Probability Surveys
Vol. 20 (2023) 291–381
ISSN: 1549-5787
https://doi.org/10.1214/23-PS16

Proof methods in random matrix theory

Michael Fleermann and Werner Kirsch

FernUniversität in Hagen,
Fakultät für Mathematik und Informatik,
Universitätsstraße 1,
58097 Hagen, Germany
e-mail: michael.fleermann@fernuni-hagen.de; werner.kirsch@fernuni-hagen.de

Abstract: In this survey article, we give an introduction to two methods of proof in random matrix theory: The method of moments and the Stieltjes transform method. We thoroughly develop these methods and apply them to show both the semicircle law and the Marchenko-Pastur law for random matrices with independent entries. The material is presented in a pedagogical manner and is suitable for anyone who has followed a course in measure-theoretic probability theory.

MSC2020 subject classifications: 60B20.
Keywords and phrases: Random matrix theory, method of moments, Stieltjes transform method, semicircle law, Marchenko-Pastur law.

Received April 2022.

Contents

arXiv: 2203.02551

1. Introduction

1.1. Intention and scope of this text

The goal of this article is to give a digestible yet concise introduction to random matrix theory. We focus on the tools and concepts that allow us to comprehend the results which marked the very beginnings of this theory: The semicircle law discovered in [58, 59] and the Marchenko-Pastur law established in [37]. These are statements pertaining to probabilistic weak convergence – namely weak convergence in expectation resp. in probability resp. almost surely – which is a framework also encountered in probability theory when studying the Glivenko-Cantelli theorem, for example. We thoroughly investigate the subtleties of probabilistic weak convergence in Chapter 2 of this text.

Statements about weak convergence – such as the central limit theorem – may be proved in numerous ways, two of them being the analysis of the moments of the distributions or the analysis of certain transforms of the distributions involved. Concerning the proof of the central limit theorem, see Chapter 30 in [10] for the use of moments, and Chapter 27 in [10] for the use of transforms. When studying statements of probabilistic weak convergence in random matrix theory, it turns out that again, moments and transforms can be employed with great success and in numerous settings. Therefore, we carefully develop the method of moments in Chapter 3 and the Stieltjes transform method in Chapter 5. We employ these methods to show both the semicircle law and the Marchenko-Pastur law in Chapters 4 and 6.

During the past decades, random matrix theory has evolved into a huge field of study. Both the results and the techniques to derive them have become rather sophisticated, making an entry into this field cumbersome. This text aims to alleviate this barrier of entry and can be followed after completing a basic course of measure-theoretic probability theory. It is based on the works [25, 26] of the first author, but has also benefitted greatly from the research endeavors of both authors. Further, the techniques presented are employed in many contemporary research articles and are thus highly relevant for researchers aiming to contribute to random matrix theory.

1.2. Cornerstones, recent developments and further reading

Before we begin with the development of the theory in Chapter 2 of this text, we would like to give pointers to other monographs and lecture notes on random matrices, and sketch a picture of recent and important developments in the last decades. The area of research on random matrices is vast. It is the goal here to discuss certain cornerstones of the theory.

To begin with, there exists a number of seminal monographs on random matrices, including [4, 5, 38, 41, 53] for general expositions, [22, 39] for expositions with an emphasis on specialized areas within random matrix theory, and [42] with an emphasis on the applications. Valuable lecture notes and surveys on random matrices are given by [33, 51] for general expositions and [13, 7] with an emphasis on specialized areas.

The general interest of study concerns *the spectrum* of random matrices, to be precise, the location of their eigenvalues and the structure of their eigenvectors. Herewith connected is the question of their invertibility, see [56]. Three main models of random matrices regularly studied are self-adjoint random matrices, sample covariance matrices and non-selfadjoint random matrices. Their respective special cases with Gaussian entries are called Gaussian Orthogonal/Unitary Ensemble (GOE/GUE), Wishart matrix and Ginibre ensemble. The restriction to the Gaussian case has many advantages. For example, there exists an explicit expression for the joint density of the eigenvalues (see [51] for details in all three cases). In these models, classical results on the behavior of the spectrum were obtained under the assumption that entries be standardized and independent, the latter up the structural constraint imposed by the matrix model. For the self-adjoint case, this led to the development of the semicircle law, pioneered by [58, 59], employing the method of moments. For the case of sample covariance matrices, the analysis led to the Marchenko-Pastur law in [37], employing a Stieltjes transform method. Lastly, in the case of non-selfadjoint random matrices, the limit distribution is known as the circular law, which was derived in [54] by use of a transform called *the logarithmic potential*, see [13] for a detailed exposition.

Apart from the global analysis of the eigenvalues of the aforementioned models, the *local behavior* of the eigenvalues also gained a lot of attention. For example, for self-adjoint and sample covariance matrices – having real-valued eigenvalues only – it would be natural to analyze the behavior of the largest eigenvalue. To be precise, does the largest eigenvalue converge (almost surely, say) to the edge of the support of the respective limiting spectral distribution, and what is its fluctuation after appropriate rescaling? The questions about almost sure convergence were answered in [6] and [29]. The fluctuations of the largest eigenvalue are given by the famous Tracy-Widom Law derived in [57, 50] for self-adjoint matrices. For the Tracy-Widom law for sample covariance matrices, see the recent paper [46] and references therein for earlier achievements. But the analysis of the local behavior of the eigenvalues of random matrices did not stop with the largest eigenvalue, but rather was succeeded by sophisticated analyses known as *local laws*. Local laws answer the question about whether

eigenvalue distributions are also well-approximated by the target distribution when restricting the analysis to smaller and smaller intervals. The technique to answer this question involves an extensive generalization of the Stieltjes transform method, namely the analysis of the resolvents of the random matrices, leading to matrix-valued limit laws, see [7, 22, 21] for didactical expositions, [23, 55] and references therein for original research, and [31, 30] for improvements thereupon. With these techniques, it was also shown that the random eigenvectors (of self-adjoint random matrices with independent entries, for example) are delocalized, meaning roughly that all vector entries are of the same order. In addition, the technical framework of the local law was used to resolve the Wigner-Dyson-Mehta (WDM) universality conjecture. For a description of this conjecture, see [21] and references therein.

The techniques to derive the limiting results mentioned so far all benefitted greatly from the independence of the matrix entries. A natural question to ask is to what extent this condition may be relaxed without jeopardizing classical limit laws. For works in this area, see [15] and [28] and references therein.

Another area of study is the spectrum of sums and products of random matrices. One possible framework to tackle these questions is called *free probability*, see [39]. Other techniques and results are included in [42], an extensive part of which is devoted to this topic. Further, in [2] a direct generalization of the method of moments is given for the case of a product of random matrices, and a generalization of the Marchenko-Pastur law is obtained.

2. Weak convergence

2.1. Spaces of continuous functions

On the set $\mathbb{R}$ of real numbers we will always consider the standard topology and the associated Borel σ-algebra $\mathcal{B}$. To study convergence of probability measures on $(\mathbb{R}, \mathcal{B})$, it is useful to get acquainted with certain spaces of functions $\mathbb{R} \to \mathbb{R}$ first. If $f : \mathbb{R} \to \mathbb{R}$ is a function, we define the *support* of f as

$$\operatorname{supp}(f) \coloneqq \overline{\{x \in \mathbb{R} : f(x) \neq 0\}}.$$

Note that by definition, the support of f is always a closed subset of $\mathbb{R}$, and it is immediate that a point $x \in \mathbb{R}$ lies in the support of f if and only if for any $\varepsilon > 0$ there is a $y \in B_\varepsilon(x)$, such that $f(y) \neq 0$. Here and later, $B_\delta(z)$ denotes the open δ-ball around the element z in a metric space which is clear from the context.

We say that a function $f : \mathbb{R} \to \mathbb{R}$ *vanishes at infinity*, if

$$\lim_{x \to \pm\infty} f(x) = 0.$$

Denote by $\mathcal{C}(\mathbb{R})$ the vector space of continuous functions $\mathbb{R} \to \mathbb{R}$. We define the three subspaces

1. $\mathcal{C}_b(\mathbb{R}) \coloneqq \{f : \mathbb{R} \to \mathbb{R} \,|\, f \text{ is continuous and bounded}\}$,

2. $\mathcal{C}_0(\mathbb{R}) \coloneqq \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous and vanishes at infinity}\}$ and
3. $\mathcal{C}_c(\mathbb{R}) \coloneqq \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous with compact support}\}$.

It is clear that

$$\mathcal{C}_c(\mathbb{R}) \subsetneq \mathcal{C}_0(\mathbb{R}) \subsetneq \mathcal{C}_b(\mathbb{R}) \subsetneq \mathcal{C}(\mathbb{R}),$$

since the function $x \mapsto \min(1, 1/|x|)$ lies in $\mathcal{C}_0(\mathbb{R})\backslash\mathcal{C}_c(\mathbb{R})$, the function $x \mapsto \mathbb{1}_{\mathbb{R}}(x)$ lies in $\mathcal{C}_b(\mathbb{R})\backslash\mathcal{C}_0(\mathbb{R})$ and the function $x \mapsto x$ lies in $\mathcal{C}(\mathbb{R})\backslash\mathcal{C}_b(\mathbb{R})$. Since all functions in $\mathcal{C}_c(\mathbb{R})$, $\mathcal{C}_0(\mathbb{R})$ and $\mathcal{C}_b(\mathbb{R})$ are bounded, we can equip these spaces with the supremum norm $\|\cdot\|_\infty$ defined by

$$\|f\|_\infty \coloneqq \sup_{x\in\mathbb{R}} |f(x)|.$$

From now on, we will always consider the spaces $\mathcal{C}_b(\mathbb{R})$, $\mathcal{C}_0(\mathbb{R})$ and $\mathcal{C}_c(\mathbb{R})$ as vector spaces normed by the supremum norm. Convergence with respect to this norm is also called *uniform convergence*. To analyze properties of these normed spaces, we introduce continuous cutoff-functions as in [34, p. 8]:

Definition 2.1. *For any real numbers $u > \ell \geq 0$ we define the function $\phi_\ell^u : \mathbb{R} \to [0, 1]$ by*

$$\phi_\ell^u(x) \coloneqq \begin{cases} 1 & \text{if } |x| \leq \ell, \\ \frac{u-|x|}{u-\ell} & \text{if } \ell < |x| < u, \\ 0 & \text{if } |x| \geq u. \end{cases}$$

Note that for any $u > \ell \geq 0$, ϕ_ℓ^u is continuous with compact support $[-u, u]$. The following theorem will summarize important properties of $\mathcal{C}_b(\mathbb{R})$, $\mathcal{C}_0(\mathbb{R})$ and $\mathcal{C}_c(\mathbb{R})$.

Theorem 2.2. *The following statements hold:*

i) $\mathcal{C}_b(\mathbb{R})$ *is complete, but not separable.*
ii) $\mathcal{C}_0(\mathbb{R})$ *is complete and separable.*
iii) $\mathcal{C}_c(\mathbb{R})$ *is not complete, but separable.*
iv) $\mathcal{C}_c(\mathbb{R})$ *is dense in* $\mathcal{C}_0(\mathbb{R})$.

Proof. i) If $(f_n)_n$ is Cauchy in $\mathcal{C}_b(\mathbb{R})$ and $x \in \mathbb{R}$, then $f_n(x)$ is Cauchy in $\mathbb{R}$, thus converges to a limit $f(x) \in \mathbb{R}$. Further, we can pick an $m \in \mathbb{N}$ such that f_m is uniformly ε-close to all f_n for n large enough, from which it follows that $f_n \to f$ uniformly. From this, it easily follows that f is bounded. It remains to show that f is continuous for which we again choose an f_m as above and utilize a standard 3ε-argument. To see that $\mathcal{C}_b(\mathbb{R})$ is not separable, we construct an uncountable subset $\mathcal{F} \subseteq \mathcal{C}_b$, such that for all $f, g \in \mathcal{F}$ with $f \neq g$ we have $\|f - g\|_\infty = 1$. To this end, denote by Z the set of 0-1-sequences, so $Z = \{0, 1\}^{\mathbb{N}}$. Note that Z is uncountable. For any sequence $z \in Z$ we define

$$\forall\, x \in \mathbb{R} : F_z(x) \coloneqq \sum_{i\in\mathbb{N}} z_i \cdot \phi_{0.1}^{0.2}(x - i)$$

and $\mathcal{F} \coloneqq \{F_z \mid z \in Z\}$. Now $\mathcal{F}$ is as desired.

iii)/iv) To show that $\mathcal{C}_c(\mathbb{R})$ is not complete, we show that it is not closed in the strict superset $\mathcal{C}_0(\mathbb{R})$. In fact, we show even more, that is, that $\mathcal{C}_c(\mathbb{R})$ is dense in $\mathcal{C}_0(\mathbb{R})$ (then since $\mathcal{C}_c(\mathbb{R}) \subsetneq \mathcal{C}_0(\mathbb{R})$, $\mathcal{C}_c(\mathbb{R})$ cannot be closed). This fact is also needed for statements ii) and iv). So let $f \in \mathcal{C}_0(\mathbb{R})$ be arbitrary. Now consider the sequence of functions $(f_n)_n$, where

$$\forall\, n \in \mathbb{N} : \forall\, x \in \mathbb{R} :\ f_n(x) \coloneqq \phi_n^{n+1}(x) f(x).$$

Then $(f_n)_n$ is a sequence in $\mathcal{C}_c(\mathbb{R})$ which converges uniformly to f. Hence, $\mathcal{C}_c(\mathbb{R})$ is dense in $\mathcal{C}_0(\mathbb{R})$. Next, we will show that $\mathcal{C}_c(\mathbb{R})$ is separable. To this end, denote by $\mathcal{P}$ the countable set of all polynomials with rational coefficients and set

$$\mathcal{Q} \coloneqq \{p \cdot \phi_n^{n+1} \,|\, p \in \mathcal{P}, n \in \mathbb{N}\}.$$

Then $\mathcal{Q}$ is easily identified as a dense countable subset of $\mathcal{C}_c(\mathbb{R})$.
ii) To show that $\mathcal{C}_0(\mathbb{R})$ is complete, let $(f_n)_n$ be an arbitrary Cauchy sequence in $\mathcal{C}_0(\mathbb{R})$. This is also a Cauchy sequence in $\mathcal{C}_b(\mathbb{R})$, so with i) we know that there is an $f \in \mathcal{C}_b(\mathbb{R})$ such that $f_n \to f$ uniformly. It is easily seen that f vanishes at infinity, so that $f \in \mathcal{C}_0(\mathbb{R})$. To see that $\mathcal{C}_0(\mathbb{R})$ is separable, note that we have already seen that $\mathcal{C}_c(\mathbb{R})$ is separable and dense in $\mathcal{C}_0(\mathbb{R})$. □

2.2. Convergence of probability measures

We will denote the set of measures on $(\mathbb{R}, \mathcal{B})$ by $\mathcal{M}(\mathbb{R})$, the set of finite measures by $\mathcal{M}_f(\mathbb{R})$, the set of probability measures by $\mathcal{M}_1(\mathbb{R})$, and the set of sub-probability measures by $\mathcal{M}_{\leq 1}(\mathbb{R})$. Here, a measure μ on $(\mathbb{R}, \mathcal{B})$ is called *sub-probability measure*, if $\mu(\mathbb{R}) \in [0, 1]$. Note that

$$\mathcal{M}_1(\mathbb{R}) \subsetneq \mathcal{M}_{\leq 1}(\mathbb{R}) \subsetneq \mathcal{M}_f(\mathbb{R}) \subsetneq \mathcal{M}(\mathbb{R}).$$

As a shorthand notation, if $\mu \in \mathcal{M}(\mathbb{R})$ and $f : \mathbb{R} \to \mathbb{R}$ is measurable, we write

$$\langle \mu, f \rangle \coloneqq \int f \,\mathrm{d}\mu$$

with the convention that when in doubt, x is the variable of integration:

$$\left\langle \mu, x^k \right\rangle = \int x^k \mu(\mathrm{d}x).$$

Definition 2.3. *Let $\mathcal{F} \subseteq \mathcal{C}_b(\mathbb{R})$ be a linear subspace, then a* positive linear bounded functional *I on $\mathcal{F}$ is a bounded $\mathbb{R}$-linear map $\mathcal{F} \to \mathbb{R}$ with $I(f) \geq 0$ for all $f \in \mathcal{F}$ with $f \geq 0$.*

Lemma 2.4. *Let $\mathcal{F} \subseteq \mathcal{C}_b(\mathbb{R})$ be a linear subspace with $\mathcal{C}_c(\mathbb{R}) \subseteq \mathcal{F}$. Then for any $\mu \in \mathcal{M}_f(\mathbb{R})$, the map*

$$\begin{aligned} I_\mu : \mathcal{F} &\longrightarrow \mathbb{R} \\ f &\longmapsto I_\mu(f) \coloneqq \langle \mu, f \rangle \end{aligned}$$

defines a positive linear bounded functional on $\mathcal{F}$ with operator norm $\mu(\mathbb{R})$.

