int_{-\infty}^{E} \varrho(\omega)\mathrm{d}\omega\Big\rceil \tag{6.32}$$

to be the index of the N-quantile closest to E. Alternatively, for any $i \in [\![1, N]\!]$ one could define $\gamma_i = \gamma_i^{(N)}$ to be the i-th N-quantile of ϱ by the relation

$$\int_{-\infty}^{\gamma_i} \varrho(\omega)\mathrm{d}\omega = \frac{i}{N},$$

then clearly $\gamma_{i(E)}$ is (one of) the closest N-quantile to E as long as E is in the bulk, $\varrho(E) > 0$.

Theorem 6.9 (Rigidity of bulk eigenvalues)**.** *Let H be a Wigner type or, more generally, a correlated random matrix, satisfying the conditions of Theorem 6.6 or Theorem 6.7, respectively. Let ϱ be the self-consistent density of states obtained from solving the corresponding Dyson equation. Fix any $\delta, \varepsilon, D > 0$. For any energy E in the bulk, $\varrho(E) \geqslant \delta$, we have*

$$\mathbb{P}\Big(|\lambda_{i(E)} - E| \geqslant \frac{N^\varepsilon}{N}\Big) \leqslant \frac{C}{N^D}. \tag{6.33}$$

Sketch of the proof. The proof of rigidity from the local law is a fairly standard procedure by now, see Chapter 11 of [39], or Lemma 5.1 [6] especially taylored to our situation. The key step is the following *Helffer-Sjöstrand formula* that expresses integrals of a compactly supported function f on the real line against a (signed) measure ν with bounded variation in terms of the Stieltjes transform of ν. (Strictly speaking we defined Stieltjes transform only for probability measures, but the concept can be easily extended since any signed measure with bounded variation can be written as a difference of two non-negative measures, and thus Stieltjes transform extends by linearity).

Let χ be a compactly supported smooth cutoff function on $\mathbb{R}$ such that $\chi \equiv 1$ on $[-1, 1]$. Then the Cauchy integral formula implies

$$f(\tau) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{i\eta f''(\sigma)\chi(\eta) + i(f(\sigma) + i\eta f'(\sigma))\chi'(\eta)}{\tau - \sigma - i\eta}\mathrm{d}\sigma\mathrm{d}\eta. \tag{6.34}$$

Thus for any real valued smooth f the **Helffer-Sjöstrand formula** states that

$$\int_{\mathbb{R}} f(\tau)\nu(\mathrm{d}\tau) = -\frac{1}{2\pi}\big(L_1 + L_2 + L_3\big) \tag{6.35}$$

with

$$L_1 = \int_{\mathbb{R}^2} \eta f''(\sigma)\chi(\eta)\,\mathrm{Im}\, m(\sigma + i\eta)\mathrm{d}\sigma\mathrm{d}\eta$$

$$L_2 = \int_{\mathbb{R}^2} f'(\sigma)\chi'(\eta)\,\mathrm{Im}\, m(\sigma + i\eta)\mathrm{d}\sigma\mathrm{d}\eta$$

$$L_3 = \int_{\mathbb{R}^2} \eta f'(\sigma)\chi'(\eta)\,\mathrm{Re}\, m(\sigma + i\eta)\mathrm{d}\sigma\mathrm{d}\eta$$

where $m(z) = m_\nu(z)$ is the Stieltjes transform of ν. Although this formula is a simple identity, it plays an essential role in various problems of spectral analysis. One may apply it to develop functional calculation (functions of a given self-adjoint operator) in terms of the its resolvents [25].

For the proof of the eigenvalue rigidity, the formula (6.35) is used for $\nu := \mu_N - \varrho$, i.e. for the difference of the empirical and the self-consistent density of states. Since the normalized trace of the resolvent is the Stieltjes transform of the empirical density of states, the averaged local law (6.25) (or (6.31) with $T = 1$) states that

$$|m_\nu(\tau + i\eta)| \leqslant \frac{N^\varepsilon}{N\eta} \qquad \eta \geqslant N^{-1+\gamma} \tag{6.36}$$

with very high probability for any τ with $\varrho(\tau) \geqslant \delta$. Now we fix two energies, τ_1 and τ_2 in the bulk and define f to be the characteristic function of the interval $[\tau_1, \tau_2]$ smoothed out on some scale η_0 at the edges, i.e.

$$f|_{[\tau_1,\tau_2]} = 1, \qquad f|_{\mathbb{R}\setminus[\tau_1-\eta_0,\tau_2+\eta_0]} = 0$$

with derivative bounds $|f'| \leqslant C/\eta_0$, $|f''| \leqslant C/\eta_0^2$ in the transition regimes

$$J := [\tau_1 - \eta_0, \tau_1] \cup [\tau_2, \tau_2 + \eta_0].$$

We will choose $\eta_0 = N^{-1+\gamma}$. Then it is easy to see that L_2 and L_3 are bounded by $N^{-1+\varepsilon+C\gamma}$ since $\chi'(\eta)$ is supported far away from 0, say on $[1, 2] \cup [-2, -1]$, hence, for example

$$|L_2| \lesssim \int_1^2 \mathrm{d}\eta \int_J \mathrm{d}\sigma \, \frac{1}{\eta_0} \, |\chi'(\eta)| \frac{N^\varepsilon}{N\eta} \leqslant N^{-1+\varepsilon+2\gamma}$$

using that $|J| \leqslant 2\eta_0 \lesssim N^{-1+\gamma}$. A similar direct estimate does not work for L_1 since it would give

$$|L_1| \lesssim \int_0^\infty \mathrm{d}\eta \int_J \mathrm{d}\sigma \, \eta \, \frac{1}{\eta_0^2} \, \chi(\eta) \frac{N^\varepsilon}{N\eta} \leqslant N^{\varepsilon+3\gamma}. \tag{6.37}$$

Even this estimate would need a bit more care since the local law (6.36) does not hold for η smaller than $N^{-1+\gamma}$, but here one uses the fact that for any positive measure μ, the (positive) function $\eta \to \eta \operatorname{Im} m_\mu(\sigma + i\eta)$ is monotonously increasing, so the imaginary part of the Stieltjes transform at smaller η-values can be controlled by those at larger η values. Here it is crucial that L_1 contains only the imaginary part of the Stieltjes transforms and not the entire Stieltjes transform. The argument (6.37), while does not cover the entire L_1, it gives a sufficient bound on the small η regime:

$$\int_0^{\eta_0} \mathrm{d}\eta \int_J \mathrm{d}\sigma \eta f''(\sigma)\chi(\eta) \operatorname{Im} m(\sigma + i\eta) \mathrm{d}\sigma \mathrm{d}\eta \leqslant \int_0^{\eta_0} \mathrm{d}\eta \int_J \mathrm{d}\sigma |f''(\sigma)| \eta_0 \operatorname{Im} m(\sigma + i\eta_0) \mathrm{d}\sigma \mathrm{d}\eta \leqslant N^{-1+\varepsilon+3\gamma}.$$

To improve (6.37) by a factor $1/N$ for $\eta \geqslant \eta_0$, we integrate by parts before estimating. First we put one σ-derivative from f'' to $m_\nu(\sigma+i\eta)$, then the ∂_σ derivate is switched to ∂_η derivative, then another integration by parts, this time in η removes the derivative from m_ν. The boundary terms, we obtain formulas similar to L_2 and L_3 that have already been estimated.

The outcome is that

$$\int_{\mathbb{R}} f(\tau) \big[\mu_N(\mathrm{d}\tau) - \varrho(\tau)\mathrm{d}\tau\big] \leqslant N^{-1+\varepsilon'} \tag{6.38}$$

for any $\varepsilon' > 0$ with very high probability, since ε and γ can be chosen arbitrarily small positive numbers in the above argument. If f were exactly the characteristic function, then (6.38) would imply that

$$\frac{1}{N} \#\big\{\lambda_j \in [\tau_1, \tau_2]\big\} = \int_{\tau_1}^{\tau_2} \varrho(\omega)\mathrm{d}\omega + O(N^{-1+\varepsilon'}) \tag{6.39}$$

i.e. it would identify the eigenvalue counting function down to the optimal scale. Estimating the effects of the smooth cutoffs is an easy technicality. Finally, (6.39) can be easily turned into (6.33), up to one more catch. So far we assumed that τ_1, τ_2 are both in the bulk since the local law was formulated in the bulk and (6.39) gave the number of eigenvalues in any interval with endpoints in the bulk. The quantiles appearing in (6.33), however, involve semi-infinite intervals, so one also needs a local law well outside of the bulk. Although in Theorems 6.6 and 6.7 we formulated local laws in the bulk, similar, and typically even easier estimates are available for energies far away from the support of ϱ. In fact, in the regime where $\operatorname{dist}(\tau, \operatorname{supp}\varrho) \geqslant \delta$ for some fixed $\delta > 0$, the analogue (6.36) is improved to

$$|m_\nu(\tau + i\eta)| \leqslant \frac{N^\varepsilon}{N} \qquad \eta > 0, \tag{6.40}$$

so the above estimates on L_j's are even easier when τ_1 or τ_2 is far away from the bulk. □

6.3.3 Universality of local eigenvalue statistics

The universality of the local distribution of the eigenvalues is the main coveted goal of random matrix theory. While local laws and rigidity are statements where random quantities are compared with deterministic ones, i.e. they are, in essence, law of large number type results (even if not always formulated in that way), the universality is about the emergence and ubiquity of a new distribution.

We will formulate universality in two forms: on the level of correlation functions and on the level of individual gaps. While these formulations are "morally" equivalent, technically they require quite different proofs.

We need to strengthen a bit the assumption on the lower bound on the variances in (6.5) for complex hermitian Wigner type matrices H. In this case we define the real symmetric 2×2 matrix

$$\sigma_{ij} := \begin{pmatrix} \mathbb{E}(\operatorname{Re} h_{ij})^2 & \mathbb{E}(\operatorname{Re} h_{ij})(\operatorname{Im} h_{ij}) \\ \mathbb{E}(\operatorname{Re} h_{ij})(\operatorname{Im} h_{ij}) & \mathbb{E}(\operatorname{Im} h_{ij})^2 \end{pmatrix}$$

for every i, j and we will demand that

$$\sigma_{ij} \geqslant \frac{c}{N} \tag{6.41}$$

with some $c > 0$ uniformly for all i, j. Similarly, for correlated matrices the flatness condition (6.18) is strengthened to the requirement that there is a constant $c > 0$ such that

$$\mathbb{E}|\operatorname{Tr} BH|^2 \geqslant c \operatorname{Tr} B^2 \tag{6.42}$$

for any real symmetric (or complex hermitian, depending on the symmetry class of H) deterministic matrix B.

Theorem 6.10 (Bulk universality)**.** *Let H be a Wigner type or, more generally, a correlated random matrix, satisfying the conditions of Theorem 6.6 or Theorem 6.7, respectively. For Wigner type matrices in the complex hermitian symmetry class we additionally assume* (6.41). *For correlated random matrices, we additionally assume* (6.42). *Let ϱ be the self-consistent density of states obtained from solving the corresponding Dyson equation. Let $k \in \mathbb{N}$, $\delta > 0, E \in \mathbb{R}$ with $\varrho(E) \geqslant \delta$ and let $\Phi\colon \mathbb{R}^k \to \mathbb{R}$ be a compactly supported smooth test function. Then for some positive constants c and C, depending on Φ, δ, k, we have the following:*

(i) [Universality of correlation functions] Denote the k-point correlation function of the eigenvalues of H by $p_N^{(k)}$ (see (1.15)*) and denote the corresponding k-point correlation function of the GOE/GUE-point process by $\Upsilon^{(k)}$. Then*

$$\left| \int_{\mathbb{R}^k} \Phi(\mathbf{t}) \left[\frac{1}{\rho(E)} p_N^{(k)} \Big(E + \frac{\mathbf{t}}{N\rho(E)} \Big) - \Upsilon_k(\mathbf{t}) \right] \mathrm{d}\mathbf{t} \right| \leqslant CN^{-c}. \tag{6.43}$$

(ii) [Universality of gap distributions] Recall that $i(E)$ is the index of the N-th quantile in the density ϱ that is closest to the energy E (6.32)*. Then*

$$\left| \mathbb{E}\Phi\Big(\big(N\rho(\lambda_{k(E)})[\lambda_{k(E)+j} - \lambda_{k(E)}] \big)_{j=1}^k \Big) - \mathbb{E}_{GOE/GUE}\Phi\Big(\big(N\rho_{sc}(0)[\lambda_{\lceil N/2 \rceil + j} - \lambda_{\lceil N/2 \rceil}] \big)_{j=1}^k \Big) \right| \leqslant CN^{-c}, \tag{6.44}$$

where the expectation $\mathbb{E}_{GOE/GUE}$ is taken with respect to the Gaussian matrix ensemble in the same symmetry class as H.

Short sketch of the proof. The main method to prove universality is the *three-step strategy* outlined in Section 1.2.4. The first step is to obtain a local law which serves as an *a priori* input for the other two steps and it is the only model dependent step. The second step is to show that a small Gaussian component in the distribution already produces the desired universality. The third step is a perturbative argument to show that removal of the Gaussian component does not change the local statistics. There have been many theorems of increasing generality to complete the second and third steps and by now very general "black-box" theorems exist that are model-independent.

The **second step** relies on the local equilibration properties of the Dyson Brownian motion introduced in (1.20). The latest and most general formulation of this idea concerns universality of deformed Wigner matrices of the form

$$H_t = V + \sqrt{t}W,$$

where V is a deterministic matrix and W is a GOE/GUE matrix. In applications V itself is a random matrix and in H_t an additional independent Gaussian component is added. But for the purpose of local equilibration of the DBM, hence for the emergence of the universal local statistics, only the randomness of W is used, hence one may condition on V. The main input of the following result is that the local eigenvalue density of V must be controlled in a sense of lower and upper bounds on the imaginary part of the Stieltjes transform m_V of the empirical eigenvalue density of V. In practice this is obtained from the local law with very high probability in the probability space of V.

Theorem 6.11 ([60,61])**.** *Choose two N-dependent parameters, L, ℓ such that $1 \gg L^2 \gg \ell \gg N^{-1}$ (here the notation $\gg$ indicates separation by an N^ε factor for an arbitrarily small $\varepsilon > 0$). Suppose that around a fixed energy E_0 in a window of size L the local eigenvalue density of V on scale ℓ is controlled, i.e.*

$$c \leqslant \operatorname{Im} m_V(E + i\eta) \leqslant C, \qquad E \in (E_0 - L, E_0 + L), \qquad \eta \in [\ell, 10]$$

(in particular, E_0 is in the bulk of V). Assume also that $\|V\| \leqslant N^C$. Then for any t with $N^\varepsilon \ell \leqslant t \leqslant N^{-\varepsilon} L^2$ the bulk universality of H_t around E_0 holds both in the sense of correlation functions at fixed energy (6.43) *and in sense of gaps* (6.44)*.*

Theorem 6.11 in this general form appeared in [61] (gap universality) and in [60] (correlation functions universality at fixed energy). These ideas have been developed in several papers. Earlier results concerned Wigner or generalized Wigner matrices and proved correlation function universality with a small energy averaging [35,36], fixed energy universality [22] and gap universality [38]. Averaged energy and gap universality for random matrices with general density profile were also proven in [37] assuming more precise information on m_V that are available from the optimal local laws.

Finally, the **third step** is to remove the small Gaussian component by realizing that the family of matrices of the form $H_t = V + \sqrt{t}W$ to which Theorem 6.11 applies is sufficiently rich so that for any given random matrix H there exists a matrix V and a small t so that the local statistics of H and $H_t = V + \sqrt{t}W$ coincide. We will use this result for some t with $t = N^{-1+\gamma}$ with a small γ. The time t has to be much larger than ℓ and ℓ has to be much larger than N^{-1} since below that scale the local density of V (given by $\operatorname{Im} m_V(E+i\eta)$) is not bounded. But t cannot be too large either otherwise the comparison result cannot hold.

Note that the local statistics is not compared directly with that of V; this would not work even for Wigner matrices V and even if we used the Ornstein Uhlenbeck process, i.e. $H_t = e^{-t/2}V + \sqrt{1 - e^{-t}}W$ (for Wigner matrices V the OU process has the advantage that it preserves not only the first but also the second moments of H_t). But for any given Wigner-type ensemble H one can find a random V and an independent Gaussian W so that the first three moments of H and $H_t = e^{-t/2}V + \sqrt{1 - e^{-t}}W$ coincide and the fourth moments are very close; this freedom is guaranteed by the lower bound on s_{ij} and σ_{ij} (6.41).

The main perturbative result is the following *Green function comparison theorem* that allows us to compare *expectations* of reasonable functions of the Green functions of two different ensembles whose first four moments (almost) match (the idea of matching four moments in random matrices was introduced in [73]). The key point is that $\eta = \operatorname{Im} z$ can be slightly below the critical threshold $1/N$: the expectation regularizes the possible singularity. Here is the prototype of such a theorem:

Theorem 6.12 (Green function comparison)**.** *[44] Consider two Wigner type ensembles H and $\widetilde{H}$ such that their first two moments are the same, i.e. the matrices of variances coincide, $S = \widetilde{S}$ and the third and fourth moments almost match in a sense that*

$$\big| E h_{ij}^s - \mathbb{E}\widehat{h}_{ij}^s \big| \leqslant N^{-2-\delta}, \qquad s = 3, 4 \tag{6.45}$$

(for the complex hermitian case all mixed moments of order 3 and 4 should match). Define a sequence of interpolating Wigner-type matrices $H_0, H_1, H_2, \ldots$ such that $H_0 = H$, then in H_1 the h_{11} matrix element is

replaced with $\widetilde{h}_{11}$, in H_2 the h_{11} and h_{12} elements are replaced with $\widetilde{h}_{11}$ and $\widetilde{h}_{12}$, etc., i.e. we replace one by one the distribution of the matrix elements. Suppose that the Stieltjes transform on scale $\eta = N^{-1+\gamma}$ is bounded for all these interpolating matrices and for any $\gamma > 0$. Set now $\eta' := N^{-1-\gamma}$ and let Φ a smooth function with moderate growth. Then

$$\Big|\mathbb{E}\Phi\big(G(E+i\eta')\big) - \widetilde{\mathbb{E}}\Phi\big(G(E+i\eta')\big)\Big| \leqslant N^{-\delta+C\gamma} \tag{6.46}$$

and similar multivariable versions also hold.

In the applications, choosing γ sufficiently small, we could conclude that the distribution of the Green functions of H and $\widetilde{H}$ on scale even **below the eigenvalue spacing** are close. On this scale local correlation functions can be identified, so we conclude that the local eigenvalue statistics of H and $\widetilde{H}$ are the same. This will conclude step 3 of the three step strategy and finish the proof of bulk universality, Theorem 6.10. □

Idea of the proof of Theorem 6.12. The proof of (6.46) is a "brute force" resolvent and Taylor expansion. For simplicity, we first replace Φ by its finite Taylor polynomial. Moreover, we consider only the linear term for illustration in this proof. We estimate the change of $\mathbb{E}G(E+i\eta')$ after each replacement; we need to bound each of them by $o(N^{-2})$ since there are of order N^2 replacements. Fix an index pair i, j. Suppose we are at the step when we change the (ij)-th matrix element h_{ij} to $\widetilde{h}_{ij}$. Let R denote the resolvent of the matrix with (ij)-th and (ji)-th elements being zero, in particular R is independent of h_{ij}. It is easy to see from the local law that $\max_{ab}|R_{ab}(E+i\eta)| \lesssim 1$ for any $\eta \geqslant N^{-1+\gamma}$ and therefore, by the monotonicity of $\eta \to \eta \operatorname{Im} m(E+i\eta)$ we find that $|R_{ab}(E+i\eta') \lesssim N^{2\gamma}$. Then simple resolvent expansion gives, schematically, that

$$G = R + Rh_{ij}R + Rh_{ij}Rh_{ij}R + Rh_{ij}Rh_{ij}Sh_{ij}R + Rh_{ij}Rh_{ij}Rh_{ij}Rh_{ij}R + \dots \tag{6.47}$$

and a similar expansion for $\widetilde{G} = G(h_{ij} \leftrightarrow \widetilde{h}_{ij})$ where all h_{ij} is replaced with $\widetilde{h}_{ij}$ (strictly speaking we need to replace h_{ij} and $h_{ji} = \bar{h}_{ij}$ simultaneously due to hermitian symmetry, but we neglect this). We do the expansion up to the fourth order terms (counting the number of h's). The naive size of a third order term, say, $Rh_{ij}Rh_{ij}Rh_{ij}R$ is of order $N^{-3/2+8\gamma}$ since every h_{ij} is of order $N^{-1/2}$. However, the difference in $\mathbb{E}$ and $\widetilde{\mathbb{E}}$-expectations of these terms are of order $N^{-2-\delta}$ by (6.45). Thus for the first four terms (fully expanded ones) in (6.47) it holds that

$$\mathbb{E}G - \widetilde{\mathbb{E}}\widetilde{G} = O(N^{-2-\delta+C\gamma}) + \text{fifth and higher order terms}$$

But all fifth and higher order terms have at least five h factors so their size is essentially $N^{-5/2}$, i.e. negligible, even without any cancellation between G and $\widetilde{G}$. Finally, we need to repeat this one by one replacement N^2 times, so we arrive at a bound of order $N^{-\delta+C\gamma}$. This proves (6.46).

□

Exercise 6.13. *For a given real symmetric matrix V let H_t solve the SDE*

$$\mathrm{d}H_t = \frac{\mathrm{d}B_t}{\sqrt{N}}, \qquad H_{t=0} = V$$

where $B_t = B(t)$ is a standard real symmetric matrix valued Brownian motion, i.e. the matrix elements $b_{ij}(t)$ for $i < j$ as well as $b_{ii}(t)/\sqrt{2}$ are independent standard Brownian motions and $b_{ij}(t) = b_{ji}(t)$. Prove that the eigenvalues of H_t satisfy the following coupled system of stochastic differential equations (Dyson Brownian motion):

$$\mathrm{d}\lambda_a = \sqrt{\frac{2}{N}}\mathrm{d}B_a + \frac{1}{N}\sum_{b\neq a}\frac{1}{\lambda_a - \lambda_b}\mathrm{d}t, \qquad a \in [\![1, N]\!]$$

where $\{B_a \ : \ a \in [\![1, N]\!]\}$ is a collection of independent standard Brownian motions with initial condition $\boldsymbol{\lambda}_a(t=0)$ given by the eigenvalues of V. Hint: Use first and second order perturbation theory to differentiate the eigenvalue equation $H\mathbf{u}_a = \lambda_a\mathbf{u}_a$ with the side condition $\langle \mathbf{u}_a, \mathbf{u}_b\rangle = \delta_{ab}$. then use Ito formula. (see Section 12.2 of [39]).