Proof. We only need to show that the operator norm is indeed $\mu(\mathbb{R})$. To see this, note that for any $k > 0$, we have $\phi_k^{k+1} \in \mathcal{F}$, $\phi_k^{k+1} \geq 0$ and $\|\phi_k^{k+1}\|_\infty = 1$. Further,

$$I_\mu(\phi_k^{k+1}) = \left\langle \mu, \phi_k^{k+1} \right\rangle \geq \mu([-k,k]).$$

Thus, the operator norm of I_μ is at least $\mu([-k,k])$ for all $k > 0$, hence at least $\mu(\mathbb{R})$. On the other hand, the operator norm is at most $\mu(\mathbb{R})$, since for any $f \in \mathcal{F}$ we find $|\langle \mu, f\rangle| \leq \langle \mu, |f|\rangle \leq \mu(\mathbb{R}) \cdot \|f\|_\infty$. □

The representation theorem of Riesz now states that *any* positive linear bounded functional I on a linear space $\mathcal{F}$ with $\mathcal{C}_c(\mathbb{R}) \subseteq \mathcal{F} \subseteq \mathcal{C}_0(\mathbb{R})$ has the form $I = I_\mu$ as in Lemma 2.4.

Theorem 2.5. *Let $\mathcal{F}$ be a linear space with $\mathcal{C}_c(\mathbb{R}) \subseteq \mathcal{F} \subseteq \mathcal{C}_0(\mathbb{R})$ and equipped with the supremum norm. Then for any positive linear bounded functional I on $\mathcal{F}$, there exists exactly one $\mu \in \mathcal{M}_f(\mathbb{R})$ with $I = I_\mu$. It then holds $\|I\|_{\mathrm{op}} = \mu(\mathbb{R})$.*

Proof. The statement is well-known, see e.g. [17] or [19]. □

The next lemma will help us infer equality of two finite measures. Notationally, if A is a subset of a topological space, we denote its boundary by ∂A.

Lemma 2.6. *Let μ and ν be two finite measures on $(\mathbb{R}, \mathcal{B})$ and let $\mathcal{F} \subseteq \mathcal{C}_c(\mathbb{R})$ be a dense subset. Then*

i) $\mu = \nu \quad \Leftrightarrow \quad \mu(I) = \nu(I)$ *for all bounded intervals I with* $\mu(\partial I) = \nu(\partial I) = 0$,

ii) $\mu = \nu \quad \Leftrightarrow \quad \forall\, f \in \mathcal{C}_c(\mathbb{R}) : \langle \mu, f\rangle = \langle \nu, f\rangle \quad \Leftrightarrow \quad \forall\, f \in \mathcal{F} : \langle \mu, f\rangle = \langle \nu, f\rangle.$

Proof. i) "⇒" is clear, and for "⇐" we show that μ and ν agree on all finite open intervals. To this end, note that for any finite measure $\rho \in \mathcal{M}_f(\mathbb{R})$, the set of atoms $A_\rho := \{x \in \mathbb{R} \,|\, \rho(x) > 0\}$ is at most countable. As a result $\mathbb{R}\backslash(A_\mu \cup A_\nu)$ is dense in $\mathbb{R}$. For arbitrary $a < b$ in $\mathbb{R}$, we find sequences $(a_n)_n$ and $(b_n)_n$ in $\mathbb{R}\backslash(A_\mu \cup A_\nu)$ with $a_n \searrow a$ and $b_n \nearrow b$ as $n \to \infty$ and $a_n < b_n$ for all $n \in \mathbb{N}$. Then we obtain with continuity of measures from below (note that μ and ν agree on all intervals (a_n, b_n)):

$$\mu((a,b)) = \lim_{n\to\infty} \mu((a_n, b_n)) = \lim_{n\to\infty} \nu((a_n, b_n)) = \nu((a,b)).$$

ii) This follows immediately with Theorem 2.5. □

We are especially interested in convergence behavior of sequences in $\mathcal{M}_1(\mathbb{R})$, where the limit may lie in $\mathcal{M}_{\leq 1}(\mathbb{R})$.

Definition 2.7. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in in $\mathcal{M}_1(\mathbb{R})$.*

i) The sequence $(\mu_n)_{n\in\mathbb{N}}$ is said to converge weakly *to an element $\mu \in \mathcal{M}_1(\mathbb{R})$, if*

$$\forall f \in \mathcal{C}_b(\mathbb{R}) : \lim_{n\to\infty} \langle \mu_n, f\rangle = \langle \mu, f\rangle\,. \tag{2.1}$$

ii) The sequence $(\mu_n)_{n\in\mathbb{N}}$ *is said to converge* vaguely *to an element* $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$, *if*

$$\forall f \in \mathcal{C}_c(\mathbb{R}) : \lim_{n\to\infty} \langle \mu_n, f\rangle = \langle \mu, f\rangle \,. \tag{2.2}$$

Remark 2.8. *We would like to shed light on the seemingly innocent Definition 2.7:*

1. *Weak convergence clearly implies vague convergence. Further, due to Lemma 2.6, weak and vague limits are unique.*
2. *In light of Theorem 2.2, it is appropriate to say that the set of test functions for weak convergence is considerably larger than the set of test functions for vague convergence. As a result, weak limits are much more restrictive than vague limits, as clarified by the next two points.*
3. *The target measures* $\mu \in \mathcal{M}(\mathbb{R})$, *for which (2.1) can be satisfied for some sequence* $(\mu_n)_n$ *of probability measures are exactly all* $\mu \in \mathcal{M}_1(\mathbb{R})$. *To see this, if (2.1) holds for some* $\mu \in \mathcal{M}(\mathbb{R})$ *and a sequence* $(\mu_n)_n$ *in* $\mathcal{M}_1(\mathbb{R})$, *then we must have* $\mu(\mathbb{R}) = 1$, *since* $\mathbb{1}_{\mathbb{R}} \in \mathcal{C}_b(\mathbb{R})$. *On the other hand, if* $\mu \in \mathcal{M}_1(\mathbb{R})$ *is arbitrary, then (2.1) is satisfied for the sequence* $(\mu_n)_n$, *where* $\mu_n = \mu$ *for all* $n \in \mathbb{N}$.
4. *The measures* $\mu \in \mathcal{M}(\mathbb{R})$, *for which (2.2) can be satisfied for some sequence* $(\mu_n)_n$ *of probability measures are (somewhat surprisingly) exactly all* $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$. *To see this, if (2.2) holds for some* $\mu \in \mathcal{M}(\mathbb{R})$ *and a sequence* $(\mu_n)_n$ *in* $\mathcal{M}_1(\mathbb{R})$, *then we have for any* $m \in \mathbb{N}$ *that* $\langle \mu_n, \phi_m^{m+1}\rangle \to_n \langle \mu, \phi_m^{m+1}\rangle$, *so* $\langle \mu, \phi_m^{m+1}\rangle \leq 1$, *which entails* $\mu([-m,m]) \leq 1$ *for all* $m \in \mathbb{N}$. *Since measures are continous from below, we conclude that also* $\mu(\mathbb{R}) \leq 1$, *so* μ *is a sub-probability measure. On the other hand, if* $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$ *is arbitrary, then define* $\alpha := 1 - \mu(\mathbb{R}) \in [0,1]$ *and for all* $n \in \mathbb{N} : \mu_n := \mu + \alpha\delta_n$. *Then* $(\mu_n)_n$ *is a sequence of probability measures and (2.2) is satisfied for the sequence* $(\mu_n)_n$. *To see this, let* $f \in \mathcal{C}_c(\mathbb{R})$ *be arbitrary and* $N \in \mathbb{N}$ *be so large that* $\operatorname{supp}(f) \subseteq [-N, N]$. *Then it holds for all* $n \geq N$ *that* $\langle \mu_n, f\rangle = \langle \mu, f\rangle + \alpha f(n) = \langle \mu, f\rangle$.
5. *As a result of points 3. and 4., the limit domains for weak and vague convergence in Definition 2.7 are exact. The probability measures lie* vaguely dense *in the sub-probability measures.*

Lemma 2.9. *Let* $(\mu_n)_{n\in\mathbb{N}}$ *be a sequence of probability measures and* μ *a sub-probability measure on* $(\mathbb{R}, \mathcal{B})$. *Then* $(\mu_n)_{n\in\mathbb{N}}$ *converges vaguely (resp. weakly) to* μ *if and only if every subsequence* $(\mu_n)_{n\in J}$, $J \subseteq \mathbb{N}$, *has a subsequence* $(\mu_n)_{n\in I}$, $I \subseteq J$, *that converges vaguely (resp. weakly) to* μ.

Proof. Of course, we only need to show "⇐". We assume the statement to be false, that is, that it is not true that $(\mu_n)_{n\in\mathbb{N}}$ converges vaguely (resp. weakly) to μ. Then we find a continuous function $f : \mathbb{R} \to \mathbb{R}$ which has compact support (resp. which is bounded) and an $\varepsilon > 0$ such that $|\langle \mu_n, f\rangle - \langle \mu, f\rangle| \geq \varepsilon$ for all $n \in J$, where $J \subseteq \mathbb{N}$ is an infinite subset. But now we find a subsequence $(\mu_n)_{n\in I}$, $I \subseteq J$ that converges vaguely (resp. weakly) to μ. In particular, we find an $n \in I \subseteq J$ such that $|\langle \mu_n, f\rangle - \langle \mu, f\rangle| < \varepsilon$, which leads to a contradiction to our assumption that the statement is false. □

Vague convergence of probability measures can also be characterized by convergence of the integrals $\langle \mu_n, f \rangle$ for all $f \in \mathcal{C}_0(\mathbb{R})$.

Lemma 2.10. *A sequence $(\mu_n)_n$ in $\mathcal{M}_1(\mathbb{R})$ converges vaguely to an element $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$, if and only if*

$$\forall f \in \mathcal{C}_0(\mathbb{R}) : \lim_{n\to\infty} \langle \mu_n, f \rangle = \langle \mu, f \rangle .$$

Proof. This follows easily with the fact that $\mathcal{C}_c(\mathbb{R}) \subseteq \mathcal{C}_0(\mathbb{R})$ is dense. □

If $\mu_n \to \mu$ weakly, we know that $\langle \mu_n, f \rangle \to \langle \mu, f \rangle$ for all $f \in \mathcal{C}_b(\mathbb{R})$. Often, we would like to be able to conclude $\langle \mu_n, f \rangle \to \langle \mu, f \rangle$ for more general functions f. The next lemma will be of great use in this respect, see also [18, p. 107].

Lemma 2.11. *Let $(\mu_n)_n$ and μ be probability measures such that $\mu_n \to \mu$ weakly as $n \to \infty$. Let $h : \mathbb{R} \to \mathbb{R}$ be continuous. Then to show*

$$\langle \mu_n, h \rangle \xrightarrow[n\to\infty]{} \langle \mu, h \rangle ,$$

it is sufficient to show that there is a strictly positive continuous function $g : \mathbb{R} \to (0, \infty)$ such that $\sup_{n\in\mathbb{N}} \langle \mu_n, g \rangle < \infty$ and h/g vanishes at infinity.

Proof. The proof follows from elementary calculations and is left to the reader, see also Exercise 3.2.5 in [18, p. 107]. □

As we just saw in Remark 2.8, vague convergence allows the escape of probability mass. The concept of tightness prevents this from happening:

Definition 2.12. *A sequence of probability measures $(\mu_n)_n$ on $(\mathbb{R}, \mathcal{B})$ is called* tight*, if for all $\varepsilon > 0$ there exists a compact subset $K \subseteq \mathbb{R}$ such that*

$$\forall\, n \in \mathbb{N} :\ \mu_n(K^c) \leq \varepsilon.$$

A sufficient condition for tightness is given in the next Lemma, which we adopted from [18, p. 106]:

Lemma 2.13. *Let $(\mu_n)_n$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$. If there exists a measurable non-negative function $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(x) \to \infty$ for $x \to \pm\infty$ and*

$$\sup_n \langle \mu_n, \phi \rangle < \infty,$$

then $(\mu_n)_n$ is tight. In particular, this holds true if

$$\sup_n \left\langle \mu_n, x^2 \right\rangle < \infty.$$

Proof. Let $C \coloneqq \sup_n \langle \mu_n, \phi \rangle < \infty$. Then it holds for any $n \in \mathbb{N}$ and $k > 0$ that

$$C \geq \langle \mu_n, \phi \rangle \geq \left\langle \mu_n, \mathbb{1}_{[-k,k]^c} \cdot \inf_{|x|>k} \phi(x) \right\rangle = \left\langle \mu_n, \mathbb{1}_{[-k,k]^c} \right\rangle \cdot \inf_{|x|>k} \phi(x).$$

Since $\inf_{|x|>k} \phi(x) \to \infty$ as $k \to \infty$, the statement follows. □

Lemma 2.14. *Let $(\mu_n)_n$ be a sequence in $\mathcal{M}_1(\mathbb{R})$ and $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$ such that $\mu_n \to \mu$ vaguely as $n \to \infty$, then the following statements are equivalent:*

i) $(\mu_n)_n$ is tight.
ii) μ is a probability measure.
iii) μ_n converges weakly to μ.