7 Analysis of the vector Dyson equation

In this section we outline the proof of a few results concerning the vector Dyson equation (6.1)

$$-\frac{1}{\mathbf{m}} = z + S\mathbf{m}, \qquad z \in \mathbb{H}, \quad \mathbf{m} \in \mathbb{H}^N, \tag{7.1}$$

where $S = S^t$ is symmetric, bounded, $\|S\| \leqslant C$ and has nonnegative entries.

We recall the convention that $1/\mathbf{m}$ denotes a vector in $\mathbb{C}^N$ with components $1/m_j$. Similarly, the relation $\mathbf{u} \leqslant \mathbf{v}$ and the product $\mathbf{uv}$ of two vectors are understood in coordinate-wise sense.

7.1 Existence and uniqueness

We sketch the existence and uniqueness result, i.e. Theorem 6.1, a detailed proof can be found in Chapter 4 [4]. To orient the reader here we only mention that it is a fix-point argument for the map

$$\Phi(\mathbf{u}) := -\frac{1}{z + S\mathbf{u}} \tag{7.2}$$

that maps $\mathbb{H}^N$ to $\mathbb{H}^N$ for any fixed $z \in \mathbb{H}$. Denoting by

$$D(\zeta, \omega) := \frac{|\zeta - \omega|^2}{(\operatorname{Im}\zeta)(\operatorname{Im}\omega)}, \qquad \zeta, \omega \in \mathbb{H}$$

the standard hyperbolic metric on the upper half plane, one may check that Φ is a contraction in this metric. More precisely, for any fixed constant η_0 we have

$$\max_j D\Big(\Phi(\mathbf{u})_j, \Phi(\mathbf{w})_j\Big) \leqslant \Big(1 + \frac{\eta_0^2}{\|S\|}\Big)^{-2} \max_j D(u_j, w_j) \tag{7.3}$$

assuming that $\operatorname{Im} z \geqslant \eta_0$ and both $\mathbf{u}$ and $\mathbf{w}$ lie in a large compact set

$$B_{\eta_0} := \Big\{\mathbf{u} \in \mathbb{H}^N \ : \ \|\mathbf{u}\| \leqslant \frac{1}{\eta_0}, \quad \inf_j \operatorname{Im} u_j \geqslant \frac{\eta_0^2}{(2 + \|S\|)^2}\Big\}, \tag{7.4}$$

that is mapped by Φ into itself. Here $\|\mathbf{u}\| = \max_j |u_j|$. Once setting up the contraction properly, the rest is a straightforward fixed point theorem. The representation (6.3) follows from the Nevanlinna's theorem as mentioned after Definition 2.1.

Given (6.3), we recall that $\varrho = \langle \boldsymbol{\nu} \rangle = \frac{1}{N}\sum_j \nu_j$ is the self-consistent density of states. We consider its harmonic extension to the upper half plane and continue to denote it by ϱ:

$$\varrho = \varrho(z) := \frac{\eta}{\pi}\int_{\mathbb{R}} \frac{\varrho(\mathrm{d}\tau)}{|x - E|^2 + \eta} = \frac{1}{\pi}\langle \operatorname{Im} \mathbf{m}(z)\rangle, \qquad z = E + i\eta. \tag{7.5}$$

Exercise 7.1. *Check directly from* (7.1) *that the solution satisfies the additional condition of the Nevanlinna's theorem, i.e. that for every j we have $i\eta m_j(i\eta) \to -1$ as $\eta \to \infty$. Moreover, check that $|m_j(z)| \leqslant 1/\operatorname{Im} z$.*

Exercise 7.2. *Prove that the support of all measures ν_i lie in $[-2\sqrt{\|S\|}, 2\sqrt{\|S\|}]$.*
Hint: suppose $|z| > 2\sqrt{\|S\|}$, then check the following implication:

$$\textit{If} \quad \|\mathbf{m}(z)\| < \frac{|z|}{2\|S\|}, \quad \textit{then} \quad \|\mathbf{m}(z)\| < \frac{2}{|z|}$$

and apply a continuity argument to conclude that $\|\mathbf{m}(z)\| < \frac{2}{|z|}$ holds unconditionally. Taking the imaginary part of (7.1) *conclude that $\operatorname{Im} \mathbf{m}(E + i\eta) \to 0$ as $\eta \to 0$ for any $|E| > 2\sqrt{\|S\|}$.*

Exercise 7.3. *Prove the inequality* (7.3)*, i.e. that Φ is indeed a contraction on B_{η_0}. Hint: Prove and then use the following properties of the metric D:*

1) The metric D is invariant under linear fractional transformations of $\mathbb{H}$ of the form

$$f(z) = \frac{az+b}{cz+d}, \qquad z \in \mathbb{H}, \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{R}).$$

2) Contraction

$$D(z+i\lambda, w+i\lambda) = \Big(1+\frac{\lambda}{\operatorname{Im} z}\Big)^{-1}\Big(1+\frac{\lambda}{\operatorname{Im} w}\Big)^{-1} D(z,w), \qquad z, w \in \mathbb{H}, \qquad \lambda > 0$$

3) Convexity: Let $\mathbf{a} = (a_1, \ldots, a_N) \in \mathbb{R}_+^N$, then

$$D\big(\sum_i a_i u_i, \sum_i a_i w_i\big) \leqslant \max_i D(u_i, w_i), \qquad \mathbf{u}, \mathbf{w} \in \mathbb{H}^N.$$

7.2 Bounds on the solution

Now we start the quantitative analysis of the solution and we start with a result on the boundedness in the bulk. We introduce the maximum norm and the ℓ^p norms on $\mathbb{C}^N$ as follows:

$$\|\mathbf{u}\| := \|\mathbf{u}\|_\infty = \max_j |u_j|, \qquad \|\mathbf{u}\|_p^p := \frac{1}{N}\sum_j |u_j|^2 = \langle |\mathbf{u}|^p \rangle.$$

The procedure to bound $\mathbf{m}$ is that we first obtain an ℓ^2-bound which usually requires less conditions. Then we enhance it to an ℓ^∞ bound. First we obtain a bound that is useful in the bulk but deteriorates as the self-consistent density vanishes, e.g. at the edges and cusps. Second, we improve this bound to one that is also useful near the edges/cusps but this requires some additional regularity condition on s_{ij}. In these notes we will not aim at the most optimal conditions, see [5] and [4] for the detailed analysis.

7.2.1 Bounds useful in the bulk

Theorem 7.4. *[Bounds on the solution] We assume the lower and upper bounds of the form (see* (6.5)*):*

$$\frac{c}{N} \leqslant s_{ij} \leqslant \frac{C}{N}. \tag{7.6}$$

Then we have

$$\|\mathbf{m}\|_2 \lesssim 1, \qquad |\mathbf{m}(z)| \leqslant \frac{1}{\varrho(z) + dist(z, supp\varrho)}, \qquad \frac{1}{|\mathbf{m}(z)|} \lesssim 1 + |z|,$$

and

$$\varrho(z) \lesssim \operatorname{Im} \mathbf{m}(z) \lesssim (1+|z|)^2 \|\mathbf{m}(z)\|^2 \varrho(z),$$

where we recall that $\lesssim$ indicates a bound up to an unspecified multiplicative constant that is independent of N (also, recall that the last three inequalities are understood in coordinate-wise sense).

Proof. For simplicity, in the proof we assume that $|z| \lesssim 1$; the large z regime is much easier and follows directly from the Stieltjes transform representation of $\mathbf{m}$. Taking the imaginary part of the Dyson equation (7.1), we have

$$\frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|^2} = \eta + S \operatorname{Im} \mathbf{m}. \tag{7.7}$$

Using the lower bound from (7.6), we get

$$S \operatorname{Im} \mathbf{m} \geqslant c \langle \operatorname{Im} \mathbf{m} \rangle \geqslant c\varrho$$

thus

$$\operatorname{Im}\mathbf{m} \geqslant c|\mathbf{m}|^2\varrho. \tag{7.8}$$

Taking the average of both sides and dividing by $\varrho > 0$, we get $\|\mathbf{m}\|_2 \lesssim 1$. Using $\operatorname{Im}\mathbf{m} \leqslant |\mathbf{m}|$, we immediately get an upper bound on $|\mathbf{m}| \lesssim 1/\varrho$. The alternative bound

$$|\mathbf{m}(z)| \lesssim \frac{1}{\operatorname{dist}(z, \operatorname{supp}\varrho)}$$

follows from the Stieltjes transform representation (6.3).

Next, we estimate the rhs. of (7.1) trivially, we have

$$\frac{1}{|m_i|} \leqslant |z| + \sum_j s_{ij}|m_j| \lesssim |z| + \|\mathbf{m}\|_1 \leqslant |z| + \|\mathbf{m}\|_2 \lesssim 1$$

using Hölder inequality in the last but one step. This gives the upper bound on $1/|\mathbf{m}|$.

Using this bound, we can conclude from (7.8) that $\varrho \lesssim \operatorname{Im}\mathbf{m}$. The upper bound on $\operatorname{Im}\mathbf{m}$ also follows from (7.7) and (7.6):

$$\frac{\operatorname{Im}\mathbf{m}}{|\mathbf{m}|^2} \leqslant \eta + S\operatorname{Im}\mathbf{m} \leqslant \eta + \langle \operatorname{Im}\mathbf{m}\rangle \lesssim \eta + C\varrho.$$

Using that

$$\varrho(z) \gtrsim \frac{\eta}{(1+|z|)^2}$$

which can be easily checked from (7.5) and the boundedness of the support of ϱ, we conclude the two-sided bounds on $\operatorname{Im}\mathbf{m}$. □

Notice two weak points when using this relatively simple argument. First, the lower bound in (7.6) was heavily used, although much less assumption is sufficient. We will not discuss these generalizations in these notes, but see Theorem 2.11 of [4] and remarks thereafter addressing this issue. Second, the upper bound on $|\mathbf{m}|$ for small η is useful only inside the self-consistent bulk spectrum or away from the support of ϱ, it deteriorates near the edges of the spectrum. In the next sections we remedy this situation.

7.2.2 Unconditional ℓ^2-bound away from zero

Next, we present a somewhat surprising result that shows that an ℓ^2-bound on the solution, $\|\mathbf{m}(z)\|_2$, away from the only critical point $z = 0$ is possible *without any condition on S*. The spectral parameter $z = 0$ is clearly critical, e.g. if $S = 0$, the solution $\mathbf{m}(z) = -1/z$ blows up. Thus to control the behavior of $\mathbf{m}$ around $z \approx 0$ one needs some non degeneracy condition on S. We will not address the issue of $z \approx 0$ in these notes, but we remark that a fairly complete picture was obtained in Chapter 6 of [4] using the concept of *fully indecomposability.*

Before presenting the ℓ^2-bound away from zero, we introduce an important object, the **saturated self-energy operator**, that will also play a key role later in the stability analysis:

Definition 7.5. *Let S be a symmetric matrix with nonnegative entries and let $\mathbf{m} = \mathbf{m}(z) \in \mathbb{H}^N$ solve the vector Dyson equation* (7.1) *for some fixed spectral parameter $z \in \mathbb{H}$. The matrix $F = (F_{ij})$ with*

$$F_{ij} := |m_i|s_{ij}|m_j|$$

acting as

$$F\mathbf{u} = |\mathbf{m}|S\big(|\mathbf{m}|\mathbf{u}\big), \quad \textit{i.e.} \quad \big(F\mathbf{u}\big)_i = |m_i|\sum_j s_{ij}|m_j|u_j \tag{7.9}$$

on any vector $\mathbf{u} \in \mathbb{H}^N$, is called the **saturated self-energy operator**.