Proof. i) ⇒ iii) Let $f \in \mathcal{C}_b(\mathbb{R})$ be arbitrary and set $s \coloneqq \max(\|f\|_\infty, 1)$. Let $\varepsilon > 0$ be arbitrary, then due to tightness of $(\mu_n)_n$ and continuity from below of μ, we find a $k > 0$ such that $\mu_n([-k,k]^c) \leq \frac{\varepsilon}{2s}$ and $\mu([-k,k]^c) \leq \frac{\varepsilon}{2s}$. Now for $n \in \mathbb{N}$ arbitrary we find

$$\begin{aligned}
&|\langle \mu_n, f\rangle - \langle \mu, f\rangle| \\
&\leq |\langle \mu_n, f\rangle - \langle \mu_n, f\phi_k^{k+1}\rangle| + |\langle \mu_n, f\phi_k^{k+1}\rangle - \langle \mu, f\phi_k^{k+1}\rangle| + |\langle \mu, f\phi_k^{k+1}\rangle - \langle \mu, f\rangle| \\
&\leq \langle \mu_n, |f| \cdot |1 - \phi_k^{k+1}|\rangle + |\langle \mu_n, f\phi_k^{k+1}\rangle - \langle \mu, f\phi_k^{k+1}\rangle| + \langle \mu, |f| \cdot |\phi_k^{k+1} - 1|\rangle \\
&\leq s \cdot \frac{\varepsilon}{2s} + |\langle \mu_n, f\phi_k^{k+1}\rangle - \langle \mu, f\phi_k^{k+1}\rangle| + s \cdot \frac{\varepsilon}{2s}
\end{aligned}$$

It follows that $\limsup_n |\langle \mu_n, f\rangle - \langle \mu, f\rangle| \leq \varepsilon$.

iii) ⇒ ii) This statement is obvious. Consider $\mathbb{1}_{\mathbb{R}} \in \mathcal{C}_b(\mathbb{R})$.

ii) ⇒ i). Let $\varepsilon > 0$ be arbitrary. Then for $k > 0$ we find

$$\mu_n([-(k+1), k+1]) \geq \langle \mu_n, \phi_k^{k+1}\rangle \geq \langle \mu, \phi_k^{k+1}\rangle - |\langle \mu, \phi_k^{k+1}\rangle - \langle \mu_n, \phi_k^{k+1}\rangle|$$

Now first choose k large enough such that the first summand on the r.h.s. is larger than $1 - \varepsilon/2$, then choose $N \in \mathbb{N}$ large enough such that for all $n > N$ the absolute value on the r.h.s. is at most $\varepsilon/2$. Then we obtain for all $n > N$ that $\mu_n([-(k+1), k+1]) \geq 1 - \varepsilon$. On the other hand, we find $k_1, \ldots, k_N > 0$ such that

$$\forall\, i \in \{1, \ldots, N\} :\ \mu_i([-k_i, k_i]) \geq 1 - \varepsilon.$$

Let $k^* \coloneqq \max\{k+1, k_1, \ldots, k_N\}$, then we obtain for all $n \in \mathbb{N}$ that $\mu_n([-k^*, k^*]) \geq 1 - \varepsilon$. Therefore, $(\mu_n)_n$ is tight. □

Lemma 2.15. *Let $(\mu_n)_n$ be a sequence in $\mathcal{M}_1(\mathbb{R})$, then the following statements hold:*

i) $(\mu_n)_n$ has a subsequence converging vaguely to some $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$.
ii) If $(\mu_n)_n$ is tight, it has a subsequence converging weakly to some $\mu \in \mathcal{M}_1(\mathbb{R})$.

Proof. i) Let $(g_m)_m$ be a dense sequence in $\mathcal{C}_c(\mathbb{R})$, then for all $m \in \mathbb{N}$, $(\langle \mu_n, g_m\rangle)_n$ is a sequence in $\mathbb{R}$ whose absolute value is bounded by $\|g_m\|_\infty < \infty$, thus has a convergent subsequence by Bolzano-Weierstrass. By a diagonal argument, we can find a subsequence $J \subseteq \mathbb{N}$, such that for all $m \in \mathbb{N}$, $(\langle \mu_n, g_m\rangle)_{n \in J}$ converges. But since $(g_m)_m$ is dense in $\mathcal{C}_c(\mathbb{R})$, $\lim_{n \in J} \langle \mu_n, f\rangle$ exists for all $f \in \mathcal{C}_c(\mathbb{R})$ (it can be shown that $(\langle \mu_n, f\rangle)_n$ is Cauchy). The function

$$I : \mathcal{C}_c(\mathbb{R}) \longrightarrow \mathbb{R}$$

$$f \longmapsto I(f) \coloneqq \lim_{n\in J} \langle \mu_n, f\rangle$$

is a linear bounded positive functional on $\mathcal{C}_c(\mathbb{R})$ with operator norm at most 1, since $|\langle \mu_n, f\rangle| \leq \|f\|_\infty$ for all $n \in \mathbb{N}$ and $f \in \mathcal{C}_c(\mathbb{R})$. With Theorem 2.5, we find an element $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$ such that $I = I_\mu$, which entails $\mu_n \to \mu$ vaguely for $n \in J$.
ii) With $i)$ we find a subsequence $J \subseteq \mathbb{N}$ and a $\mu \in \mathcal{M}_{\leq 1}(\mathbb{R})$ such that $(\mu_n)_{n\in J}$ converges to μ vaguely. But Lemma 2.14 yields that $\mu \in \mathcal{M}_1(\mathbb{R})$ and $\mu_n \to \mu$ weakly for $n \in J$. □

Note that statement $i)$ of Lemma 2.15 is the well-known Helly's selection theorem contained in most standard books on probability theory, see [18] or [35], for example. However, we give a new proof here that differs completely from the standard proofs which utilize distribution functions.

So far we have discussed the intricacies of weak and vague convergence of probability measures. Our next goal is to better understand the topology of weak convergence on $\mathcal{M}_1(\mathbb{R})$, which will deepen our understanding of stochastic weak convergence to be discussed in the next section. Our first goal will be to reduce the number of test functions for weak convergence to a countable subset of $\mathcal{C}_b(\mathbb{R})$. However, $(\mathcal{C}_b(\mathbb{R}), \|\cdot\|_\infty)$ is large; it is not even separable. But there is no reason for despair, since the following theorem holds, which we adopted from our previous work [25].

Theorem 2.16. *Fix a sequence $(g_k)_{k\in\mathbb{N}}$ in $\mathcal{C}_c(\mathbb{R})$ which lies dense in $\mathcal{C}_c(\mathbb{R})$. Then the following statements hold:*

i) Let μ, $(\mu_n)_n \in \mathcal{M}_1(\mathbb{R})$, then the following statements are equivalent:

a) $\mu_n \to \mu$ weakly.

b) $\forall\, k \in \mathbb{N} : \langle \mu_n, g_k\rangle \xrightarrow[n\to\infty]{} \langle \mu, g_k\rangle$.

ii) Define for all $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$:

$$d_M(\mu,\nu) \coloneqq \sum_{k\in\mathbb{N}} \frac{|\langle \mu, g_k\rangle - \langle \nu, g_k\rangle|}{2^k \cdot (1 + |\langle \mu, g_k\rangle - \langle \nu, g_k\rangle|)}.$$

Then d_M forms a metric on $\mathcal{M}_1(\mathbb{R})$ which metrizes weak convergence. That is, a sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathcal{M}_1(\mathbb{R})$ converges weakly to $\mu \in \mathcal{M}_1(\mathbb{R})$ iff $d_M(\mu_n, \mu) \to 0$ as $n \to \infty$.

iii) $(\mathcal{M}_1(\mathbb{R}), d_M)$ is a separable, but not complete, metric space.

Proof. i) Let $(\mu_n)_{n\in\mathbb{N}}$ and μ be probability measures. If $\mu_n \to \mu$ weakly, then surely we have for all $k \in \mathbb{N}$ that $\langle \mu_n, g_k\rangle \to \langle \mu, g_k\rangle$ as $n \to \infty$. If on the other hand we have for all $k \in \mathbb{N}$ that $\langle \mu_n, g_k\rangle \to \langle \mu, g_k\rangle$ as $n \to \infty$, then one easily sees that μ_n converges vaguely to μ, and then also weakly by Lemma 2.14.

ii) and iii): From Lemma 2.6, we find for any $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ that

$$\mu = \nu \;\Leftrightarrow\; \forall\, k \in \mathbb{N} : \langle \mu, g_k\rangle = \langle \nu, g_k\rangle\,.$$

Next, we will inspect the space $\mathbb{R}^{\mathbb{N}}$ endowed with the product topology. With respect to this topology, a sequence $(z_n)_n$ in $\mathbb{R}^{\mathbb{N}}$ converges to a $z \in \mathbb{R}^{\mathbb{N}}$ iff for all $i \in \mathbb{N}$ the coordinates $z_n(i)$ in $\mathbb{R}$ converge to $z(i)$ as $n \to \infty$. Further, it is well-known that the topology on $\mathbb{R}^{\mathbb{N}}$ is metrizable through the metric ρ with

$$\forall\, x, y \in \mathbb{R}^{\mathbb{N}} : \quad \rho(x,y) \coloneqq \sum_{k\in\mathbb{N}} \frac{|x(k) - y(k)|}{2^k \cdot (1 + |x(k) - y(k)|)}.$$

This follows (for example) with 3.5.7 in [48, p. 121] in combination with Theorem 4.2.2 in [20, p. 259]. Further, $(\mathbb{R}^{\mathbb{N}}, \rho)$ is a *separable* metric space (Theorem 16.4 in [60, p. 109]).

We now define the following map (see [40, p. 43]):

$$\begin{aligned} T : \mathcal{M}_1(\mathbb{R}) &\longrightarrow \mathbb{R}^{\mathbb{N}} \\ \mu &\longmapsto (\langle \mu, g_1 \rangle, \langle \mu, g_2 \rangle, \ldots) \end{aligned}$$

Then surely, T is injective, since if $T(\mu) = T(\nu)$, then also for all $k \in \mathbb{N}$: $\langle \mu, g_k \rangle = \langle \nu, g_k \rangle$ and then $\mu = \nu$. Additionally, we have for all $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ that

$$d_M(\mu,\nu) = \sum_{k\in\mathbb{N}} \frac{|\langle \mu, g_k \rangle - \langle \nu, g_k \rangle|}{2^k \cdot (1 + |\langle \mu, g_k \rangle - \langle \nu, g_k \rangle|)} = \rho(T(\mu), T(\nu)). \tag{2.3}$$

Since T injective and ρ is a metric, d_M is a metric as well, so that $(\mathcal{M}_1(\mathbb{R}), d_M)$ is a metric space. With equation (2.3) we see that $T : (\mathcal{M}_1(X,d), d_M) \longrightarrow \mathbb{R}^{\mathbb{N}}$ is not only injective, but even isometric, especially continuous and a homeomorphism onto its image. Surely, the image is separable as a subspace of a separable metric space. Thus, $(\mathcal{M}_1(\mathbb{R}), d_M)$, being homeomorphic to a separable space, is also separable (Corollary 1.4.11 in [20, p. 31]).

With what we have shown so far we obtain for arbitrary $(\mu_n)_{n\in\mathbb{N}}, \mu \in \mathcal{M}_1(\mathbb{R})$:

$$\begin{aligned} &\mu_n \text{ converges weakly to } \mu \\ \Leftrightarrow\ &\forall\, k \in \mathbb{N} : \langle \mu_n, g_k \rangle \xrightarrow[n\to\infty]{} \langle \mu, g_k \rangle \\ \Leftrightarrow\ &T(\mu_n) \xrightarrow[n\to\infty]{} T(\mu) \text{ in } \mathbb{R}^{\mathbb{N}} \\ \Leftrightarrow\ &\rho(T(\mu_n), T(\mu)) \xrightarrow[n\to\infty]{} 0 \\ \Leftrightarrow\ &d_M(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

We showed the first equivalence in the first part of this proof, the second equivalence holds per definition of T and the above mentioned characterization of convergence in $\mathbb{R}^{\mathbb{N}}$, the third equivalence follows with the metrizability of $\mathbb{R}^{\mathbb{N}}$ through ρ, and the last equivalence follows from above equation (2.3). What is left to show is that $(\mathcal{M}_1(\mathbb{R}), d_M)$ is not complete. To this end, let $(\mu_n)_n$ be any sequence in $\mathcal{M}_1(\mathbb{R})$ which converges vaguely to a sub-probability measure ν with $\nu(\mathbb{R}) < 1$. Then for all $k \in \mathbb{N}$, $\langle \mu_n, g_k \rangle \to \langle \nu, g_k \rangle$ as $n \to \infty$. Thus, $d_M(\mu_n, \nu) \to 0$ as $n \to \infty$ (the function d_M makes sense even with sub-probability measures as arguments). Since for any $n, m \in \mathbb{N}$, $d_M(\mu_n, \mu_m) \le d_M(\mu_n, \nu) + d_M(\mu_m, \nu)$, we find that $(\mu_n)_n$ is a Cauchy sequence in $(\mathcal{M}_1(\mathbb{R}), d_M)$ that does not converge weakly to an element in $\mathcal{M}_1(\mathbb{R})$. □

2.3. Random probability measures on $(\mathbb{R}, \mathcal{B})$

As we saw in Theorem 2.16, the set $\mathcal{M}_1(\mathbb{R})$ can be metrized in such a way that the resulting convergence is exactly "weak convergence of probability measures." This shows that Definition 2.7 was adequate in the sense that it defined weak convergence for sequences of probability measures rather than for nets. The reason is that in metric spaces (or more generally, in spaces which satisfy the first axiom of countability, which means that any point has a countable neighborhood basis), the topology can be reconstructed from the knowledge of convergent sequences rather than nets. This is due to the fact that a set in such a space is closed iff any limit of a convergent *sequence* in the set is an element of the set.

From now on, we will always view $\mathcal{M}_1(\mathbb{R})$ as equipped with the topology of weak convergence and the associated Borel σ-algebra. We know that $\mathcal{M}_1(\mathbb{R})$ is separable and that d_M as in Theorem 2.16 is a metric yielding the topology of weak convergence. It is then a triviality that for any $f \in \mathcal{C}_b(\mathbb{R})$, the function

$$\begin{aligned} I_f : \mathcal{M}_1(\mathbb{R}) &\longrightarrow \mathbb{R} \\ \mu &\longmapsto I_f(\mu) \coloneqq \langle \mu, f \rangle \end{aligned}$$

is continuous on $\mathcal{M}_1(\mathbb{R})$.

Since $\mathcal{M}_1(\mathbb{R})$ is now considered also as a measurable space, we can study $\mathcal{M}_1(\mathbb{R})$-valued random variables, which is the subject of this section.

Definition 2.17. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.*

i) A random probability measure *on $(\mathbb{R}, \mathcal{B})$ is a measurable map $\mu : \Omega \to \mathcal{M}_1(\mathbb{R})$, $\omega \mapsto \mu(\omega, \cdot)$.*

ii) A stochastic kernel *from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$ is a map $\mu : \Omega \times \mathcal{B} \longrightarrow \mathbb{R}$, so that the following holds:*

a) For all $\omega \in \Omega$, $\mu(\omega, \cdot)$ is a probability measure on $(\mathbb{R}, \mathcal{B})$.

b) For all $B \in \mathcal{B}$, $\mu(\cdot, B)$ is $\mathcal{A}$-$\mathcal{B}$-measurable.

Lemma 2.18. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.*

i) A map $\mu : \Omega \times \mathcal{B} \longrightarrow \mathbb{R}$ is a random probability measure iff it is a stochastic kernel.

ii) If μ is a stochastic kernel from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$ and $f : \mathbb{R} \to \mathbb{R}$ is measurable and bounded, then $\omega \mapsto \langle \mu(\omega), f \rangle$ is measurable and bounded by $\|f\|_\infty$.

Proof. We first show *ii)*: Surely, the indicated map is bounded by $\|f\|_\infty$, since we have for all $\omega \in \Omega$:

$$|\langle \mu(\omega), f \rangle| \leq \langle \mu(\omega), |f| \rangle \leq \langle \mu(\omega), \|f\|_\infty \rangle \leq \|f\|_\infty.$$

To show measurability, we employ a standard extension argument: $\omega \mapsto \mu(\omega, B)$ is measurable for all $B \in \mathcal{B}$. Let f be a simple function on $(\mathbb{R}, \mathcal{B})$, that is, $f = \sum_{i=1}^n \alpha_i \cdot \mathbb{1}_{B_i}$ for some $n \in \mathbb{N}$, $\alpha_i \in [0, \infty)$ and $B_i \in \mathcal{B}$, $i = 1, \ldots, n$, then also $\omega \mapsto \langle \mu(\omega), f \rangle = \sum_{i=1}^n \alpha_i \cdot \mu(\omega, B_i)$ is measurable as a linear combination of finitely many measurable functions. Now let $f \geq 0$ be measurable

and bounded, then there exists sequence of simple functions $(f_n)_{n\in\mathbb{N}}$ such that $f_n \nearrow_n f$ pointwise. For $\omega \in \Omega$ arbitrary it follows per monotone convergence that $\langle \mu(\omega), f_n\rangle \nearrow_n \langle \mu(\omega), f\rangle$, so also $\omega \mapsto \langle \mu(\omega), f\rangle$ is measurable as a pointwise limit of measurable functions. Now if $f : \mathbb{R} \longrightarrow \mathbb{R}$ is measurable and bounded, then also the positive and negative parts f_+ and f_- (then $f_+, f_- \geq 0$ with $f = f_+ - f_-$). Then $\omega \mapsto \langle \mu(\omega), f\rangle = \langle \mu(\omega), f_+\rangle - \langle \mu(\omega), f_-\rangle$ is measurable as a difference of measurable functions.

We now show i):

"$\Leftarrow$" We have just shown that for all $f \in \mathcal{C}_b(\mathbb{R})$ the map $\omega \mapsto \langle \mu(\omega), f\rangle$ is measurable. Then we obtain for all $\nu \in \mathcal{M}_1(\mathbb{R})$ that the map $\omega \mapsto d_M(\mu(\omega), \nu)$ is measurable as a limit of measurable functions, since

$$d_M(\mu(\omega), \nu) = \sum_{k\in\mathbb{N}} \frac{|\langle \mu(\omega), g_k\rangle - \langle \nu, g_k\rangle|}{2^k \cdot (1 + |\langle \mu(\omega), g_k\rangle - \langle \nu, g_k\rangle|)}.$$

To show the measurability of $\omega \mapsto \mu(\omega, \cdot)$, it suffices to show that preimages of open balls from $(\mathcal{M}_1(\mathbb{R}), d_M)$ are measurable, since the σ-algebra on $\mathcal{M}_1(\mathbb{R})$ is generated by the topology which is generated by the metric d_M, and the space $\mathcal{M}_1(\mathbb{R})$ is separable with respect to the topology of weak convergence. So let $\nu \in \mathcal{M}_1(\mathbb{R})$ and $\varepsilon > 0$ be arbitrary, then it holds with $B_\varepsilon^{\mathcal{M}_1(\mathbb{R})}(\nu) \coloneqq \{\nu' \in \mathcal{M}_1(\mathbb{R}) : d_M(\nu', \nu) < \varepsilon\}$:

$$\mu^{-1}\left(B_\varepsilon^{\mathcal{M}_1(\mathbb{R})}(\nu)\right) = \{\omega \in \Omega : d_M(\mu(\omega), \nu) < \varepsilon\} = d_M(\mu(\cdot), \nu)^{-1}([0, \varepsilon)) \in \mathcal{A},$$

since above we already recognized $d_M(\mu(\cdot), \nu)$ as measurable.

"$\Rightarrow$" If μ is a random probability measure, then for all $\omega \in \Omega$, $\mu(\omega, \cdot)$ is a probability measure on $(\mathbb{R}, \mathcal{B})$. We now argue that for any $B \in \mathcal{B}$, $\omega \mapsto \mu(\omega, B)$ is measurable. We first prove this for all open bounded intervals in $\mathbb{R}$, since these intervals generate $\mathcal{B}$. So let $a < b \in \mathbb{R}$ be arbitrary and define $\varepsilon \coloneqq (b-a)/4$. Then define for all $n \in \mathbb{N}$ the function $\phi_n : \mathbb{R} \to \mathbb{R}$ so that $\phi_n \equiv 1$ on $[a + \frac{1}{n}\varepsilon, b - \frac{1}{n}\varepsilon]$, $\phi_n \equiv 0$ on $(a, b)^c$ and ϕ_n is affine on the intervals $[a, a + \frac{1}{n}\varepsilon]$ and $[b - \frac{1}{n}a, b]$ in such a way that it is continuous. Then ϕ_n is bounded, continuous and $\phi_n(x) \nearrow_n \mathbb{1}_{(a,b)}(x)$ for all $x \in \mathbb{R}$. We know that for all $n \in \mathbb{N}$, $\omega \mapsto \langle \mu(\omega), \phi_n\rangle$ is measurable as a composition of a measurable and a continuous map (see remark before Definition 2.17). Now for any $\omega \in \Omega$:

$$\lim_{n\to\infty} \langle \mu(\omega), \phi_n\rangle = \left\langle \mu(\omega), \mathbb{1}_{(a,b)}\right\rangle = \mu(\omega, (a, b)),$$

by monotone convergence. As a result, $\mu(\cdot, (a, b))$ is $\mathcal{A}$-$\mathcal{B}$-measurable as the pointwise limit of measurable functions. Now define the set

$$\mathcal{G} \coloneqq \{B \in \mathcal{B} \,|\, \omega \mapsto \mu(\omega, B) \text{ is measurable}\}.$$

Surely, all open intervals lie in $\mathcal{G}$ as we have just shown. If we can show that $\mathcal{G}$ is a Dynkin system we can conclude that $\mathcal{G} = \mathcal{B}$, which is our goal. First of all, $\emptyset, \mathbb{R} \in \mathcal{G}$, since constant functions are always measurable. Second, since

$\mu(\cdot, B^c) = 1 - \mu(\cdot, B)$, we have that $B^c \in \mathcal{G}$ whenever $B \in \mathcal{G}$. Third, if $(B_n)_n$ is a sequence of pairwise disjoint sets in $\mathcal{G}$, then $\mu(\cdot, \cup_n B_n) = \sum_n \mu(\cdot, B_n)$, so since all $\mu(\cdot, B_n)$ are measurable, then so is $\mu(\cdot, \cup_n B_n)$ as a pointwise limit of a sequence of measurable functions. This shows that $\cup_n B_n \in \mathcal{G}$ so that $\mathcal{G}$ is indeed a Dynkin system. □

Random probability measures are not so uncommon in probability theory. Consider the next example:

Example 2.19. *Let $Y_1, \dots, Y_n$ be real-valued random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then*

$$\rho := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$$

is a random probability measure on $(\mathbb{R}, \mathcal{B})$, which we call empirical distribution *(of the Y_i). Indeed, for any $\omega \in \Omega$,*

$$\rho(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i(\omega)}$$

is a convex combination of probability measures and thus again a probability measure on $(\mathbb{R}, \mathcal{B})$. On the other hand, if $B \in \mathcal{B}$ is arbitrary, then

$$\omega \mapsto \rho(\omega, B) = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i(\omega)}(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_B(Y_i(\omega))$$

is certainly measurable. Thus, we recognize the empirical distribution ρ as a random probability measure on $(\mathbb{R}, \mathcal{B})$ via Lemma 2.18. For any measurable set B, $\rho(B)$ yields the proportion of the Y_i's that fall into the set B. Connected to the empirical distribution ρ is its empirical distribution function $F_\rho(x) := \rho((-\infty, x])$ defined for all $x \in \mathbb{R}$. This is a random distribution function and the protagonist of the famous Glivenko-Cantelli theorem and the Dvoretzky–Kiefer–Wolfowitz inequality, see [47, p. 321] or [61, p. 553].

Now, let us resume our study. If μ is a random probability measure and $B \in \mathcal{B}$, then $\mu(B)$ is a bounded random variable. It is natural to consider its expectation $\mathbb{E}\mu(B)$ as the expected mass that μ prescribes to the set B. But as it turns out, $B \mapsto \mathbb{E}\mu(B)$ is yet another (deterministic) probability measure:

Theorem 2.20. *Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space and μ be a random probability measure on $(\mathbb{R}, \mathcal{B})$. Then the following statements hold:*

i) The map

$$\begin{aligned} \bar{\mu} : \mathcal{B} &\longrightarrow [0,1] \\ B &\longmapsto \bar{\mu}(B) := \int_\Omega \mu(\omega, B)\, \mathbb{P}(d\omega) = \mathbb{E}\mu(B) \end{aligned}$$

is an element of $\mathcal{M}_1(\mathbb{R})$, the so called expected measure *of μ.*

ii) Any non-negative measurable function $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ is $\bar{\mu}$-integrable iff $\langle \mu, f \rangle$ is $\mathbb{P}$-integrable, and in this case it holds

$$\langle \bar{\mu}, f \rangle = \int_{\mathbb{R}} f(x)\, \bar{\mu}(dx) = \int_{\Omega} \int_{\mathbb{R}} f(x)\, \mu(\omega, dx)\, \mathbb{P}(d\omega) = \mathbb{E}\, \langle \mu, f \rangle .$$

In particular, this equation is valid for any bounded measurable function $f : \mathbb{R} \to \mathbb{R}$.

iii) If $f : \mathbb{R} \to \mathbb{R}$ is $\bar{\mu}$-integrable, then $\langle \mu, f \rangle$ is $\mathbb{P}$-integrable and $\langle \bar{\mu}, f \rangle = \mathbb{E}\, \langle \mu, f \rangle$.

iv) Heed must be taken: If $f : \mathbb{R} \to \mathbb{R}$ is measurable and such that $\langle \mu, f \rangle$ is $\mathbb{P}$-integrable so that $\mathbb{E}\, \langle \mu, f \rangle$ is well-defined, f need not be $\bar{\mu}$-integrable, so that it is not true *that $\langle \bar{\mu}, f \rangle = \mathbb{E}\, \langle \mu, f \rangle$ whenever one of the two exists. In particular, statement ii) cannot be generalized to arbitrary measurable functions $f : \mathbb{R} \to \mathbb{R}$.*

Due to these interrelations we will also write $\mathbb{E}\mu$ instead of $\bar{\mu}$, and with what we have seen so far it holds for all function f with $\mathbb{E}\, \langle \mu, |f| \rangle < \infty$ that f is $\mathbb{E}\mu$-integrable with

$$\langle \mathbb{E}\mu, f \rangle = \langle \bar{\mu}, f \rangle = \mathbb{E}\, \langle \mu, f \rangle .$$

Proof. *i)* Clearly, $\mathbb{E}\mu(\emptyset) = 0$ and $\mathbb{E}\mu(\mathbb{R}) = 1$. Now if $(B_n)_n$ is a sequence of pairwise disjoint elements in $\mathcal{B}$, then

$$\mathbb{E}\mu(\cup_n B_n) = \mathbb{E} \sum_n \mu(B_n) = \sum_n \mathbb{E}\mu(B_n),$$

where in the last step we used dominated convergence. This shows that $\bar{\mu}$ is indeed a probability measure.
ii) Let $f : \mathbb{R} \to \mathbb{R}$ be a simple function, that is, $f = \sum_{i=1}^n \alpha_i \cdot \mathbb{1}_{B_i}$ for some $n \in \mathbb{N}$, $\alpha_i \in [0, \infty)$ and $B_i \in \mathcal{B}$, $i = 1, \ldots, n$. Then

$$\langle \bar{\mu}, f \rangle = \sum_{i=1}^n \alpha_i \cdot \bar{\mu}(B_i) = \mathbb{E} \sum_{i=1}^n \alpha_i \cdot \mu(B_i) = \mathbb{E}\, \langle \mu, f \rangle .$$

Now let $f : \mathbb{R} \to \mathbb{R}$ be non-negative and measurable witnessed by a sequence of simple functions $(f_n)_n$ with $f_n \nearrow_n f$ pointwise, then clearly

$$\langle \bar{\mu}, f \rangle = \lim_{n \to \infty} \langle \bar{\mu}, f_n \rangle = \lim_{n \to \infty} \mathbb{E}\, \langle \mu, f_n \rangle = \mathbb{E}\, \langle \mu, f \rangle ,$$

where in the first and the last step we used monotone convergence. In particular, the non-negative f is $\bar{\mu}$-integrable iff $\langle \mu, f \rangle$ is $\mathbb{P}$-integrable and in this case it holds $\langle \bar{\mu}, f \rangle = \mathbb{E}\, \langle \mu, f \rangle$. Now if $f : \mathbb{R} \to \mathbb{R}$ is bounded, then there exists a $C \in \mathbb{R}$ such that $f + C$ is non-negative (and of course, it remains bounded, thus integrable). Then we immediately obtain $\langle \bar{\mu}, f \rangle = \langle \bar{\mu}, f + C \rangle - C = \mathbb{E}\, \langle \mu, f + C \rangle - C = \mathbb{E}\, \langle \mu, f \rangle$.
iii) If now $f : \mathbb{R} \to \mathbb{R}$ is $\bar{\mu}$-integrable, then $f = f_+ - f_-$ where $f_+, f_- \geq 0$ are $\bar{\mu}$-integrable. By *ii)*, the non-negative random variables $\langle \mu, f_+ \rangle$ and $\langle \mu, f_- \rangle$

are both $\mathbb{P}$-integrable. Then their difference $\langle \mu, f_+ \rangle - \langle \mu, f_- \rangle = \langle \mu, f \rangle$ is also $\mathbb{P}$-integrable and we obtain with *ii*):

$$\langle \bar{\mu}, f \rangle = \langle \bar{\mu}, f_+ \rangle - \langle \bar{\mu}, f_- \rangle = \mathbb{E}\langle \bar{\mu}, f_+ \rangle - \mathbb{E}\langle \bar{\mu}, f_- \rangle = \mathbb{E}\langle \bar{\mu}, f \rangle .$$

iv) Unfortunately, this point appears to be overlooked in the literature. We need to construct a counter-example to show what we state. To this end, consider the random probability measure μ on $(\mathbb{R}, \mathcal{B})$ with

$$\forall\, n \in \mathbb{N} : \mathbb{P}\left(\mu = \frac{1}{2}\delta_{-n} + \frac{1}{2}\delta_n\right) = \frac{1}{cn^2},$$

where $c := \sum_n \frac{1}{n^2} < \infty$. Further, let f be the identity on $\mathbb{R}$, that is, $f(x) = x$ for all $x \in \mathbb{R}$. Then surely, f is measurable, and since almost all realizations of μ are symmetric measures, we have $\langle \mu, f \rangle = 0$ almost surely, which is $\mathbb{P}$-integrable with $\mathbb{E}\langle \mu, f \rangle = 0$. We now assume that f is $\bar{\mu}$-integrable and lead this to a contradiction: If f were $\bar{\mu}$-integrable, then so would $|f|$ and by *ii*) we would have $\langle \bar{\mu}, |f| \rangle = \mathbb{E}\langle \mu, |f| \rangle < \infty$. But with probability $\frac{1}{cn^2}$, μ takes the value $\frac{1}{2}\delta_{-n} + \frac{1}{2}\delta_n$, so $\langle \mu, |f| \rangle$ takes the value n, leading to the calculation

$$\mathbb{E}\langle \mu, |f| \rangle = \sum_{n \in \mathbb{N}} \frac{n}{cn^2} = \infty,$$

which is a contradiction. □

In the remainder of this section, we will derive and discuss three notions of convergence of random probability measures on $(\mathbb{R}, \mathcal{B})$, namely weak convergence in expectation, weak convergence in probability and weak convergence almost surely.