Suppose that S has strictly positive entries. Since $m_i \neq 0$ from (7.1), clearly F has also positive entries, and $F = F^*$. Thus the Perron-Frobenius theorem applies to F, and it guarantees that F has a single largest eigenvalue r (so that for any other eigenvalue λ we have $|\lambda| < r$) and the corresponding eigenvector $\mathbf{f}$ has positive entries:

$$F\mathbf{f} = r\mathbf{f}, \qquad \mathbf{f} > 0.$$

Moreover, since F is symmetric, we have $\|F\| = r$ for the usual Euclidean matrix norm of F.

Proposition 7.6. *Suppose that S has strictly positive entries and let $\mathbf{m}$ solve (7.1) for some $z = E + i\eta \in \mathbb{H}$. Then the norm of the saturated self-energy operator is given by*

$$\|F\| = 1 - \eta \frac{\langle \mathbf{f}|\mathbf{m}|\rangle}{\langle \mathbf{f} \frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|}\rangle}, \tag{7.10}$$

in particular $\|F\| < 1$. Moreover,

$$\|\mathbf{m}(z)\|_2 \leqslant \frac{2}{|z|}. \tag{7.11}$$

We remark that for the bounds $\|F\| < 1$ and (7.11) it is sufficient if S has nonnegative entries instead of positive entries; the proof requires a bit more care, see Lemma 4.5 [4].

Proof. Taking the imaginary part of (7.1) and multiplying it by $|\mathbf{m}|$, we have

$$\frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|} = \eta|\mathbf{m}| + |\mathbf{m}|S\Big(|\mathbf{m}|\frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|}\Big) = \eta|\mathbf{m}| + F\,\frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|}. \tag{7.12}$$

Scalar multiply this equation by $\mathbf{f}$, use the symmetry of F and $F\mathbf{f} = \|F\|\mathbf{f}$ to get

$$\Big\langle \mathbf{f}, \frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|}\Big\rangle = \eta\langle \mathbf{f}|\mathbf{m}|\rangle + \Big\langle \mathbf{f}, F\frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|}\Big\rangle = \eta\langle \mathbf{f}|\mathbf{m}|\rangle + \|F\|\Big\langle \mathbf{f}, \frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|}\Big\rangle,$$

which is equivalent to (7.10) (note that $\langle , \rangle$ as a binary operation is the scaler product while $\langle \cdot \rangle$ is the averaging).

For the bound on $\mathbf{m}$, we write (7.1) as $-z\mathbf{m} = 1 + \mathbf{m}S\mathbf{m}$, so taking the ℓ^2-norm, we have

$$\|\mathbf{m}\|_2 \leqslant \frac{1}{|z|}\big(1 + \|\mathbf{m}S\mathbf{m}\|_2\big) \leqslant \frac{1}{|z|}\big(1 + \big\||\mathbf{m}|S|\mathbf{m}|\big\|_2\big) = \frac{1}{|z|}\big(1 + \|F\mathbf{1}\|_2\big) \leqslant \frac{2}{|z|},$$

where $\mathbf{1} = (1, 1, 1, \ldots)$, note that $\|\mathbf{1}\|_2 = 1$ and we used (7.10) in the last step. □

7.2.3 Bounds valid uniformly in the spectrum

In this section we introduce an extra regularity assumption that enables us to control $\mathbf{m}$ uniformly throughout the spectrum, including edges and cusps. For simplicity, we restrict our attention to the special case when s_{ij} originates from a piecewise continuous nonnegative profile function $S(x, y)$ defined on $[0, 1] \times [0, 1]$, i.e. we assume

$$s_{ij} = \frac{1}{N} S\Big(\frac{i}{N}, \frac{j}{N}\Big). \tag{7.13}$$

We will actually need that S is piecewise 1/2-Hölder continuous (6.8). The theorem holds under a weaker condition called *component regularity*, see Assumption (C) in [5] but we omit these details.

Theorem 7.7. *Assume that s_{ij} is given by (7.13) with a piecewise Hölder-1/2 continuous function S with uniform lower and upper bounds $c \leqslant S(x, y) \leqslant C$. Then for any $R > 0$ and for any $|z| \leqslant R$ we have*

$$|\mathbf{m}(z)| \sim 1, \qquad \operatorname{Im} m_i(z) \sim \operatorname{Im} m_j(z),$$

where the implicit constants in the $\sim$ relation depend only on c, C and R. In particular, all components of $\operatorname{Im} \mathbf{m}$ *are comparable, hence*

$$\operatorname{Im} m_i \sim \langle \operatorname{Im} \mathbf{m} \rangle = \varrho. \tag{7.14}$$

The theorem also holds under weaker conditions: the uniform lower bound can be replaced with a condition called *diagonal positivity* see Assumption (A) in [5] called *component regularity*, see Assumption (C) in [5] but we omit these generalizations here.

Proof. We have already obtained an ℓ^2-bound $\|\mathbf{m}\|_2 \lesssim 1$ in Theorem 7.4. Now we consider any two indices i, j, evaluate (7.1) at these points and subtract them. From

$$-\frac{1}{m_i} = z + (S\mathbf{m})_i, \qquad -\frac{1}{m_j} = z + (S\mathbf{m})_j$$

we thus obtain

$$\Big|\frac{1}{m_i}\Big| \leqslant \Big|\frac{1}{m_j}\Big| + \sum_k |s_{ik} - s_{jk}||m_k| \leqslant \Big|\frac{1}{m_j}\Big| + \|\mathbf{m}\|_2 \Big(N \sum_k |s_{ik} - s_{jk}|^2\Big)^{1/2}.$$

Using (7.13) and the Hölder continuity (for simplicity assume $n = 1$), we have

$$N \sum_k |s_{ik} - s_{jk}|^2 \leqslant \frac{1}{N} \sum_k \Big| S\Big(\frac{i}{N}, \frac{k}{N}\Big) - S\Big(\frac{j}{N}, \frac{k}{N}\Big)\Big|^2 \leqslant C \frac{|i-j|}{N},$$

thus

$$\Big|\frac{1}{m_i}\Big| \leqslant \Big|\frac{1}{m_j}\Big| + C' \sqrt{\frac{|i-j|}{N}}.$$

Taking the reciprocal and squaring it we have for every fixed j that

$$\frac{1}{N} \sum_i \left[\frac{1}{\Big|\frac{1}{m_j}\Big| + C' \sqrt{\frac{|i-j|}{N}}} \right]^2 \leqslant \frac{1}{N} \sum_i |m_i|^2 = \|\mathbf{m}\|_2^2 \lesssim 1.$$

The left hand side is can be estimated from below by

$$\frac{1}{N} \sum_i \left[\frac{1}{\Big|\frac{1}{m_j}\Big| + C' \sqrt{\frac{|i-j|}{N}}} \right]^2 \gtrsim \frac{1}{N} \sum_i \frac{1}{\frac{1}{|m_j|^2} + \frac{|i-j|}{N}} \gtrsim \log |m_j|.$$

Combining the last two inequalities, this shows the uniform upper bound

$$|\mathbf{m}| \lesssim 1.$$

The lower bound is obtained from

$$\Big|\frac{1}{m_i}\Big| = \big|z + \sum_j s_{ij} m_j\big| \leqslant |z| + \frac{C}{N} \sum |m_j| \lesssim 1$$

using the upper bound $|\mathbf{m}| \lesssim 1$ and $s_{ij} \lesssim 1/N$. This proves $|\mathbf{m}| \sim 1$.

The comparability of the components of $\operatorname{Im} \mathbf{m}$ now follows from the imaginary part of (7.1), $|\mathbf{m}| \sim 1$ and that $S(\operatorname{Im} \mathbf{m}) \sim \langle \operatorname{Im} \mathbf{m} \rangle$:

$$\frac{\operatorname{Im} \mathbf{m}}{|\mathbf{m}|^2} = \eta + S\big(\operatorname{Im} \mathbf{m}\big) \quad \Longrightarrow \quad \operatorname{Im} \mathbf{m} \sim \eta + \langle \operatorname{Im} \mathbf{m} \rangle$$

completing the proof. □

7.3 Regularity of the solution and the stability operator

In this section we prove some parts of the regularity Theorem 6.3. We will not go into the details of the edge and cusp analysis here, see [5] for a shorter qualitative analysis and [4] for the full quantitative analysis of all possible singularities. Here we will only show the 1/3-Hölder regularity (6.10). We will use this opportunity to introduce and analyze the key stability operator of the problem which then will also be used in the random matrix part of our analysis.

It is to keep in mind that the small $\eta = \operatorname{Im} z$ regime is critical; typically bounds of order $1/\eta$ or $1/\eta^2$ are easy to obtain but these are useless for local analysis (recall that η indicates the scale of the problem). For the fine regularity properties of the solution, one needs to take $\eta \to 0$ with uniform controls. For the random matrix part, we will take η down to $N^{-1+\gamma}$ for any small $\gamma > 0$, so any $1/\eta$ bound would not be affordable.

Proof of (i) and (iii) from Theorem 6.3. We differentiate (7.1) with respect to z (note that $\mathbf{m}(z)$ is real analytic by (6.3) for any $z \in \mathbb{H}$).

$$-\frac{1}{\mathbf{m}} = z + S\mathbf{m} \quad \Longrightarrow \quad \frac{\partial_z \mathbf{m}}{\mathbf{m}^2} = 1 + S\partial_z \mathbf{m} \quad \Longrightarrow \quad \partial_z \mathbf{m} = \frac{1}{1 - \mathbf{m}^2 S}\mathbf{m}^2. \tag{7.15}$$

The inverse of the (z-dependent) linear operator $1 - \mathbf{m}^2 S$ is called the **stability operator**. We will later prove the following main bound on this operator:

Lemma 7.8 (Bound on the stability operator)**.** *Suppose that for any $z \in \mathbb{H}$ with $|z| \leqslant C$ we have $|\mathbf{m}(z)| \sim 1$. Then*

$$\Big\| \frac{1}{1 - \mathbf{m}^2 S} \Big\| \lesssim \frac{1}{\varrho(z)^2} = \frac{1}{\langle \operatorname{Im} \mathbf{m} \rangle^2}. \tag{7.16}$$

By Theorem 7.7 we know that under conditions of Theorem 6.3, we have $\|\mathbf{m}\| \sim 1$, so the lemma is applicable. Assuming this lemma for the moment, and using that $\mathbf{m}$ is analytic on $\mathbb{H}^N$, we conclude that

$$|\partial_z \operatorname{Im} \mathbf{m}| = \frac{1}{2}|\partial_z \mathbf{m}| \lesssim \frac{1}{\langle \operatorname{Im} \mathbf{m} \rangle^2} \sim \frac{1}{(\operatorname{Im} \mathbf{m})^2},$$

i.e. the derivative of $(\operatorname{Im} \mathbf{m}(z))^3$ is bounded. Thus $z \to \operatorname{Im} \mathbf{m}(z)$ is a 1/3-Hölder regular function on the open upper half plane with a uniform Hölder constant. Therefore $\operatorname{Im} \mathbf{m}(z)$ extends to the real axis as a 1/3-Hölder continuous function. This proves (6.10). Moreover, it is real analytic away from the edges of the self-consistent spectrum $\mathfrak{S} = \{\tau \in \mathbb{R} \, : \, \varrho(\tau) > 0\}$; indeed on $\mathfrak{S}$ it satisfies an analytic ODE with bounded coefficients by (7.16) while outside of the closure of $\mathfrak{S}$ the density is zero. □

Exercise 7.9. *Assume the conditions of Theorem 6.3, i.e.* (6.5) *and that S is piecewise Hölder continuous* (6.8)*. Prove that the saturated self-energy operator has norm 1 on the imaginary axis exactly on the support of the self-consistent density of states, i.e.*

$$\lim_{\eta \to 0+} \|F(E + i\eta)\| = 1 \quad \textit{if and only if} \quad E \in \operatorname{supp}\varrho.$$

Hint: First prove that the Stieltjes transform of a 1/3-Hölder continuous function with compact support is itself 1/3-Hölder continuous up to the real line.