Definition 2.21. *Let $(\mu_n)_{n\in\mathbb{N}}$ and μ be random probability measures on $(\mathbb{R}, \mathcal{B})$, then we say that $(\mu_n)_n$ converges weakly in expectation to μ, if the sequence of expected measures $(\mathbb{E}\mu_n)_{n\in\mathbb{N}}$ converges weakly to the expected measure $\mathbb{E}\mu$, so if:*

$$\forall\, f \in \mathcal{C}_b(\mathbb{R}) : \langle \mathbb{E}\mu_n, f \rangle \xrightarrow[n\to\infty]{} \langle \mathbb{E}\mu, f \rangle ,$$

which is equivalent to (see Theorem 2.20)

$$\forall\, f \in \mathcal{C}_b(\mathbb{R}) : \mathbb{E}\langle \mu_n, f \rangle \xrightarrow[n\to\infty]{} \mathbb{E}\langle \mu, f \rangle .$$

The concept of weak convergence in expectation is extremely important for investigations in the field of random matrix theory, since it lies the foundation for stronger convergence types. This is due to the fact that weak convergence $\mathbb{P}$-almost surely or in probability will also imply weak convergence in expectation, so the latter convergence type is a necessary condition for stronger convergence types (see also Theorem 3.9). The exact interrelations between the three concepts of convergence for random probability measures are summarized in the end of this section in Theorem 2.29.

Before turning to the next convergence types, we wish to remind the reader what convergence in probability and almost surely means for random variables in separable metric spaces, which we implicitly assume to be equipped with Borel-σ-algebras:

Definition 2.22. *Let $(Y_n)_{n\in\mathbb{N}}$ and Y be random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, which take values in a separable metric space $(\mathcal{X}, d)$.*

i) We say that $(Y_n)_{n\in\mathbb{N}}$ converges to Y in probability, if $d(Y_n, Y)$ converges to 0 in probability.
ii) We say that $(Y_n)_{n\in\mathbb{N}}$ converges to Y almost surely, if $d(Y_n, Y)$ converges to 0 almost surely.

Let us collect a quick lemma:

Lemma 2.23. *Let $(Y_n)_{n\in\mathbb{N}}$ and Y be random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, which take values in a separable metric space $(\mathcal{X}, d)$. If $(Y_n)_{n\in\mathbb{N}}$ converges to Y almost surely, then also in probability.*

Proof. Let $(Y_n)_{n\in\mathbb{N}}$ converge to Y almost surely. This means that the sequence of real-valued random variables $(d(Y_n, Y))_n$ converges to 0 almost surely. But this implies that $(d(Y_n, Y))_n$ converges to 0 in probability, which is precisely what it means for $(Y_n)_n$ to converge to Y in probability. □

Now let us define and analyze what it means for random probability measures to converge in probability and almost surely. Since random probability measures are nothing but random variables into the separable metric space $\mathcal{M}_1(\mathbb{R})$, we know what to do:

Definition 2.24. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, μ and $(\mu_n)_{n\in\mathbb{N}}$ be random probability measures on $(\mathbb{R}, \mathcal{B})$.*

i) We say that $(\mu_n)_n$ converges weakly to μ in probability, if $d_M(\mu_n, \mu)$ converges to 0 in probability.
ii) We say that $(\mu_n)_n$ converges weakly to μ almost surely, if $d_M(\mu_n, \mu)$ converges to 0 almost surely.

Although stochastic types of weak convergence can be defined solidly as in Definition 2.24, this definition is not convenient to work with in practice. In addition, we would like to see that these convergence concepts do *not* depend on the choice of the metric that metrizes weak convergence on $\mathcal{M}_1(\mathbb{R})$.

Theorem 2.25. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, μ and $(\mu_n)_{n\in\mathbb{N}}$ be random probability measures on $(\mathbb{R}, \mathcal{B})$.*

i) The following statements are equivalent:
   a) $(\mu_n)_n$ converges weakly to μ in probability, that is, $d_M(\mu_n, \mu) \to 0$ in probability.
   b) If d is any metric on $\mathcal{M}_1(\mathbb{R})$ that metrizes weak convergence, then $d(\mu_n, \mu) \to 0$ in probability.

c) For all $f \in \mathcal{C}_b(\mathbb{R})$, the sequence of bounded real-valued random variables $(\langle \mu_n, f\rangle)_n$ converges in probability to $\langle \mu, f\rangle$, so

$$\forall\, f \in \mathcal{C}_b(\mathbb{R}) : \forall\, \varepsilon > 0 : \mathbb{P}(|\langle \mu_n, f\rangle - \langle \mu, f\rangle| > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

ii) The following statements are equivalent:

a) $(\mu_n)_n$ converges weakly to μ almost surely, that is, $d_M(\mu_n, \mu) \to 0$ almost surely.

b) For $\mathbb{P}$-almost all $\omega \in \Omega$, $\mu_n(\omega)$ converges weakly to $\mu(\omega)$.

c) If d is any metric on $\mathcal{M}_1(\mathbb{R})$ that metrizes weak convergence, then $d(\mu_n, \mu) \to 0$ almost surely.

d) For all $f \in \mathcal{C}_b(\mathbb{R})$, $\langle \mu_n, f\rangle$ converges almost surely to $\langle \mu, f\rangle$, that is,

$$\forall\, f \in \mathcal{C}_b(\mathbb{R}) : \left[\langle \mu_n, f\rangle \xrightarrow[n\to\infty]{} \langle \mu, f\rangle \ \text{ almost surely}\right].$$

e) Almost surely we find that for all $f \in \mathcal{C}_b(\mathbb{R})$, $\langle \mu_n, f\rangle$ converges to $\langle \mu, f\rangle$, that is,

$$\left[\forall\, f \in \mathcal{C}_b(\mathbb{R}) : \langle \mu_n, f\rangle \xrightarrow[n\to\infty]{} \langle \mu, f\rangle\right] \quad \text{almost surely.}$$

Remark 2.26. *1. Note that in Theorem 2.25 ii) d) and e) we used careful bracketing $[\ldots]$ when it comes to almost sure convergence of multiple objects. This is done to avoid ambiguity. For example, questions could arise whether we find a set of measure 1 on which all objects converge (as in e)), or if for each object, we find a set of measure 1,* possibly depending on that object*, on which the considered object converges (as in d)).*

2. We consider Theorem 2.25 i) as equivalent definitions for the concept "weak convergence in probability", and ii) as equivalent definitions for "weak convergence almost surely." After the proof of the theorem, we will keep on working with this characterization without always referring to Theorem 2.25.

Before we begin with the proof of Theorem 2.25, we will introduce two tools which we will make use of. For later use, we will formulate the lemmas in greater generality, that is, for complex-valued random variables.

Lemma 2.27. *Let $(X_n)_n$ and X be complex-valued random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then $(X_n)_{n\in\mathbb{N}}$ converges to X in probability iff any subsequence $J \subseteq \mathbb{N}$ has another subsequence $I \subseteq J$ so that $(X_n)_{n\in I}$ converges to X almost surely.*

Proof. The proof can be found in [18, p. 58]. □

The next extremely useful lemma generalizes the previous one by finding a *simultaneous almost surely convergent subsequence* for a countable number of sequences of random variables.

Lemma 2.28. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and for all $k \in \mathbb{N}$ let $X^{(k)}$ and $(X_n^{(k)})_{n\in\mathbb{N}}$ be complex-valued random variables. Then the following statements are equivalent:*

i) For all $k \in \mathbb{N}$, $(X_n^{(k)})_n$ converges to $X^{(k)}$ in probability.
ii) For any subsequence $J \subseteq \mathbb{N}$, we find a subsequence $I \subseteq J$ and a set $N \in \mathcal{A}$ with $\mathbb{P}(N) = 0$ such that

$$\forall\, \omega \in \Omega\backslash N : \forall\, k \in \mathbb{N} : X_n^{(k)}(\omega) \xrightarrow[n\in I]{} X^{(k)}(\omega).$$

Proof. The part $ii) \Rightarrow i)$ follows immediately with Lemma 2.27. So we only need to show $i) \Rightarrow ii)$: For $k = 1$ we find that $(X_n^{(1)})_{n\in J}$ converges in probability to $X^{(1)}$. Therefore, we find a subsequence $I_1 \subseteq J$ such that

$$X_n^{(1)} \xrightarrow[n\in I_1]{} X^{(1)} \quad \mathbb{P}\text{-a.s. witnessed by a set of measure zero } N_1.$$

Since $(X_n^{(2)})_{n\in I_1}$ converges to $X^{(2)}$ in probability, we find a subsequence $I_2 \subseteq I_1$ with $\min(I_2) > \min(I_1)$ such that

$$X_n^{(2)} \xrightarrow[n\in I_2]{} X^{(2)} \quad \mathbb{P}\text{-a.s. witnessed by a set of measure zero } N_2.$$

We continue this approach for all $k \in \mathbb{N}$ and obtain subsequences

$$\mathbb{N} \supseteq J \supseteq I_1 \supseteq I_2 \supseteq \ldots \supseteq I_k \supseteq \ldots$$

such that for all $k \in \mathbb{N}$ we have $\min(I_{k+1}) > \min(I_k)$ and

$$X_n^{(k)} \xrightarrow[n\in I_k]{} X^{(k)} \quad \mathbb{P}\text{-a.s. witnessed by a set of measure zero } N_k.$$

We set $N \coloneqq \cup_{k\in\mathbb{N}} N_k$ and for all $k \in \mathbb{N} : i_k \coloneqq \min(I_k)$, then we obtain that $(i_k)_{k\in\mathbb{N}}$ is strictly increasing in $\mathbb{N}$ and

$$\forall\, \omega \in \Omega\backslash N : \forall\, l \in \mathbb{N} : X_{i_k}^{(l)}(\omega) \xrightarrow[k\in\mathbb{N}]{} X^{(l)}(\omega).$$

To see this, let $\omega \in \Omega\backslash N$ and $l \in \mathbb{N}$ be arbitrary. Then we have that $\omega \in \Omega\backslash N_l$ and $i_k = \min(I_k) \in I_l$ for all $k \geq l$, so that indeed

$$X_{i_k}^{(l)}(\omega) \xrightarrow[k\in\mathbb{N}]{} X^{(l)}(\omega).$$

The proof is completed by setting $I \coloneqq \{i_k \,|\, k \in \mathbb{N}\}$. □

Now we are ready to prove Theorem 2.25:

Proof of Theorem 2.25. We show *ii)* first.

Clearly, *a)*, *b)* and *c)* are equivalent, since the metrics metrize weak convergence. Also, *e)* is just a reformulation of *b)*, thus equivalent. In addition, *d)* follows immediately from *e)*, so we have

$$a) \;\Leftrightarrow\; b) \;\Leftrightarrow\; c) \;\Leftrightarrow\; e) \;\Rightarrow\; d)$$

We now show *d)* $\Rightarrow$ *b)*: For each $k \in \mathbb{N}$ we have that $\langle \mu_n, g_k \rangle$ converges to $\langle \mu, g_k \rangle$ almost surely on a set A_k of measure 1 (the functions $(g_k)_k$ are as in Theorem 2.16). Then the set $\Omega_1 \coloneqq \cap_k A_k$ has measure 1 and for all $\omega \in \Omega_1$ we find that

$$\forall\, k \in \mathbb{N} : \langle \mu_n(\omega), g_k \rangle \xrightarrow[n\to\infty]{} \langle \mu(\omega), g_k \rangle\,.$$

Therefore, with Theorem 2.16, we have for all $\omega \in \Omega_1$ that $\mu_n(\omega) \to \mu(\omega)$ weakly as $n \to \infty$ and hence *b)*.

We now show *i)*:

a) $\Leftrightarrow$ *b)* By exact symmetry in the argument, we will only argue *a)* $\Rightarrow$ *b)*: Let $\mu_n \to \mu$ weakly in probability, that is, $(d_M(\mu_n, \mu))_{n\in\mathbb{N}}$ converges to 0 in probability. We want to show that also $(d(\mu_n, \mu))_{n\in\mathbb{N}}$ converges to 0 in probability. To use Lemma 2.27, let $J \subseteq \mathbb{N}$ be an arbitrary subsequence. Then we find a subsequence $I \subseteq J$ such that $(d_M(\mu_n, \mu))_{n\in I}$ converges to 0 almost surely. With part *ii)* this means that also $(d(\mu_n, \mu))_{n\in I}$ converges to 0 almost surely. But then $(d(\mu_n, \mu))_{n\in\mathbb{N}}$ converges to 0 in probability.

a) $\Rightarrow$ *c)* If $(\mu_n)_n$ converges weakly to μ in probability, then this means that $d_M(\mu_n, \mu)$ converges to 0 in probability. Let $f \in \mathcal{C}_b(\mathbb{R})$ be arbitrary. We must show that $\langle \mu_n, f \rangle$ converges to $\langle \mu, f \rangle$ in probability. To this end, let $J \subseteq \mathbb{N}$ be an arbitrary subsequence. Then there is a subsequence $I \subseteq J$ such that $(d_M(\mu_n, \mu))_{n\in I}$ converges to 0 almost surely on a measurable subset $\Omega_1 \subseteq \Omega$ with measure 1. Then it holds in particular for any $\omega \in \Omega_1$ that $(\langle \mu_n(\omega), f \rangle)_{n\in I}$ converges to $\langle \mu(\omega), f \rangle$, so $(\langle \mu_n, f \rangle)_{n\in I}$ converges to $\langle \mu, f \rangle$ almost surely. The statement follows with Lemma 2.27.

c) $\Rightarrow$ *a)* We find that for all $k \in \mathbb{N}$, $(\langle \mu_n, g_k \rangle)_{n\in\mathbb{N}}$ converges to $\langle \mu, g_k \rangle$ in probability. We must show that $d_M(\mu_n, \mu)$ converges to zero in probability. Let $J \subseteq \mathbb{N}$ be any subsequence. With Lemma 2.28, we find a subsequence $I \subseteq J$ and a measurable set $\Omega_1 \subseteq \Omega$ of measure 1, such that

$$\forall\, \omega \in \Omega_1 : \forall\, k \in \mathbb{N} : \langle \mu_n(\omega), g_k \rangle \xrightarrow[n\in I]{} \langle \mu(\omega), g_k \rangle$$

With Theorem 2.16, this entails that for all $\omega \in \Omega_1$, $(d_M(\mu_n(\omega), \mu(\omega)))_{n\in I}$ converges to 0. With Lemma 2.27, this means that $(d_M(\mu_n, \mu))_{n\in\mathbb{N}}$ converges to zero in probability. □

So, what we have seen so far is that random probability measures can converge in three different ways, namely weakly in expectation, weakly in probability and weakly almost surely. We have solidly defined and then characterized these convergence concepts. At last, we point out a hierarchy among them:

Theorem 2.29. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mu_n)_{n\in\mathbb{N}}$ and μ be random probability measures on $(\mathbb{R}, \mathcal{B})$.*

i) If $\mu_n \to \mu$ weakly almost surely, then also weakly in probability.
ii) If $\mu_n \to \mu$ weakly in probability, then also weakly in expectation.

Proof. *i)* This follows directly with Lemma 2.23.

ii) If $\mu_n \to \mu$ weakly in probability, per Theorem 2.25 this means that for all

$f \in \mathcal{C}_b(\mathbb{R})$ we find $\langle \mu_n, f\rangle \to \langle \mu, f\rangle$ in probability, thus $\mathbb{E}\langle \mu_n, f\rangle \to \mathbb{E}\langle \mu, f\rangle$ by the following Lemma 2.30, since $|\langle \mu_n, f\rangle| \leq \|f\|_\infty$ and $|\langle \mu, f\rangle| \leq \|f\|_\infty$. □

Lemma 2.30. *Let $(X_n)_{n\in\mathbb{N}}$ and X be complex-valued random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $C \in \mathbb{R}$ such that $|X_n| \leq C$ for all $n \in \mathbb{N}$ and $|X| \leq C$. Then $X_n \to X$ in probability implies $\mathbb{E}|X_n - X| \to 0$, in particular $\mathbb{E}X_n \to \mathbb{E}X$.*

Proof. Let $\varepsilon > 0$ be arbitrary, then we calculate:

$$\begin{aligned}\mathbb{E}|X_n - X| &= \mathbb{E}|X_n - X|\mathbb{1}_{\{|X_n - X| \leq \varepsilon\}} + \mathbb{E}|X_n - X|\mathbb{1}_{\{|X_n - X| > \varepsilon\}} \\ &\leq \varepsilon + \mathbb{P}(|X_n - X| > \varepsilon) \cdot 2C.\end{aligned}$$

Therefore, we conclude

$$\limsup_{n\to\infty} \mathbb{E}|X_n - X| \leq \varepsilon.$$

□

2.4. Limit laws in random matrix theory

We will now introduce the types of random probability measures which we would like to investigate, namely the empirical spectral distribution of random matrices. To this end, let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and denote by $(\mathrm{Mat}_n(\mathbb{K}), \|\cdot\|_{\mathrm{op}})$ the normed $\mathbb{K}$-vector space of $n \times n$-matrices with $\mathbb{K}$-valued entries, where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm with respect to the euclidian norm $\|\cdot\|$ on $\mathbb{K}^n$, that is,

$$\forall\, X \in \mathrm{Mat}_n(\mathbb{K}) : \|X\|_{\mathrm{op}} = \sup\{\|Xv\| : v \in \mathbb{K}^n, \|v\| = 1\}.$$

It is immediate that $(\mathrm{Mat}_n(\mathbb{K}), \|\cdot\|_{\mathrm{op}})$ is a Banach-space, and a sequence of matrices $(X_m)_m$ converges to a matrix X in $\mathrm{Mat}_n(\mathbb{K})$ iff all entries $X_m(i,j)$ converge to $X(i,j)$ in $\mathbb{K}$ as $m \to \infty$. If $X \in \mathrm{Mat}_n(\mathbb{K})$ we denote its *adjoint* by X^*, which is just the transpose of X if $\mathbb{K} = \mathbb{R}$ and the conjugate transpose of X if $\mathbb{K} = \mathbb{C}$. A matrix $X \in \mathrm{Mat}_n(\mathbb{K})$ is called *self-adjoint* if $X^* = X$ (then X is also called *symmetric* if $\mathbb{K} = \mathbb{R}$ and *Hermitian* if $\mathbb{K} = \mathbb{C}$) and we denote the subset of all self-adjoint matrices of $\mathrm{Mat}_n(\mathbb{K})$ by $\mathrm{SMat}_n(\mathbb{K})$. Then $\mathrm{SMat}_n(\mathbb{K}) \subseteq \mathrm{Mat}_n(\mathbb{K})$ is a closed subset, since $X \mapsto X^*$ is continuous. Further, $\mathrm{SMat}_n(\mathbb{K})$ is closed under $\mathbb{R}$-linear combinations. To introduce more notation, if $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are arbitrary, we denote by $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ the diagonal matrix $D \in \mathrm{SMat}_n(\mathbb{K})$ with entries $D(i,i) = \lambda_i$ for all $i \in \{1, \ldots, n\}$. Further, we denote by tr the trace functional $\mathrm{Mat}_n(\mathbb{K}) \longrightarrow \mathbb{K}$, that is,

$$\forall\, X \in \mathrm{Mat}_n(\mathbb{K}) : \mathrm{tr}\, X = \sum_{t=1}^{n} X(t,t).$$

The trace has some interesting properties, which are summarized in the following lemma:

Lemma 2.31. *The trace* tr *is a continuous linear functional on $(Mat_n(\mathbb{K}), \|\cdot\|_{\mathrm{op}})$. Further, if $X, S \in Mat_n(\mathbb{K})$ are arbitrary, where S is invertible, then* $\mathrm{tr}(X) = \mathrm{tr}(S^{-1}XS)$.

Proof. It is immediate that the trace is a continuous linear functional. The equality $\mathrm{tr}(X) = \mathrm{tr}(S^{-1}XS)$ is due to the fact that X and $S^{-1}XS$ have the same characteristic polynomial. The trace is the $(n-1)$th coefficient of the characteristic polynomial (multiplied by $(-1)^{n-1}$). For details we refer the reader to [44] or [24]. □

The next lemma clarifies the eigenvalue structure of self-adjoint matrices:

Lemma 2.32. *For any matrix $X \in SMat_n(\mathbb{K})$ we find an invertible matrix $S \in Mat_n(\mathbb{K})$ and real numbers $\lambda_1^X \leq \ldots \leq \lambda_n^X$, such that $S^{-1}XS = \mathrm{diag}(\lambda_1^X, \ldots, \lambda_n^X)$. In particular, X has exactly n real eigenvalues (counting multiplicities), and all eigenvalues are real.*

Proof. We refer the reader to [44] or [24]. □

In general, if Y is a self-adjoint $n \times n$ matrix, we will denote its n real eigenvalues by $\lambda_1^Y \leq \ldots \leq \lambda_n^Y$. The next theorem is a very versatile tool in random matrix theory. For example, it can be used to derive that eigenvalues are continuous functions of the entries of the matrix (Corollary 2.34), or it can be used to analyze asymptotic equivalence of empirical spectral distributions via the bounded Lipschitz metric.

Theorem 2.33 (Hoffman-Wielandt)**.** *For all $n \in \mathbb{N}$ and $X, Y \in SMat_n(\mathbb{K})$ it holds:*

$$\sum_{i=1}^{n} |\lambda_i^X - \lambda_i^Y|^2 \leq \mathrm{tr}(X-Y)^*(X-Y) = \mathrm{tr}(X-Y)^2.$$

Proof. See [62, p. 320] or [32, p. 217]. □

We can immediately conclude that eigenvalues are continuous functions of the matrices.

Corollary 2.34. *Let $n \in \mathbb{N}$ and $l \in \{1, \ldots, n\}$ be arbitrary, then*

$$\begin{aligned} \mathrm{Eig}_l : SMat_n(\mathbb{K}) &\longrightarrow \mathbb{R} \\ X &\longmapsto \lambda_l^X \end{aligned}$$

is continuous.

Proof. Let $(X_m)_{m\in\mathbb{N}}$ and X in $\mathrm{SMat}_n(\mathbb{K})$ so that $X_m \to X$ for $m \to \infty$ (which means convergence in operator norm, or equivalently, entry-wise convergence). Then we find with Theorem 2.33 and Lemma 2.31 that

$$|\lambda_l^{X_m} - \lambda_l^X|^2 \leq \sum_{i=1}^{N} |\lambda_i^{X_m} - \lambda_i^X|^2 \leq \mathrm{tr}(X_m - X)^2 \xrightarrow[m\to\infty]{} 0.$$ □

Having studied eigenvalues of self-adjoint matrices, let us turn our attention to random matrices.

Definition 2.35. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $n \in \mathbb{N}$ be arbitrary then a ($n \times n$ self-adjoint) random matrix is a measurable map $X : (\Omega, \mathcal{A}) \to (SMat_n(\mathbb{K}), \mathcal{B}_s^{(n^2)})$, where $\mathcal{B}_s^{(n^2)}$ denotes Borel σ-algebra on $SMat_n(\mathbb{K})$.*

From here on out, unless stated otherwise, we will always understand a *random matrix* to be self-adjoint and with entries in $\mathbb{K}$ if the underlying field is not further specified. It is clear that a map $X : (\Omega, \mathcal{A}) \to (\mathrm{SMat}_n(\mathbb{K}), \mathcal{B}_s^{(n^2)})$ is measurable iff all entries $X(i,j) : (\Omega, \mathcal{A}) \to (\mathbb{K}, \mathcal{B}_{\mathbb{K}})$ are measurable, where $\mathcal{B}_{\mathbb{K}}$ denotes the Borel σ-algebra on $\mathbb{K}$. If X is an $n \times n$ random matrix on $(\Omega, \mathcal{A}, \mathbb{P})$, then for all $\omega \in \Omega$, $X(\omega) \in \mathrm{SMat}_n(\mathbb{K})$, such that $X(\omega)$ possesses eigenvalues $\lambda_1^{X(\omega)} \leq \ldots \leq \lambda_n^{X(\omega)}$. We wish to see that the maps $\omega \mapsto \lambda_l^{X(\omega)}$ for $l = 1, \ldots, n$ are measurable.

Lemma 2.36. *Let X be an $n \times n$ random matrix on $(\Omega, \mathcal{A}, \mathbb{P})$ and $l \in \{1, \ldots, n\}$ be arbitrary, then*

$$\begin{aligned} \lambda_l^X : (\Omega, \mathcal{A}) &\longrightarrow (\mathbb{R}, \mathcal{B}) \\ \omega &\longmapsto \lambda_l^{X(\omega)} \end{aligned}$$

is measurable, thus a real-valued random variable.

Proof. We know by Corollary 2.34 that

$$\begin{aligned} \mathrm{Eig}_l : \mathrm{SMat}_n(\mathbb{K}) &\longrightarrow \mathbb{R} \\ X &\longmapsto \lambda_l^X \end{aligned}$$

is continuous, in particular measurable. Further, $X : \Omega \longrightarrow \mathrm{SMat}_n(\mathbb{K})$ is measurable per definition, hence the composition $\lambda_l^X \coloneqq \mathrm{Eig}_l \circ X$ is measurable as well. □

Lemma 2.36 allows us to study eigenvalues of random matrices in the context of probability theory. One aspect which gains a lot of attention is the behavior of the empirical distribution of the eigenvalues (see also Example 2.19).

Definition 2.37. *Let X be an $n \times n$ random matrix on $(\Omega, \mathcal{A}, \mathbb{P})$, then the* empirical spectral distribution (ESD) *σ_n of X is the random probability measure on $(\mathbb{R}, \mathcal{B})$ given by*

$$\begin{aligned} \sigma_n : \Omega \times \mathcal{B} &\longrightarrow [0,1] \\ (\omega, B) &\longmapsto \sigma_n(\omega, B) \coloneqq \frac{1}{n} \sum_{l=1}^{n} \delta_{\lambda_l^{X(\omega)}}(B) \end{aligned}$$

It follows from our discussion in Example 2.19 that σ_n really is a random probability measure. How is σ_n to be interpreted? For any interval $I \subseteq \mathbb{R}$, the random variable $\sigma_n(I)$ tells us the proportion of the n eigenvalues that fall into the interval I. Thus, σ_n carries information on the location of the eigenvalues, and it is of particular interest where the eigenvalues are located in the limit, that is, for $n \to \infty$.

Wigner's semicircle law

It is a famous theorem by Wigner that allows us to conclude under fairly weak assumptions (mainly independence of matrix entries and uniformly bounded moments) that in the limit, eigenvalues will be spread according to the semicircle distribution:

Definition 2.38. *The semicircle distribution σ is the probability measure on $(\mathbb{R}, \mathcal{B})$ with Lebesgue-density f_σ where*

$$\begin{aligned} f_\sigma : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto f_\sigma(x) := \frac{1}{2\pi}\sqrt{4-x^2}\mathbb{1}_{[-2,2]}(x). \end{aligned}$$

Here and throughout this text, we will denote the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ by λ. With respect to Definition 2.38, we have to prove that $f_\sigma \lambda$ is actually a probability measure. We see immediately that the measure is finite, since f_σ is bounded and has compact support. We will postpone the proof that the Lebesgue integral over f_σ is 1 to Lemma 3.11. Since convergence to the semicircle distribution is an important and ubiquitous concept, we make the following definition.

Definition 2.39. *If $(\sigma_n)_n$ are the ESDs of random matrices $(X_n)_n$ and $\sigma_n \to \sigma$ weakly in expectation resp. in probability resp. almost surely, then we say that* the semicircle law holds for $(X_n)_n$ *in expectation resp. in probability resp. almost surely.*

We now turn to Wigner's semicircle law. Notationally, for all $n \in \mathbb{N}$ we define the index set $[n]^2 := [n] \times [n] = \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$.