7.4 Bound on the stability operator: proof of Lemma 7.8

The main mechanism for the stability bound (7.16) goes through the operator $F = |\mathbf{m}|S\big(|\mathbf{m}| \cdot\big)$ defined in (7.9). We know that F has a single largest eigenvalue, but in fact under the condition (7.6) this matrix has a substantial gap in its spectrum below the largest eigenvalue. To make this precise, we start with a definition:

Definition 7.10. *For a compact hermitian matrix T the* **spectral gap** *Gap(T) is the difference between the two largest eigenvalues of $|T| = \sqrt{TT^*}$. If $\|T\|_2$ is a degenerate eigenvalue then the gap is zero by definition.*

The following simple lemma shows that matrices with nonnegative entries tend to have a positive gap:

Lemma 7.11. *Let $T = T^*$ have nonnegative entries, $t_{ij} = t_{ji} \geqslant 0$ and let $\mathbf{h}$ be the Perron-Frobenius eigenvector, $T\mathbf{h} = \|T\|\mathbf{h}$ with $\mathbf{h} \geqslant 0$. Then*

$$Gap(T) \geqslant \Big(\frac{\|\mathbf{h}\|_2}{\|\mathbf{h}\|_\infty}\Big) \cdot \min_{ij} t_{ij}.$$

Exercise 7.12. *Prove this lemma. Hint: Set $\|T\| = 1$ and take a vector $\mathbf{u} \perp \mathbf{h}$, $\|\mathbf{u}\|_2 = 1$. Verify that*

$$\langle \mathbf{u}, (1 \pm T)\mathbf{u}\rangle = \frac{1}{2}\sum_{ij} t_{ij}\Big[u_i\Big(\frac{h_j}{h_i}\Big)^{1/2} \pm u_i\Big(\frac{h_i}{h_j}\Big)^{1/2}\Big]^2$$

and estimate it from below.

Applying this lemma to F, we have the following:

Lemma 7.13. *Assume* (7.6) *and let $|z| \leqslant C$. Then F has norm of order one, it has uniform spectral gap;*

$$\|F\| \sim 1, \qquad Gap(F) \sim 1;$$

and its ℓ^2-normalized Perron-Frobenius eigenvector, $\mathbf{f}$ with $F\mathbf{f} = \|F\|\mathbf{f}$, has comparable components

$$\mathbf{f} \sim 1.$$

Proof. We have already seen that $\|F\| \leqslant 1$. The lower bound $\|F\| \gtrsim 1$ follows from $F_{ij} = |m_i|s_{ij}|m_j| \gtrsim 1/N$, in fact $F_{ij} \sim N^{-1}$, thus $\|F\mathbf{1}\|_2 \gtrsim 1$. For the last statement, we write $\mathbf{f} = \|F\|^{-1}F\mathbf{f} \sim F\mathbf{f} \sim \langle \mathbf{f}\rangle$ and then by normalization $1 = \|\mathbf{f}\|_2 \sim \langle \mathbf{f}\rangle \sim \mathbf{f}$. Finally the statement on the gap follows from Lemma 7.11 and that $\|\mathbf{f}\|_\infty \sim \|\mathbf{f}\|_2$. □

Armed with these information on F, we explain how F helps to establish a bound on the stability operator. Using the polar decomposition $\mathbf{m} = e^{i\boldsymbol{\varphi}}|\mathbf{m}|$, we can write for any vector $\mathbf{w}$

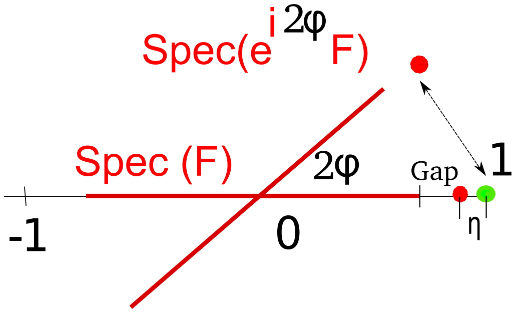
$$(1 - \mathbf{m}^2 S)\mathbf{w} = |\mathbf{m}|\big(1 - e^{2i\boldsymbol{\varphi}}F\big)|\mathbf{m}|^{-1}\mathbf{w}. \tag{7.17}$$

Since $|\mathbf{m}| \sim 1$, it is sufficient to invert $1 - e^{2i\boldsymbol{\varphi}}F$ or $e^{-2i\boldsymbol{\varphi}} - F$. Since F has a real spectrum, this latter matrix should intuitively be invertible unless $\sin 2\boldsymbol{\varphi} \approx 0$. This intuition is indeed correct if $\mathbf{m}$ and thus $e^{2i\boldsymbol{\varphi}}$ were constant; the general case is more complicated.

Assume first that we are in the generalized Wigner case, when $\mathbf{m} = m_{sc} \cdot \mathbf{1}$, i.e. the solution is a constant vector with components $m := m_{sc}$. Writing $m = |m|e^{i\varphi}$ with some phase φ, we see that

$$1 - m^2 S = 1 - e^{2i\varphi}F.$$

Since F is hermitian and has norm bounded by 1, it has spectrum in $[-1, 1]$. So without the phase the inverse of $1 - F$ would be quite singular (basically $\|F\| \approx 1 - c\eta$, see (7.10) at least in the bulk spectrum). The phase $e^{2i\varphi}$ however rotates F out of the real axis, see the picture.



The distance of 1 from the spectrum of F is tiny, but from the spectrum of $e^{2i\varphi}F$ is comparable with $\varphi \sim \operatorname{Im} m = \varrho$:

$$\Big\|\frac{1}{1-m^2S}\Big\| = \Big\|\frac{1}{1-e^{2i\varphi}F}\Big\| \sim \frac{C}{|\varphi|} \sim \frac{C}{\varrho}$$

in the regime where $|\varphi| \leqslant \pi/2$ thanks to the gap in the spectrum of F both below 1 and above -1. In fact this argument indicates a better bound of order $1/\varphi \sim 1/\varrho$ and not only its square in (7.16).

For the general case, when $\mathbf{m}$ is not constant, such a simple argument does not work, since the rotation angles φ_j from $m_j = e^{i\varphi_j}|m_j|$ now depend on the coordinate j, so there is no simple geometric relation between the spectrum of F and that of $\mathbf{m}^2S$. In fact the optimal bound in general is $1/\varrho^2$ and not $1/\varrho$.

To obtain it, we still use the identity

$$(1-\mathbf{m}^2S)\mathbf{w} = e^{2i\boldsymbol{\varphi}}|\mathbf{m}|\big(e^{-2i\boldsymbol{\varphi}} - F\big)|\mathbf{m}|^{-1}\mathbf{w}, \tag{7.18}$$

and focus on inverting $e^{-2i\boldsymbol{\varphi}} - F$. We have the following general lemma:

Lemma 7.14. *Let T be hermitian with $\|T\| \leqslant 1$ and with top normalized eigenvector $\mathbf{f}$, i.e. $T\mathbf{f} = \|T\|\mathbf{f}$. For any unitary operator U we have*

$$\Big\|\frac{1}{U-T}\Big\| \leqslant \frac{C}{\mathit{Gap}(T)\ \cdot \big|1-\|T\|\langle \mathbf{f}, U\mathbf{f}\rangle\big|}. \tag{7.19}$$

A simple calculation shows that this lemma applied to $T = F$ and $U = \big(|\mathbf{m}|/\mathbf{m}\big)^2$ yields the bound C/ϱ^2 for the inverse of $e^{-2i\boldsymbol{\varphi}} - F$ since

$$\big|1-\|T\|\langle \mathbf{f}, U\mathbf{f}\rangle\big| \geqslant \operatorname{Re}\big[1 - \big\langle \frac{\mathbf{m}^2\mathbf{f}^2}{|\mathbf{m}|^2}\big\rangle\big] = 2\big\langle\frac{(\operatorname{Im}\mathbf{m})^2\mathbf{f}^2}{|\mathbf{m}|^2}\big\rangle \sim \langle \operatorname{Im}\mathbf{m}\rangle^2.$$

This proves Lemma 7.8. □

The proof of Lemma 7.14 can be found in Appendix B of [5]. The idea is that one needs a lower bound on $\|(U-T)\mathbf{w}\|$ for any normalized $\mathbf{w}$. Split $\mathbf{w}$ as $\mathbf{w} = \langle \mathbf{f}, \mathbf{w}\rangle \mathbf{f} + P\mathbf{w}$, where P is the orthogonal projection to the complement of $\mathbf{f}$. We will frequently use that

$$\|TP\mathbf{w}\| \leqslant \big[\|T\| - \operatorname{Gap}(T)\big]\|P\mathbf{w}\|, \tag{7.20}$$

following from the definition of the gap. Setting $\alpha := \big|1-\|T\|\langle \mathbf{f}, U\mathbf{f}\rangle\big|$, we distinguish three cases

(i) $16\|P\mathbf{w}\|^2 \geqslant \alpha$;

(ii) $16\|P\mathbf{w}\|^2 < \alpha$ and $\alpha \geqslant \|PU\mathbf{f}\|^2$;

(iii) $16\|P\mathbf{w}\|^2 < \alpha$ and $\alpha < \|PU\mathbf{f}\|^2$.

In regime (i) we use a crude triangle inequality $\|(U-T)\mathbf{w}\| \geqslant \|\mathbf{w}\| - \|T\mathbf{w}\|$, the splitting of w and (7.20). In regime (ii) we first project $(U-T)\mathbf{w}$ onto the $\mathbf{f}$ direction: $\|(U-T)\mathbf{w}\| \geqslant |\langle \mathbf{f}, (1-U^*T)\mathbf{w}\rangle|$ and estimate. Finally in regime (iii) we first project $(U-T)\mathbf{w}$ onto the P direction $\|(U-T)\mathbf{w}\| \geqslant \|P(U-T)\mathbf{w}\|$ and estimate.

Exercise 7.15. *Complete the analysis of all three regimes and finish the proof of Lemma 7.14.*

□

8 Analysis of the matrix Dyson equation

In this section we analyze the matrix Dyson equation introduced in (6.14)

$$I + (z + \mathcal{S}[M])M = 0, \qquad \operatorname{Im} M > 0, \quad \operatorname{Im} z > 0, \qquad (MDE) \tag{8.1}$$

where we assume that $\mathcal{S} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is a symmetric and positivity preserving linear map. In many aspects the analysis goes parallel to that of the vector Dyson equation and we will highlight only the main complications due to the matrix character of this problem.

The proof of the existence and uniqueness result, Theorem 6.4, is analogous to the vector case using the Caratheodory metric, so we omit it, see [53]. The Stieltjes transform representation (6.16) can also be proved by reducing it to the scalar case (Exercise 8.2). The self-consistent density of states is defined as before:

$$\varrho(\mathrm{d}\tau) = \frac{1}{\pi}\langle V(\mathrm{d}\tau)\rangle = \frac{1}{\pi N}\operatorname{Tr} V(\mathrm{d}\tau),$$

and its harmonic extension is again denoted by $\varrho(z) = \langle M(z)\rangle$.

From now on we assume the flatness condition (6.18) on $\mathcal{S}$. We have the analogue of Theorem 7.4 on various bounds on M that can be proven in a similar manner. The role of the ℓ^2-norm, $\|m\|_2$ in the vector case will be played by the Hilbert-Schmidt norm, the natural scalar product structure on matrices. The role of the "absolute value" $|\mathbf{m}|$ in the vector case will be played by the operator norm $\|M\|$ in the matrix case.

Theorem 8.1. *[Bounds on M] Assuming the flatness condition* (6.18)*, we have*

$$\|M\|_{hs} \lesssim 1, \quad \|M(z)\| \lesssim \frac{1}{\varrho(z) + dist(z, supp(\varrho))}, \qquad \|M^{-1}(z)\| \lesssim 1 + |z| \tag{8.2}$$

and

$$\varrho(z) \lesssim \operatorname{Im} M(z) \lesssim (1+|z|)^2 \|M(z)\|^2 \varrho(z) \tag{8.3}$$

where $\|T\|_{hs} := \big(\frac{1}{N}\operatorname{Tr} TT^*\big)^{1/2}$ *is the normalized Hilbert-Schmidt norm.*

Exercise 8.2. *Prove that if $M(z)$ is an analytic matrix-valued function on the upper half plane, $z \in \mathbb{H}$, such that $\operatorname{Im} M(z) > 0$, and $i\eta M(i\eta) \to -I$ as $\eta \to \infty$, then $M(z)$ has a Stieltjes transform representation of the form* (6.16)*. Hint: Reduce the problem to the scalar case by considering the quadratic form $\langle \mathbf{w}, M(z)\mathbf{w}\rangle$ for $\mathbf{w} \in \mathbb{C}$.*

Exercise 8.3. *Prove Theorem 8.1 by mimicking the corresponding proof for the vector case but watching out for the non commutativity of the matrices.*

8.1 The saturated self-energy matrix

We have seen in the vector Dyson equation that the stability operator $(1 - \mathbf{m}^2 S)^{-1}$ played a central role both in establishing regularity of the self-consistent density of states and also in establishing the local law. What is the matrix analogue of this operator? Is there any analogue for the saturated self-energy operator F defined in Definition 7.5 ?

The matrix responsible for the stability can be easily found, mimicking the calculation (7.15) by differentiating (8.1) wrt. z

$$I + (z + \mathcal{S}[M])M = 0 \quad \Longrightarrow \quad (I + \mathcal{S}[\partial_z M])M + (z + \mathcal{S}[M])\partial_z M = 0 \quad \Longrightarrow \quad \partial_z M = (1 - M\mathcal{S}[\cdot]M)^{-1}M^2. \tag{8.4}$$

where we took the inverse of the "super operator" $1 - M\mathcal{S}[\cdot]M$. We introduce the notation $\mathcal{C}_T$ for the operator **"sandwiching by a matrix** T**"**, that acts on any matrix R as

$$\mathcal{C}_T[R] := TRT.$$

With this notation we have $1 - M\mathcal{S}[\cdot]M = 1 - \mathcal{C}_M\mathcal{S}$ that acts on $N \times N$ matrices as $(1 - \mathcal{C}_M\mathcal{S})[R] = R - M\mathcal{S}(R)M$.

The boundedness of the stability operator, the inverse of $1 - \mathbf{m}^2S$ in the vector case, relied crucially on finding a symmetrized version of the operator $\mathbf{m}^2S$, the saturated self-energy operator (Definition 7.9), for which spectral theory can be applied, see the identity (7.18). This will be the heart of the proof in the following section where we control the spectral norm of the stability operator. Note that spectral theory in the matrix setup means to work with the Hilbert space of matrices, equipped with the Hilbert-Schmidt scalar product. We denote by $\|\cdot\|_{sp} := \|\cdot\|_{hs\to hs}$ the corresponding norm of superoperators viewed as linear maps on this Hilbert space.

8.2 Bound on the stability operator

The key technical result of the analysis of the MDE is the following lemma:

Lemma 8.4. *Assuming the flatness condition* (6.18)*, we have, for* $|z| \leqslant C$,

$$\Big\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\Big\|_{sp} \lesssim \frac{1}{\big[\varrho(z) + dist(z, supp(\varrho))\big]^C} \tag{8.5}$$

with some universal constant ($C = 100$ would do).

Similarly to the argument in Section 7.3 for the vector case, the bound (8.5) directly implies Hölder regularity of the solution and it implies (6.19). It is also the key estimate in the random matrix part of the proof of the local law.

Proof of Lemma 8.4. In the vector case, the saturated self-energy matrix F naturally emerged from taking the imaginary part of the Dyson equation and recognizing a Perron-Frobenius type eigenvector of the form $\frac{\operatorname{Im}\mathbf{m}}{|\mathbf{m}|}$, see (7.12). This structure was essential to establish the bound $\|F\| \leqslant 1$. We proceed similarly for the matrix case to find the analogous super operator $\mathcal{F}$ that has to be symmetric and positivity preserving in addition to having a "useful" Perron-Frobenius eigenequation. The imaginary part of the MDE in the form

$$-\frac{1}{M} = z + \mathcal{S}[M]$$

is given by

$$\frac{1}{M^*}\operatorname{Im} M\frac{1}{M} = \eta + \mathcal{S}[\operatorname{Im} M], \quad \Longrightarrow \quad \operatorname{Im} M = \eta M^*M + M^*\mathcal{S}[\operatorname{Im} M]M. \tag{8.6}$$

What is the analogue of $\frac{\operatorname{Im}\mathbf{m}}{|\mathbf{m}|}$ in this equation that is positive, but this time as a matrix? "Dividing by $|M|$" is a quite ambiguous operation, not just because the matrix multiplication is not commutative, but also for the fact that for non-normal matrices, the absolute value of a general matrix R is not defined in a canonical way. The standard definition is $|R| = \sqrt{R^*R}$, which leads to the polar decomposition of the usual form $R = U|R|$ with some unitary U, but the alternative definition $\sqrt{RR^*}$ would also be equally justified. Just they are not the same, and this ambiguity would destroy the symmetry of the attempted super operator $\mathcal{F}$ if done naively.

Instead of guessing the right form, we just look for the matrix version of $\frac{\operatorname{Im}\mathbf{m}}{|\mathbf{m}|}$ in the form $\frac{1}{Q}\operatorname{Im} M\frac{1}{Q^*}$ with some matrix Q yet to be found. Then we can rewrite (8.6) (for $\eta = 0$ for simplicity) as

$$\frac{1}{Q}\operatorname{Im} M\frac{1}{Q^*} = \frac{1}{Q}M^*\frac{1}{Q^*}Q^*\mathcal{S}\Big[Q\,\frac{1}{Q}(\operatorname{Im} M)\frac{1}{Q^*}\,Q^*\Big]Q\frac{1}{Q}M\frac{1}{Q^*}.$$

We write it in the form

$$X = Y^*Q^*\mathcal{S}[QXQ^*]QY, \quad \text{with} \quad X := \frac{1}{Q}\,(\operatorname{Im} M)\frac{1}{Q^*}, \quad Y := \frac{1}{Q}M\frac{1}{Q^*}$$

i.e.

$$X = Y^* \mathcal{F}[X] Y, \qquad \text{with} \quad \mathcal{F}[\cdot] := Q^* \mathcal{S}[Q \cdot Q^*] Q.$$

With an appropriate Q, this operator will be the correct saturated self-energy operator.

To get the Perron-Frobenius structure, we need to "get rid" of the Y and Y^* in this equation; we have a good chance if we require that Y be unitary, $YY^* = Y^*Y = I$. The good news is that $X = \operatorname{Im} Y$ and if Y is unitary, then X and Y commute (check this fact). We thus arrive at

$$X = \mathcal{F}[X].$$

Thus the Perron-Frobenius argument applies and we get that $\mathcal{F}$ is bounded in spectral norm:

$$\|\mathcal{F}\|_{sp} \leqslant 1$$

Actually, if $\eta > 0$, then we get a strict inequality.

Using the definition of $\mathcal{F}$ and that $M = QYQ^*$ with some unitary Y, we can also write the operator $\mathcal{C}_M \mathcal{S}$ appearing in the stability operator in terms of $\mathcal{F}$. Indeed, for any matrix R

$$M\mathcal{S}[R]M = QYQ^* \mathcal{S}\Big[Q \frac{1}{Q} R \frac{1}{Q^*} Q^*\Big] QYQ^* = QY\mathcal{F}\Big[\frac{1}{Q} R \frac{1}{Q^*}\Big] YQ^*$$

so

$$R - M\mathcal{S}[R]M = Q\Big(1 - Y\mathcal{F}[\cdot]Y\Big)\big[\frac{1}{Q} R \frac{1}{Q^*}\big] Q^*.$$

Thus

$$I - \mathcal{C}_M \mathcal{S} = \mathcal{K}_Q (I - \mathcal{C}_Y \mathcal{F}) \mathcal{K}_Q^{-1}, \tag{8.7}$$

where for any matrix T we defined the super operator $\mathcal{K}_T$ acting on any matrix R as $\mathcal{K}_T[R] := TRT^*$ to be the symmetrized analogue of the sandwiching operator $\mathcal{C}_T$. The formula (8.7) is the matrix analogue of (7.17).

Thus, assuming that $Q \sim 1$ in a sense that $\|Q\| \lesssim 1$ and $\|Q^{-1}\| \lesssim 1$, we have

$$I - \mathcal{C}_M \mathcal{S} \text{ is stable} \iff I - \mathcal{C}_Y \mathcal{F} \text{ is stable} \iff \mathcal{C}_{Y^*} - \mathcal{F} \text{ is stable}$$

and this would bring our stability operator into the form of a "unitary minus bounded self-adjoint" to which Lemma 7.14 (in the Hilbert space of matrices) will apply.

To complete this argument, all we need is a "symmetric polar decomposition" of M in the form $M = QYQ^*$, where Y is unitary and $Q \sim 1$ knowing that $M \sim 1$. We will give this decomposition explicitly. Write $M = A + iB$ with $A := \operatorname{Re} M$ and $B := \operatorname{Im} M > 0$. Then we can write

$$M = \sqrt{B}\Big(\frac{1}{\sqrt{B}} A \frac{1}{\sqrt{B}} + i\Big)\sqrt{B}$$

and now we make the middle factor unitary by dividing its absolute value:

$$M = \sqrt{B} W Y W \sqrt{B} =: QYQ^*$$

$$W := \left[1 + \Big(\frac{1}{\sqrt{B}} A \frac{1}{\sqrt{B}}\Big)^2\right]^{\frac{1}{4}}, \quad Y := \frac{\frac{1}{\sqrt{B}} A \frac{1}{\sqrt{B}} + i}{W^2}.$$

In the regime, where $c \leqslant B \leqslant C$ and $\|A\| \leqslant C$, we have

$$Q = \sqrt{B} W \sim 1$$

in the sense that $\|Q\| \lesssim 1$ and $\|Q^{-1}\| \lesssim 1$. In our application, we use the upper bound (8.2) for $\|M\|$ and the lower bound on $B = \operatorname{Im} M$ from (8.3). This gives a control on both $\|Q\|$ and $\|Q^{-1}\|$ as a certain power of $\varrho(z)$ and this will be responsible for parts of the powers collected in the right hand side of (8.5). In this proof here we focus only on the bulk, so we do not intend to gain the additional term $\mathrm{dist}(z, \mathrm{supp}\varrho)$ that requires a slightly different argument. The result is

$$\Big\|\frac{1}{1-\mathcal{C}_M\mathcal{S}}\Big\|_{sp} \leqslant \frac{1}{\varrho(z)^C}\Big\|\frac{1}{\mathcal{U}-\mathcal{F}}\Big\|_{sp} \tag{8.8}$$

with $\mathcal{U} := \mathcal{C}_{Y^*}$.