Definition 2.40. *Let for all $n \in \mathbb{N}$, $X_n = (X_n(i,j))_{(i,j)\in[n]^2}$ be a family of real-valued random variables, then the sequence $(X_n)_n$ is called* Wigner scheme, *if the following holds:*

i) All random variables have uniformly bounded absolute moments, that is: For all $q \in \mathbb{N}$ there exists a constant $L_q \in (0, \infty)$ such that for all $n \in \mathbb{N}$ and all $(i,j) \in [n]^2$: $\mathbb{E}|X_n(i,j)|^q \leq L_q$.

ii) All random variables are standardized, that is: For all $n \in \mathbb{N}$ and all $(i,j) \in [n]^2$: $\mathbb{E}X_n(i,j) = 0$ and $\mathbb{V}X_n(i,j) = 1$.

iii) The families X_n are symmetric, that is: For all $n \in \mathbb{N}$ and $(i,j) \in [n]^2$ we have $X_n(i,j) = X_n(j,i)$.

iv) For all $n \in \mathbb{N}$ the family $(X_n(i,j))_{1\leq i\leq j\leq n}$ is independent.

Note in particular that in Definition 2.40 we *do not* require that the whole family $((X_n(i,j))_{1\leq i\leq j\leq n})_{n\in\mathbb{N}}$ be independent. A very simple Wigner scheme is given in the following example:

Example 2.41. *Let $(X(i,j))_{1\leq i\leq j}$ be an i.i.d. family of real-valued random variables such that $\mathbb{E}|X(1,1)|^q < \infty$ for all $q \in \mathbb{N}$, $\mathbb{E}X(1,1) = 0$ and $\mathbb{V}X(1,1) = 1$. Further, set $X(i,j) := X(j,i)$ for all $1 \leq j < i$. Now set for all $n \in \mathbb{N}$*

and all $(i,j) \in [n]^2$*:* $X_n(i,j) \coloneqq X(i,j)$*. Roughly speaking,* X_n *is the* $n \times n$ *submatrix of the infinite matrix* X*. Then clearly,* $(X_n)_n$ *is a Wigner scheme as in Definition 2.40.*

The following Theorem is called "Wigner's semicircle law."

Theorem 2.42. *Let* $(X_n)_n$ *be a Wigner scheme defined on a probability space* $(\Omega, \mathcal{A}, \mathbb{P})$*. Define for all* $n \in \mathbb{N}$ *the Wigner matrix* W_n *by*

$$\forall\,(i,j) \in [n]^2 : \; W_n(i,j) \coloneqq \frac{1}{\sqrt{n}} X_n(i,j).$$

Then the semicircle law holds for $(W_n)_n$ *almost surely.*

We will prove Theorem 2.42 in various ways: In Section 4.2 we will employ the method of moments to prove this theorem, whereas in Section 6.2 we use the Stieltjes transform method.

The Marchenko-Pastur law

Another class of random matrix models besides the Wigner schemes fall into the category of covariance matrices. Assume we have n observations $x_1, \dots, x_n$, each with p real-valued covariates, where $n, p \in \mathbb{N}$, so that $x_i = (x_i(1), \dots, x_i(p))^T$ for all $i \in \{1, \dots, n\}$. Define the $p \times n$ data matrix $X_n \coloneqq (x_1 | x_2 | \dots | x_n)$. The sample covariance matrix is then defined by

$$\tilde{S}_n \coloneqq \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})^T \quad \left(= \frac{n}{n-1} \cdot \left(\frac{1}{n} \sum_{k=1}^{n} x_k x_k^T - \bar{x}\bar{x}^T \right) \right),$$

which is of dimension $p \times p$. Here, the vector $\bar{x}$ denotes the arithmetic mean of the vectors x_k. Assuming that the data stems from n i.i.d. realizations of an $\mathbb{R}^p$-valued random vector X with $\mathcal{L}_2$-entries, $\tilde{S}_n$ is an unbiased estimator for its covariance matrix

$$\mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T = \begin{pmatrix} \mathbb{V}X(1) & \cdots & \operatorname{Cov}(X(1), X(p)) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X(p), X(1)) & \cdots & \mathbb{V}X(p) \end{pmatrix}.$$

Many test statistics are based on the eigenvalues of the sample covariance matrix. When analyzing these eigenvalues in the limit, it suffices to consider

$$S_n \coloneqq \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T = \frac{1}{n} X_n X_n^T, \tag{2.4}$$

since $\bar{x}\bar{x}^T$ is of rank 1. Also, we will assume that the number of covariates p grows with the number of observations n, so $p = p_n$. We further assume that $p/n \longrightarrow y \in (0, \infty)$, that is, p grows proportionally with n. This leads to the definition of a Marchenko-Pastur scheme.

Definition 2.43. *Let for all $n \in \mathbb{N}$, $p = p_n \in \mathbb{N}$ and $(X_n(i,j))_{i\in[p],j\in[n]}$ be a family of real-valued random variables. Then the sequence $(X_n)_n$ is called* Marchenko-Pastur scheme*, if the following holds:*

i) All random variables have uniformly bounded absolute moments, that is: For all $q \in \mathbb{N}$ there exists a constant $L_q \in (0,\infty)$ such that for all $n \in \mathbb{N}$ and all $(i,j) \in [p] \times [n]$: $\mathbb{E}|X_n(i,j)|^q \leq L_q$.
ii) All random variables are standardized, that is: For all $n \in \mathbb{N}$ and all $(i,j) \in [p] \times [n]$: $\mathbb{E}X_n(i,j) = 0$ and $\mathbb{V}X_n(i,j) = 1$.
iii) For all $n \in \mathbb{N}$ the family $(X_n(i,j))_{i\in[p],j\in[n]}$ is independent.
iv) There exists a constant $y \in (0,\infty)$ such that $p/n \to y$ as $n \to \infty$.

We will see that eigenvalues of covariance matrices which are based on MP-schemes will spread according to the Marchenko-Pastur distribution:

Definition 2.44. *The (standard) MP distribution with ratio index $y \in (0,\infty)$ is the probability measure μ^y on $(\mathbb{R},\mathcal{B})$ given by*

$$\mu^y = \frac{1}{2\pi xy}\sqrt{((1+\sqrt{y})^2 - x)(x-(1-\sqrt{y})^2)}\mathbb{1}_{(0,\infty)}(x)\lambda(dx) + \left(1-\frac{1}{y}\right)\delta_0\mathbb{1}_{y>1},$$

where λ denotes the Lebesgue measure on $(\mathbb{R},\mathcal{B})$ and δ_0 denotes the Dirac measure in 0.

Definition 2.45. *Let $y \in (0,\infty)$. If $(\mu_n)_n$ are the ESDs of random matrices $(V_n)_n$ and $\mu_n \to \mu^y$ weakly in expectation resp. in probability resp. almost surely, then we say that* the Marchenko-Pastur law holds for *$(V_n)_n$ in expectation resp. in probability resp. almost surely.*

The following Theorem is called "Marchenko-Pastur law."

Theorem 2.46. *Let $(X_n)_n$ be an MP-scheme defined on a probability space $(\Omega,\mathcal{A},\mathbb{P})$. Define for all $n \in \mathbb{N}$ the MP-matrix V_n by*

$$V_n := \frac{1}{n}X_nX_n^T.$$

Then the MP-law holds for $(V_n)_n$ almost surely.

We will prove Theorem 2.46 in various ways: In Section 4.3 we will employ the method of moments to prove this theorem, whereas in Section 6.3 we use the Stieltjes transform method.

Outlook

Of course, a valid question is how to prove Theorem 2.42 and Theorem 2.46. We see that certain conditions are formulated for entries of these matrix models. In order to use these conditions in our analysis, how can we relate the ESDs σ_n and μ_n back to the entries of their respective random matrices? And lastly, how can we conclude (stochastic) weak convergence of these ESDs? There are

(at least) two standard ways to achieve this, namely the method of moments and the Stieltjes transform method. These methods will be discussed in depth in the following sections. We will also use these methods to prove the almost sure semicircle law and the Marchenko-Pastur law.

3. The method of moments

In Chapter 2 we have studied in depth the concepts of weak convergence of probability measures and random probability measures. In this chapter we want to discuss a tool which helps us to infer weak convergence: The method of moments. We will carefully develop this method for both deterministic and random probability measures. To be able to use this method correctly, we also need to delve into the moment problem. But let us first define what the moments of a measure are:

Definition 3.1. *Let μ be a measure on $(\mathbb{R}, \mathcal{B})$ and $k \in \mathbb{N}_0$. If $\langle \mu, |x^k| \rangle < \infty$ (where $x^0 = 1 \, \forall \, x \in \mathbb{R}$) we call the real number*

$$m_k \coloneqq \langle \mu, x^k \rangle$$

the k-th moment of μ. *In this case, we say that μ has a finite k-th moment. On the other hand, if $\langle \mu, |x^k| \rangle = \infty$, we say the k-th moment of μ does not exist.*

3.1. The moment problem

In numerous applications it is important to know the moments of a probability measure or at least some properties of the moments. In the Hamburger moment problem (see [43, p. 145] and [49], for example), the question is reversed. Given a sequence of real numbers $(m_k)_{k \in \mathbb{N}_0}$, what can be said about the existence and uniqueness of a measure μ on $(\mathbb{R}, \mathcal{B})$ with moments $(m_k)_{k \in \mathbb{N}_0}$? To be more precise, does there exist a measure μ on $(\mathbb{R}, \mathcal{B})$ with moments $(m_k)_{k \in \mathbb{N}_0}$, and if so, is it the only measure with those moments? Of course, if such a measure exists, it is a probability measure iff $m_0 = 1$. It is rather surprising that the existence of such a measure can be nicely characterized:

Theorem 3.2. *A sequence of real numbers $(m_k)_{k \in \mathbb{N}_0}$ constitutes the moments of at least one measure on $(\mathbb{R}, \mathcal{B})$, if and only if for all $N \in \mathbb{N}$ the corresponding Hankel matrix*

$$\begin{pmatrix} m_0 & m_1 & m_2 & \dots & m_N \\ m_1 & m_2 & m_3 & \dots & m_{N+1} \\ m_2 & m_3 & m_4 & \dots & m_{N+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_N & m_{N+1} & m_{N+2} & \dots & m_{2N} \end{pmatrix}$$

is positive semi-definite, that is, if for all $N \in \mathbb{N}_0$ and all $\beta_0, \dots, \beta_N \in \mathbb{R}$ it holds:

$$\sum_{r,s=0}^{N} \beta_r \beta_s m_{r+s} \geq 0.$$

Proof. See [43, p. 145] in combination with the fact that a real symmetric matrix is positive definite in the real sense iff it is positive definite in the complex sense. □

Oftentimes it will not be of interest if a sequence of numbers $(m_k)_{k\in\mathbb{N}_0}$ really belongs to a probability measure, since we automatically obtain this result when employing the method of moments, see Theorem 3.5. Theorem 3.2 still has two important applications: On the one hand, if the researcher is a priori assuming the target distribution to have specific moments, Theorem 3.2 can be used to check whether this is a plausible assumption and can spare the researcher from trying to prove convergence to a non-existing probability measure. On the other hand, if one has already employed the method of moments and the moments of the target distribution have been calculated, one can a posteriori evaluate the plausibility of the calculations via Theorem 3.2. Indeed, this is not uncommon practice, see [14, p. 15], for example. In any case, what will be essential for the method of moments is the knowledge about the uniqueness of a distribution with given moments, that is, the answer to the question whether there is *at most* one distribution with a given sequence of moments.

Theorem 3.3. *Let $(m_k)_{k\in\mathbb{N}}$ be a sequence of real numbers. If one of the following three conditions holds, there is at most one probability measure on $(\mathbb{R},\mathcal{B})$ with moments $(m_k)_{k\in\mathbb{N}}$:*

i) $\sum_{k=1}^{\infty}\frac{1}{\sqrt[2k]{m_{2k}}}=\infty$ *(Carleman condition),*
ii) $\limsup_{k\to\infty}\frac{\sqrt[2k]{m_{2k}}}{2k}<\infty$,
iii) $\exists\, C,D\geq 1:\ \forall\, k\in\mathbb{N}:|m_k|\leq C\cdot D^k\cdot k!$.

Further, it holds that iii) $\Rightarrow$ ii) $\Rightarrow$ i), that is, the Carleman condition is the weakest of the three.

Proof. *i)*: See [1, p. 85].
ii): See [18, p. 123].
iii): See [43, p. 205].
Additional statement: The additional statement also proves that *ii)* and *iii)* are sufficient when knowing that *i)* is sufficient.

We assume that *ii)* holds. Let for all $k\in\mathbb{N}:\alpha_k:=\sqrt[2k]{m_{2k}}\geq 0$, then we have to show $\sum_{k=1}^{\infty}\frac{1}{\alpha_k}=\infty$ under the condition that $r:=\limsup_{k\to\infty}\frac{\alpha_k}{2k}<\infty$. But there exists a $K\in\mathbb{N}$ such that for all $k\geq K$ we find $\frac{\alpha_k}{2k}\leq r+1$, thus $\alpha_k\leq 2k\cdot(r+1)$. Due to divergence of the harmonic series we obtain:

$$\sum_{k=1}^{\infty}\frac{1}{\alpha_k}\geq\sum_{k\geq K}\frac{1}{2k\cdot(r+1)}=\infty.$$

Therefore, *i)* follows from *ii)*. Now if *iii)* holds, we find for all $k\in\mathbb{N}$:

$$\frac{\sqrt[2k]{m_{2k}}}{2k}\leq\frac{\sqrt[2k]{C\cdot D^{2k}\cdot(2k)!}}{2k}\leq C\cdot D\cdot\frac{\sqrt[2k]{(2k)!}}{2k}\leq C\cdot D,$$

since $(2k)^{2k}\geq(2k)!$ yields $2k\geq\sqrt[2k]{(2k)!}$ for all $k\in\mathbb{N}$. Thus, *ii)* holds. □

In the next corollary we will see that the moments of probability measures with compact support possess moments of all orders, and that they are uniquely determined by their moments.

Corollary 3.4. *Let ν be a probability measure on $(\mathbb{R}, \mathcal{B})$ with compact support which lies in $[-a, a]$ for some $a \in \mathbb{N}$. Then*

i) ν has moments of all orders.
ii) For all $k \in \mathbb{N}$: $|\langle \nu, x^k \rangle| \leq a^k$.
iii) ν is uniquely determined by its moments.

Proof. We calculate for $k \in \mathbb{N}$ arbitrary:

$$\left|\langle \nu, x^k \rangle\right| = \langle \nu, |x|^k \rangle = \langle \nu, \mathbb{1}_{[-a,a]} |x|^k \rangle \leq a^k.$$

This shows *i)* and *ii)*, and *iii)* follows immediately with Theorem 3.3 *iii)*. □

3.2. The method of moments for probability measures

Now we are well-prepared to introduce the method of moments, which is a means to infer weak convergence of a sequence of distributions from the convergence of their moments.

Theorem 3.5. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{M}_1(\mathbb{R})$, so that all moments of every μ_n exist. If there exists a sequence of real numbers $(m_k)_{k\in\mathbb{N}}$, so that*

$$\forall\, k \in \mathbb{N} : \lim_{n\to\infty} \langle \mu_n, x^k \rangle = m_k, \tag{3.1}$$

the following statements hold:

There exists a $\mu \in \mathcal{M}_1(\mathbb{R})$ and a subsequence of $(\mu_n)_{n\in\mathbb{N}}$, which converges weakly to μ. Then $\forall\, k \in \mathbb{N} : m_k = \langle \mu, x^k \rangle$. In particular, the $(m_k)_{k\in\mathbb{N}}$ are moments of a probability measure on $(\mathbb{R}, \mathcal{B})$. Further: If μ is uniquely determined by its moments, then the entire sequence $(\mu_n)_n$ converges weakly to μ.