We remark that $\mathcal{F}$ can also be written as follows:

$$\mathcal{F} = \mathcal{K}_Q^*\mathcal{S}\mathcal{K}_Q = \mathcal{C}_W\mathcal{C}_{\sqrt{\mathrm{Im}M}}\mathcal{S}\mathcal{C}_{\sqrt{\mathrm{Im}M}}\mathcal{C}_W. \tag{8.9}$$

Finally, we need to invert $\mathcal{C}_{Y^*} - \mathcal{F}$ effectively with the help of Lemma 7.14. The Perron-Frobenius type theorem (called the Krein-Rutman theorem in more general Banach spaces) applied to $\mathcal{F}$ yields that it has a normalized *eigenmatrix* F with eigenvalue $\|\mathcal{F}\|_{sp} \leqslant 1$. The following lemma collects information on $\mathcal{F}$ and F, similarly to Lemma 7.13:

Lemma 8.5. *Assume the flatness condition* (6.18) *and let* $\mathcal{F}$ *be defined by* (8.9). *Then* $\mathcal{F}$ *has a unique normalized eigenmatrix corresponding to its largest eigenvalue*

$$\mathcal{F}[F] = \|\mathcal{F}\|_{sp}F, \qquad \|F\|_{hs} = 1, \qquad \|\mathcal{F}\|_{sp} \leqslant 1.$$

Furthermore

$$\|\mathcal{F}\|_{sc} = 1 - \frac{\langle F, \mathcal{C}_W[\mathrm{Im}\, M]\rangle}{\langle F, W^{-2}\rangle}\,\mathrm{Im}\, z,$$

the eigenmatrix F *has bounds*

$$\frac{1}{\|M\|^7} \leqslant F \leqslant \|M\|^6$$

and $\mathcal{F}$ *has a spectral gap:*

$$Spec\big(\mathcal{F}/\|\mathcal{F}\|_{sc}\big) \subset [-1+\theta, 1-\theta] \cup \{1\}, \qquad \theta \geqslant \|M\|^{-42} \tag{8.10}$$

(the explicit powers do not play any significant role).

We omit the proof of this lemma (see Lemma 4.6 of [7]), its proof is similar but more involved than that of Lemma 7.13, especially the noncommutative analogue of Lemma 7.11 needs substantial changes (this is given in Lemma A.3 in [7]).

Armed with the bounds on $\mathcal{F}$ and F, we can use Lemma 7.14 with T playing the role of $\mathcal{F}$ and $\mathcal{U} := \mathcal{C}_{Y^*}$ playing the role of U:

$$\Big\|\frac{1}{\mathcal{U}-\mathcal{F}}\Big\|_{sp} \lesssim \frac{1}{\mathrm{Gap}(\mathcal{F})\ \big|1 - \|\mathcal{F}\|\langle F, \mathcal{U}(F)\rangle\big|}.$$

We already had a bound on the gap of $\mathcal{F}$ in (8.10). As a last step, we prove the estimate

$$\big|1 - \langle F, \mathcal{U}(F)\rangle\big| = \big|1 - \langle F, Y^*FY^*\rangle\big| \geqslant \frac{\varrho^2(z)}{\|M\|^4} \geqslant \varrho(z)^6.$$

Exercise 8.6. *Prove this last bound by using* $\big|1 - \langle F, Y^*FY^*\rangle\big| \geqslant \langle F, \mathcal{C}_{\mathrm{Im}\, Y^*}F\rangle$*, using the definition of* Y *and various bounds on* M *from Theorem 8.1.*

Combining (8.8) with these last bounds and with the bound on the gap of $\mathcal{F}$ (8.10) we complete the proof of Lemma 8.4 (without the $\mathrm{dist}(z, \mathrm{supp}\varrho)$ part). □

9 Ideas of the proof of the local laws

In this section we sketch the proof of the local laws. We will present the more general correlated case, i.e. Theorem 6.7 and we will focus on the entrywise local law (6.30). In Section 3 around (3.19) we already outlined the main idea. Starting from $HG = I + zG$, we have the identity

$$I + (z + \mathcal{S}[G])G = D, \qquad D := HG + \mathcal{S}[G]G, \tag{9.1}$$

and we compare it with the matrix Dyson equation

$$I + (z + \mathcal{S}[M])M = 0. \tag{9.2}$$

The first (probabilistic) part of the proof is a good bound on D, the second (deterministic) part is to use the stability of the MDE to conclude from these two equations that $G - M$ is small. The first question is in which norm should one estimate these quantities?

Note that D is still random and it is not realistic to estimate it in operator norm, in fact $\|D\| \gtrsim 1/\eta$ with high probability. To see this, consider the simplest Wigner case, then

$$D = I + \Big(z + \frac{1}{N}\operatorname{Tr} G\Big)G.$$

Let λ be the closest eigenvalue to $\operatorname{Re} z$ with normalized eigenvector $\mathbf{u}$. Note that typically $|\operatorname{Re} z - \lambda| \lesssim 1/N$ and $\eta \gg 1/N$, thus $\|G\mathbf{u}\| = 1/|\lambda - z| \sim 1/\eta$ (suppose that $\operatorname{Re} z$ is away from zero). From the local law we know that $\frac{1}{N}\operatorname{Tr} G \sim m_{sc} \sim 1$ and $z + m_{sc} \sim 1$. Thus

$$\|D\mathbf{u}\| = \Big\|\mathbf{u} + \Big(z + \frac{1}{N}\operatorname{Tr} G\Big)G\mathbf{u}\Big\| \sim \|G\mathbf{u}\| \sim 1/\eta.$$

The appropriate weaker norm is the entrywise maximum norm defined for any matrix T as

$$\|T\|_{\max} := \max_{ij} |T_{ij}|.$$

In this norm we have the following

Theorem 9.1. *Under the conditions of Theorem 6.7, for any $\gamma, \varepsilon, D > 0$ we have the following high probability statement:*

$$\mathbb{P}\Big(\|D(z)\|_{\max} \geqslant \frac{N^{\varepsilon}}{\sqrt{N\eta}} \ : \textit{ for some } z = E + i\eta \textit{ with } |z| \leqslant 1000,\ \eta \geqslant N^{-1+\gamma}\Big) \leqslant \frac{C}{N^D}, \tag{9.3}$$

i.e. all matrix elements D_{ij} are small simultaneously for all spectral parameters.

We will omit the proof, which is a tedious calculation and whose basic ingredients were sketched in Section 3. For the Wigner type matrices or for correlated matrices with fast (exponential) correlation decay as in [7] one may use the Schur complement method together with concentration estimates on quadratic functionals of independent or essentially independent random vectors (Section 3.1.1). For more general correlations or if nonzero exception of H is allowed, then we may use the cumulant method (Section 3.1.2). In both cases, one establishes a high moment bound on $\mathbb{E}|D_{ij}|^p$ via a detailed expansion and then one concludes a high probability bound via Markov inequality.

In the second (deterministic) part of the proof we compare (9.1) and (9.2). From these two equations we have

$$(I - M\mathcal{S}[\cdot]M)[G - M] = MD + M\mathcal{S}[G - M](G - M), \tag{9.4}$$

so by inverting the super operator $I - M\mathcal{S}[\cdot]M = I - \mathcal{C}_M\mathcal{S}$, we get

$$G - M = \frac{1}{I - \mathcal{C}_M\mathcal{S}}\big[MD\big] + \frac{1}{I - \mathcal{C}_M\mathcal{S}}\Big[M\mathcal{S}[G - M](G - M)\Big]. \tag{9.5}$$

We will need the additional information that not only $\|M\|$ is bounded, see (8.2), but also

$$\|M\|_\infty := \max_i \sum_j |M_{ij}| \qquad \text{and} \quad \|M\|_1 := \max_j \sum_i |M_{ij}| \tag{9.6}$$

are bounded. This information is obvious for Wigner type matrices, when M is diagonal. For correlated matrices with fast correlation decay it requires a somewhat involved additional proof that we do not repeat here, see Theorem 2.5 of [7]. Slow decay needs another argument [42].

Furthermore, we know that in the bulk spectrum the stability operator is bounded in spectral norm (8.5), i.e. when the stability operator is considered mapping matrices with Hilbert Schmidt norm. We may also consider its norm in the other two natural norms, i.e. when the space of matrices is equipped with the maximum norm (9.6) and the Euclidean matrix norm $\|\cdot\|$. We find the boundedness of the stability operator in these two other norms as well since we can prove (see Exercise 9.2)

$$\Big\|\frac{1}{I-\mathcal{C}_M\mathcal{S}}\Big\|_{\infty\to\infty} + \Big\|\frac{1}{I-\mathcal{C}_M\mathcal{S}}\Big\|_{\|\cdot\|\to\|\cdot\|} \lesssim \Big\|\frac{1}{I-\mathcal{C}_M\mathcal{S}}\Big\|_{sp}. \tag{9.7}$$

Using all these information, we easily obtain from (9.5) that

$$\|G-M\|_{\max} \lesssim \|D\|_{\max} + \|G-M\|_{\max}^2,$$

where $\lesssim$ includes factors to $\varrho(z)^{-C}$, which are harmless in the bulk. From this quadratic inequality we easily obtain that

$$\|G-M\|_{\max} \lesssim \|D\|_{\max}, \tag{9.8}$$

assuming a weak bound $\|G-M\|_{\max} \ll 1$. This latter information is obtained by a continuity argument in the imaginary part of the spectral parameter. We fix an E in the bulk, $\varrho(E) > 0$ and consider $(G-M)(E+i\eta)$ as a function of η. For large η we know that both G and M are bounded by $1/\eta$, hence they are small, so the weak bound $\|G-M\|_{\max} \ll 1$ holds. Then we conclude that (9.8) holds for large η. Since $\|D\|_{\max}$ is small, at least with very high probability, see (9.3), we obtain that the strong bound

$$\|G-M\|_{\max} \lesssim \frac{N^\varepsilon}{\sqrt{N\eta}} \tag{9.9}$$

also holds. Now we may reduce the value of η a bit since the function $\eta \to (G-M)(E+i\eta)$ is Lipschitz continuous with Lipschitz constant C/η^2. So we know that $\|G-M\|_{\max} \ll 1$ for this smaller η value as well. Thus (9.8) can again be applied and together with (9.3) we get the strong bound (9.9) for this reduced η as well. We continue this "small-step" reduction as long as the strong bound implies the weak bound, i.e. as long as $N\eta \gg N^{2\varepsilon}$, i.e. $\eta \gg N^{-1+2\varepsilon}$. Since $\varepsilon > 0$ is arbitrary we can go down to the scales $\eta \geqslant N^{-1+\gamma}$ for any $\gamma > 0$. Some care is needed in this argument, since the smallness of $\|D\|_{\max}$ holds only with high probability, so in every step we lose a set of small probability. This is, however, affordable by the union bound since the probability of the events where D is not controlled is very small, see (9.3).

The proof of the averaged law (6.31) is similar. Instead of the maximum norm, we use averaged quantities of the form $\langle TD\rangle = \frac{1}{N}\operatorname{Tr} TD$. In the first, probabilistic step instead of (9.3) we prove that for any fixed deterministic matrix T we have

$$\big|\langle TD\rangle\big| \leqslant \frac{N^\varepsilon}{N\eta}\|T\|$$

with very high probability. Notice that averaged quantities can be estimated with an additional $(N\eta)^{-1/2}$ power better; this is the main reason why averaged law (6.31) has a stronger control than the entrywise or the isotropic laws.

Exercise 9.2. *Prove* (9.7)*. Hint: use the identity*

$$\frac{1}{I-\mathcal{C}_M\mathcal{S}} = I + \mathcal{C}_M\mathcal{S} + \mathcal{C}_M\mathcal{S}\frac{1}{I-\mathcal{C}_M\mathcal{S}}\mathcal{C}_M\mathcal{S}$$

and the smoothing properties of the self-energy operation $\mathcal{S}$ following from (6.18) *and the boundedness of M in all three relevant norms.*

References

[1] E. Abrahams, P.W. Anderson, D.C. Licciardello, and T.V. Ramakrishnan. Scaling theory of localization: Absence of quantum diffusion in two dimensions. *Phys. Rev. Lett.*, 42.

[2] Amol Aggarwal. Bulk universality for generalized wigner matrices with a few moments. *Preprint Arxiv 1612.00421*, 2016.

[3] M. Aizenman and S. Molchanov. Localization at large disorder and at extreme energies: an elementary derivation. *Comm. Math. Phys.*, 157(2):245–278, 1993.

[4] O. Ajanki, L. Erdős, and T. Krüger. Quadratic vector equations on complex upper half-plane. *arXiv:1506.05095*, June 2015.

[5] Oskari Ajanki, László Erdős, and Torben Krüger. Singularities of solutions to quadratic vector equations on complex upper half-plane. *arxiv:1512.03703.*

[6] Oskari Ajanki, László Erdős, and Torben Krüger. Universality for general Wigner-type matrices. *arXiv:1506.05098.*

[7] Oskari Ajanki, László Erdős, and Torben Krüger. Stability of the matrix Dyson equation and random matrices with correlations. *arXiv:1604.08188*, 2016.

[8] G. Akemann, J. Baik, and P. Di Francesco, editors. *The Oxford handbook of random matrix theory.* Oxford University Press, Oxford, 2011.

[9] Ariel Amir, Naomichi Hatano, and David R. Nelson. Non-Hermitian localization in biological networks. *Phys. Rev. E*, 93:042310, 2016.

[10] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 2010.

[11] P. W. Anderson. Absence of diffusion in certain random lattices. *Phys. Rev.*, 109:1492–1505, Mar 1958.

[12] Z. Bao and L. Erdős. Delocalization for a class of random block band matrices. *Probab. Theory Related Fields*, pages 1–104, 2016.

[13] R. Bauerschmidt, J. Huang, A. Knowles, and H.-T. Yau. Bulk eigenvalue statistics for random regular graphs. *arXiv:1505.06700*, May 2015.

[14] R. Bauerschmidt, A. Knowles, and H.-T. Yau. Local semicircle law for random regular graphs. *arXiv:1503.08702*, March 2015.

[15] F. Bekerman, A. Figalli, and A. Guionnet. Transport maps for β-matrix models and universality. *Comm. Math. Phys.*, 338(2):589–619, 2015.

[16] Florent Benaych-Georges and Sandrine Péche. Localization and delocalization for heavy tailed band matrices. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques, Institute Henri Poincaré*, 50(4):1385–1403, 2014.

[17] P. Bleher and A. Its. Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model. *Ann. of Math. (2)*, 150(1):185–266, 1999.

[18] Charles Bordenave and Alice Guionnet. Delocalization at small energy for heavy-tailed random matrices. *Preprint Arxiv:1603.08845*, 2016.

[19] P. Bourgade, L. Erdős, and H.-T. Yau. Bulk universality of general β-ensembles with non-convex potential. *J. Math. Phys.*, 53(9):095221, 19, 2012.

[20] P. Bourgade, L. Erdős, and H.-T. Yau. Edge universality of beta ensembles. *Comm. Math. Phys.*, 332(1):261–353, 2014.

[21] P. Bourgade, L. Erdős, and H.-T. Yau. Universality of general β-ensembles. *Duke Math. J.*, 163(6):1127–1190, 2014.

[22] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin. Fixed energy universality for generalized Wigner matrices. *Comm. Pure Appl. Math.*, pages 1–67, 2015.

[23] E. Brézin and S. Hikami. Correlations of nearby levels induced by a random potential. *Nuclear Phys. B*, 479(3):697–706, 1996.

[24] E. Brézin and S. Hikami. Spectral form factor in a random matrix theory. *Phys. Rev. E*, 55(4):4067–4083, 1997.

[25] E. B. Davies. The functional calculus. *J. London Math. Soc.*, 52(1):166–176, 1995.

[26] P. Deift. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, volume 3 of *Courant Lecture Notes in Mathematics*. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1999.

[27] P. Deift and D. Gioev. Universality at the edge of the spectrum for unitary, orthogonal, and symplectic ensembles of random matrices. *Comm. Pure Appl. Math.*, 60(6):867–910, 2007.

[28] P. Deift, T. Kriecherbauer, K. T-R McLaughlin, S. Venakides, and X. Zhou. Strong asymptotics of orthogonal polynomials with respect to exponential weights. *Comm. Pure Appl. Math.*, 52(12):1491–1552, 1999.

[29] F. J. Dyson. A Brownian-motion model for the eigenvalues of a random matrix. *J. Math. Phys.*, 3:1191–1198, 1962.

[30] L. Erdős and A. Knowles. Quantum diffusion and delocalization for band matrices with general distribution. *Ann. Henri Poincaré*, 12(7):1227–1319, 2011.

[31] L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14(8):1837–1926, 2013.

[32] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Delocalization and diffusion profile for random band matrices. *Comm. Math. Phys.*, 323(1):367–416, 2013.

[33] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Spectral statistics of Erdős-Rényi graphs I: Local semicircle law. *Ann. Probab.*, 41(3B):2279–2375, 2013.

[34] L. Erdős, B. Schlein, and H.-T. Yau. Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Not.*, 2010(3):436–479, 2010.

[35] L. Erdős, B. Schlein, and H.-T. Yau. Universality of random matrices and local relaxation flow. *Invent. Math.*, 185(1):75–119, 2011.

[36] L. Erdős, B. Schlein, H.-T. Yau, and J. Yin. The local relaxation flow approach to universality of the local statistics for random matrices. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(1):1–46, 2012.

[37] L. Erdős and K. Schnelli. Universality for Random Matrix Flows with Time-dependent Density. *arXiv:1504.00650*, April 2015.

[38] L. Erdős and H.-T. Yau. Gap universality of generalized Wigner and β-ensembles. *J. Eur. Math. Soc. (JEMS)*, 17(8):1927–2036, 2015.

[39] L. Erdős and H.-T. Yau. *Dynamical Approach to Random Matrix Theory*, volume 28. Courant Lecture Notes in Mathematics, 2017.

[40] László Erdős, A Knowles, Horng-Tzer Yau, and J Yin. Spectral statistics of Erdős-Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues. *Commun. Math. Phys.*, 314(3):587–640, 2012.

[41] László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Delocalization and Diffusion Profile for Random Band Matrices. *Commun. Math. Phys.*, 323:367–416, 2013.

[42] László Erdős, Torben Krüger, and Dominik Schröder. Random matrices with slow correlation decay. *Preprint Arxiv:1705.10661*, 2017.

[43] László Erdős and H.-T. Yau. Universality of local spectral statistics of random matrices. *Bull. Amer. Math. Soc*, 49:377–414, 2012.

[44] László Erdős, Horng-Tzer Yau, and Jun Yin. Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields*, 154(1-2):341–407, 2011.

[45] A. S. Fokas, A. R. Its, and A. V. Kitaev. The isomonodromy approach to matrix models in 2D quantum gravity. *Comm. Math. Phys.*, 147(2):395–430, 1992.

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

[47] J. Fröhlich and T. Spencer. Absence of diffusion in the Anderson tight binding model for large disorder or low energy. *Comm. Math. Phys.*, 88(2):151–184, 1983.

[48] Y. V. Fyodorov and A. D. Mirlin. Scaling properties of localization in random band matrices: a σ-model approach. *Phys. Rev. Lett.*, 67(18):2405–2409, 1991.

[49] J. Geronimo and T Hill. Necessary and sufficient condition that the limit of stieltjes transforms is a stieltjes transform. *J. Approx. Theory*, 121:54–60, 2003.

[50] I. Ya. Goldsheid, S. A. Molchanov, and L. A. Pastur. A pure point spectrum of the stochastic one-dimensional schrödinger equation. *Funkt. Anal. Appl.*, 11:1–10, 1977.

[51] F. Götze, A. Naumov, and A. Tikhomirov. Local semicircle law under moment conditions. Part I: The Stieltjes transform. *arXiv:1510.07350*, October 2015.

[52] Yukun He, Antti Knowles, and Ron Rosenthal. Isotropic self-consistent equations for mean-field random matrices. *Preprint Arxiv 1611.05364*, 2016.

[53] J. William Helton, Reza Rashidi Far, and Roland Speicher. Operator-valued Semicircular Elements: Solving A Quadratic Matrix Equation with Positivity Constraints. *Internat. Math. Res. Notices*, 2007, 2007.

[54] J. Huang, B. Landon, and H.-T. Yau. Bulk universality of sparse random matrices. *J. Math. Phys.*, 56(12):123301, 19, 2015.

[55] C. Itzykson and J. B. Zuber. The planar approximation. II. *J. Math. Phys.*, 21(3):411–421, 1980.

[56] K. Johansson. Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. *Comm. Math. Phys.*, 215(3):683–705, 2001.

[57] K. Johansson. Universality for certain Hermitian Wigner matrices under weak moment conditions. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(1):47–79, 2012.

[58] A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.*, 66(11):1663–1750, 2013.

[59] A. Knowles and J. Yin. Anisotropic local laws for random matrices. *arXiv:1410.3516*, October 2014.

[60] B. Landon, Philippe Sosoe, and H.-T. Yau. Fixed energy universality of Dyson Brownian motion. *arXiv:1609.09011*, 2016.

[61] Benjamin Landon and Horng-Tzer Yau. Convergence of local statistics of Dyson Brownian motion. *arXiv:1504.03605.*

[62] J. O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau. Bulk universality for deformed Wigner matrices. *Ann. Probab.*, 44(3):2349–2425, 2016.

[63] M. L. Mehta. *Random matrices.* Academic Press, Inc., Boston, MA, second edition, 1991.

[64] Sean O'Rourke and Van Vu. Universality of local eigenvalue statistics in random matrices with external source. *Random Matrices: Theory and Applications*, 03(02), 2014.

[65] L. Pastur and M. Shcherbina. *Eigenvalue distribution of large random matrices*, volume 171 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2011.

[66] J. Schenker. Eigenvector localization for random band matrices with power law band width. *Comm. Math. Phys.*, 290(3):1065–1097, 2009.

[67] M. Shcherbina. Change of variables as a method to study general β-models: bulk universality. *J. Math. Phys.*, 55(4):043504, 23, 2014.

[68] T. Shcherbina. On the second mixed moment of the characteristic polynomials of 1D band matrices. *Comm. Math. Phys.*, 328(1):45–82, 2014.

[69] Tatyana Shcherbina. On the second mixed moment of the characteristic polynomials of 1D band matrices. *Commun. Math. Phys.*, 328(1):45–82, 2014.

[70] Tatyana Shcherbina. Universality of the local regime for the block band matrices with a finite number of blocks. *J. Stat. Phys.*, 155(3):466–499, 2014.

[71] S. Sodin. The spectral edge of some random band matrices. *Ann. of Math. (2)*, 172(3):2223–2251, 2010.

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

[73] T. Tao and V. Vu. Random matrices: universality of local eigenvalue statistics. *Acta Math.*, 206(1):127–204, 2011.

[74] D. Vollhardt and P. Wölfle. Diagrammatic, self-consistent treatment of the anderson localization problem in $d \leqslant 2$ dimensions. *Phys. Rev. B*, 22:4666–4679, 1980.

[75] J Weidmann. *Linear Operators in Hilbert Spaces.* Springer Verlag, New York, 1980.

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

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