Proof. With (3.1) it follows with $k = 2$ and Lemma 2.13 that $(\mu_n)_{n\in\mathbb{N}}$ is tight. Therefore, with Lemma 2.15 there exists a $\mu \in \mathcal{M}_1(\mathbb{R})$ and a subsequence $J \subseteq \mathbb{N}$ such that $(\mu_n)_{n\in J}$ converges weakly to μ. With Lemma 2.11, we then obtain for all $k \in \mathbb{N}$ that $(\langle \mu_n, x^k \rangle)_{n\in J}$ converges to $\langle \mu, x^k \rangle$, since the sequence $(\langle \mu_n, 1 + x^{2k} \rangle)_{n\in J}$ is bounded and the function $x \mapsto \frac{x^k}{1+x^{2k}}$ vanishes at infinity. We conclude with (3.1) that for all $k \in \mathbb{N}$ we have $\langle \mu, x^k \rangle = m_k$, so $(m_k)_k$ are indeed moments of a probability measure.

Now, if μ is uniquely determined by its moments, then the entire sequence $(\mu_n)_{n\in\mathbb{N}}$ – and not just a subsequence – converges weakly to μ. To see this, let $(\mu_n)_{n\in I}$ be an arbitrary subsequence. By Lemma 2.9, it suffices to show that this subsequence has another subsequence that converges weakly to μ. But as above (with swapped roles of I and $\mathbb{N}$) we find a probability measure ν on $(\mathbb{R}, \mathcal{B})$ and a subsequence $J' \subseteq I$, such that that $(\mu_n)_{n\in J'}$ converges weakly to ν and the numbers $(m_k)_{k\in\mathbb{N}}$ are the moments of ν. Since μ is uniquely determined by these moments, we must have $\mu = \nu$. □

Remark 3.6. *A converse statement of Theorem 3.5 is not true in general, that is, there are probability measures $(\mu_n)_n$ and μ with*

1. *All moments of μ and of all μ_n exist.*
2. *μ_n converges weakly to μ.*
3. *The moments of μ_n do not converge to the moments of μ.*

The construction is rather simple: Pick $\mu := \delta_0$ and

$$\forall\, n \in \mathbb{N}: \quad \mu_n := \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_{e^n}$$

Then surely, conditions 1. and 2. are satisfied, but 3. as well, since for all $k \in \mathbb{N}$:

$$\left\langle \mu_n, x^k \right\rangle = \frac{1}{n} e^{kn} \to \infty \neq 0 = \left\langle \mu, x^k \right\rangle.$$

3.3. The method of moments for random probability measures

The next theorem will generalize the method of moments to the convergence types of random probability measures, namely to weak convergence in expectation, in probability and almost surely. Although this could be presented in greater generality, we will restrict our attention to convergence of random probability measures to a *deterministic* probability measure. This is the type of convergence we will encounter in our analyses ahead.

Theorem 3.7. *Let $(\mu_n)_{n\in\mathbb{N}}$ be random probability measures on $(\mathbb{R}, \mathcal{B})$ and μ be a deterministic probability measure on $(\mathbb{R}, \mathcal{B})$ which is uniquely determined by its moments. Then assuming that all following expressions (random moments, expected random moments) are well-defined and finite, we conclude:*

i) If $\forall\, k \in \mathbb{N}: \mathbb{E}\left\langle \mu_n, x^k \right\rangle \xrightarrow[n\to\infty]{} \left\langle \mu, x^k \right\rangle$, then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly in expectation.

ii) If $\forall\, k \in \mathbb{N}: \left\langle \mu_n, x^k \right\rangle \xrightarrow[n\to\infty]{} \left\langle \mu, x^k \right\rangle$ in probability, then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly in probability.

iii) If $\forall\, k \in \mathbb{N}: \left[\left\langle \mu_n, x^k \right\rangle \xrightarrow[n\to\infty]{} \left\langle \mu, x^k \right\rangle \ \mathbb{P}\text{-a.s.}\right]$, then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly almost surely.

Proof. i) With Theorem 3.5 it suffices to show that for all $k \in \mathbb{N}$, $\left\langle \mathbb{E}\mu_n, x^k \right\rangle \to \left\langle \mu, x^k \right\rangle$ as $n \to \infty$. Therefore, all we must argue is that for all $k \in \mathbb{N}$, $\left\langle \mathbb{E}\mu_n, x^k \right\rangle = \mathbb{E}\left\langle \mu_n, x^k \right\rangle$. But for $k \in \mathbb{N}$ arbitrary we find

$$\left\langle \mathbb{E}\mu_n, |x^k| \right\rangle^2 \leq \left\langle \mathbb{E}\mu_n, x^{2k} \right\rangle = \mathbb{E}\left\langle \mu_n, x^{2k} \right\rangle < \infty,$$

where we used Theorem 2.20 ii), the fact that $x \mapsto x^{2k}$ is non-negative, and the assumption in the statement of the theorem that all expected moments exist. Therefore, $\mathbb{E}\mu_n$ has existing moments of all orders, so with Theorem 2.20 iii) we obtain $\left\langle \mathbb{E}\mu_n, x^k \right\rangle = \mathbb{E}\left\langle \mu_n, x^k \right\rangle$.

ii) We want to show that $\mu_n \to \mu$ weakly in probability, which means that for all $f \in \mathcal{C}_b(\mathbb{R})$, $\langle \mu_n, f \rangle$ converges to $\langle \mu, f \rangle$ in probability. To this end, let $f \in \mathcal{C}_b(\mathbb{R})$ be arbitrary. To show that $(\langle \mu_n, f \rangle)_{n\in\mathbb{N}}$ converges to $\langle \mu, f \rangle$ in probability we will show that any subsequence has an almost surely convergent subsequence: Let $J \subseteq \mathbb{N}$ be a subsequence. Applying Lemma 2.28 we find a subsequence $I \subseteq J$ and a measurable set $\Omega_1 \subseteq \Omega$ of measure 1, such that

$$\forall\, \omega \in \Omega_1 : \forall\, k \in \mathbb{N} : \left\langle \mu_n(\omega), x^k \right\rangle \xrightarrow[n\in I]{} \left\langle \mu, x^k \right\rangle .$$

In particular, with Theorem 3.5 we find that for all $\omega \in \Omega_1$, $\mu_n(\omega)$ converges weakly to μ for $n \in I$, so that in particular, $\langle \mu_n(\omega), f \rangle \to \langle \mu, f \rangle$ for $n \in I$. Therefore, $\langle \mu_n, f \rangle \to \langle \mu, f \rangle$ almost surely for $n \in I$.
iii) For all $k \in \mathbb{N}$ we find a measurable set $\Omega_k \subseteq \Omega$ with measure 1 such that for all $\omega \in \Omega_k : \left\langle \mu_n(\omega), x^k \right\rangle \to \left\langle \mu, x^k \right\rangle$ as $n \to \infty$. Then $\Omega' \coloneqq \cap_{k\in\mathbb{N}} \Omega_k$ has measure 1 and for all $\omega \in \Omega'$ we find that $\left\langle \mu_n(\omega), x^k \right\rangle \to \left\langle \mu, x^k \right\rangle$ for all $k \in \mathbb{N}$, so that with Theorem 3.5, for all $\omega \in \Omega'$ we have that $\mu_n(\omega)$ converges weakly to μ. Therefore, μ_n converges weakly to μ almost surely. □

We refer the reader to Remark 2.26 for an explanation on the use of brackets [...] in Theorem 3.7 *iii*).

Remark 3.8. *The method of moments for random probability measures (Theorem 3.7) works as follows: To show weak convergence of random probability measures in expectation, in probability or almost surely, it suffices to show that the random moments converge in expectation, in probability or almost surely. This is a very useful theorem, in particular considering we do not make any assumptions on the target measure μ except those mentioned in Theorem 3.7. In particular, we do not require the target probability measure to have compact support. In the literature on random matrices, this condition is often used to justify the method of moments, see [4, p. 11], for example.*

The next theorem will help us determine when the conditions for Theorem 3.7 are met, to be more precise, when we are able to confirm convergence of the moments in probability or almost surely. Further, it does not assume a priori the knowledge of the target measure $\mu \in \mathcal{M}_1(\mathbb{R})$. In summary, this is the theorem that is used when applying the method of moments to random matrix theory, see also Theorems 3.18 and 3.20.

Theorem 3.9. *Let $(\mu_n)_{n\in\mathbb{N}}$ be random probability measures on $(\mathbb{R}, \mathcal{B})$ and $(m_k)_{k\in\mathbb{N}}$ be a sequence of real numbers, so that there is at most one probability measure on $(\mathbb{R}, \mathcal{B})$ with moments $(m_k)_{k\in\mathbb{N}}$. We formulate the following conditions, where we assume that all expressions (random moments, expectations and variances) are finite:*

(M1) For all $k \in \mathbb{N}$,

$$\mathbb{E}\left\langle \mu_n, x^k \right\rangle \xrightarrow[n\to\infty]{} m_k.$$

For the following assumptions we assume that for all $k \in \mathbb{N}$ we can find a finite decomposition

$$\langle \mu_n, x^k \rangle = D_n^{(k,1)} + \ldots + D_n^{(k,\ell_k)},$$

such that for all $k \in \mathbb{N}$ and all $i \in [\ell_k]$, $\mathbb{E}D_n^{(k,i)}$ converges to a constant as $n \to \infty$.

(M2) For all $k \in \mathbb{N}$ and $i \in [\ell_k]$,

$$\exists\, z \in \mathbb{N} : \ \mathbb{E}\left| D_n^{(k,i)} - \mathbb{E}D_n^{(k,i)} \right|^z \xrightarrow[n\to\infty]{} 0,$$

(M3) For all $k \in \mathbb{N}$ and $i \in [\ell_k]$,

$$\exists\, z \in \mathbb{N} : \ \mathbb{E}\left| D_n^{(k,i)} - \mathbb{E}D_n^{(k,i)} \right|^z \xrightarrow[n\to\infty]{} 0 \quad \text{summably fast.}$$

Then we conclude:

i) If (M1) *holds, then there is a $\mu \in \mathcal{M}_1(\mathbb{R})$ with moments $(m_k)_{k\in\mathbb{N}}$, so that $\mathbb{E}\mu_n \to \mu$ weakly (that is, $\mu_n \to \mu$ weakly in expectation). In particular, the numbers $(m_k)_{k\in\mathbb{N}}$ are the moments of a probability measure.*

ii) If (M1) *and* (M2) *hold, we conclude*

$$\forall\, k \in \mathbb{N} : \langle \mu_n, x^k \rangle \xrightarrow[n\to\infty]{} \langle \mu, x^k \rangle \ \text{in probability}$$

and thus $\mu_n \to \mu$ weakly in probability via Theorem 3.7.

iii) If (M1) *and* (M3) *hold, we conclude*

$$\forall\, k \in \mathbb{N} : \left[\langle \mu_n, x^k \rangle \xrightarrow[n\to\infty]{} \langle \mu, x^k \rangle \ \ \mathbb{P}\text{-a.s.} \right]$$

and thus $\mu_n \to \mu$ weakly almost surely via Theorem 3.7.

Proof. i) As we saw in the proof of Theorem 3.7, we find that for all $n \in \mathbb{N}$, the expected measure $\mathbb{E}\mu_n$ has moments of all orders and that for all $k \in \mathbb{N}$: $\langle \mathbb{E}\mu_n, x^k \rangle = \mathbb{E}\langle \mu_n, x^k \rangle$. Now given (M1), statement *i)* follows directly with Theorem 3.5.

ii)/iii) If (M1) holds, then (M2) (resp. (M3)) together with Lemma 3.10 shows that for all $k \in \mathbb{N}$ and $i \in [\ell_k]$, $D_n^{(k,i)}$ converges to a constant in probability (resp. almost surely) as $n \to \infty$, so that by (M1),

$$\langle \mu_n, x^k \rangle = D_n^{(k,1)} + \ldots + D_n^{(k,\ell_k)} \xrightarrow[n\to\infty]{} m_k$$

in probability (resp. almost surely). □

Lemma 3.10. *Let $z \in \mathbb{N}$ and $(Y_n)_n$ be random variables with $\mathbb{E}|Y_n|^z < \infty$ for all $n \in \mathbb{N}$. If $\mathbb{E}Y_n \to y$ and $\mathbb{E}|Y_n - \mathbb{E}Y_n|^z \to 0$, then $Y_n \to y$ in probability. If in addition, $\mathbb{E}|Y_n - \mathbb{E}Y_n|^z$ is summable, then $Y_n \to y$ almost surely.*

Proof. Using Markov's inequality, we calculate for $\varepsilon > 0$ arbitrary:

$$\begin{aligned}\mathbb{P}(|Y_n - y| > \varepsilon) &\leq \mathbb{P}\left(|Y_n - \mathbb{E}Y_n| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|\mathbb{E}Y_n - y| > \frac{\varepsilon}{2}\right) \\ &\leq \frac{2^z}{\varepsilon^z}\mathbb{E}|Y_n - \mathbb{E}Y_n|^z + \mathbb{P}\left(|\mathbb{E}Y_n - y| > \frac{\varepsilon}{2}\right).\end{aligned}$$

The statement follows (also using Borel-Cantelli), since the very last summand vanishes for all n large enough. □

3.4. The moments of the semicircle distribution

In random matrix theory, the probability measure that appears as the limit of the empirical spectral distribution is typically the semicircle distribution as defined in Definition 2.38. What we mean by *typically* is that it appears in Wigner's semicircle law, Theorem 2.42, which is the simplest non-trivial random matrix ensemble, for it has standardized entries which are independent up to the symmetry constraint. It is safe to say that the role of the semicircle distribution in random matrix theory resembles the role of the standard normal distribution in probability theory. To remind the reader, the semicircle distribution σ is the probability measure on $(\mathbb{R}, \mathcal{B})$ with Lebesgue-density f_σ where

$$\begin{aligned}f_\sigma : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto f_\sigma(x) \coloneqq \frac{1}{2\pi}\sqrt{4 - x^2}\mathbb{1}_{[-2,2]}(x).\end{aligned}$$

Since we would like to apply the method of moments to random matrix theory, we will proceed to derive the moments of the semicircle distribution. As it turns out, we will obtain that $\langle \sigma, x^0 \rangle = 1$, so that σ is identified as a probability measure, which we still owed to the reader.

Lemma 3.11. *The moments of the semicircle distribution σ are given by*

$$\text{For all } k \in \mathbb{N}_0 : \; m_{2k}^\sigma = \frac{(2k)!}{k!(k+1)!} \quad \text{and} \quad m_{2k+1}^\sigma = 0 \tag{3.2}$$

Proof. We follow the short proof in [5, p. 16]. To this end, note that the integrand is compactly supported and bounded. Further, for odd moments the integrand is odd, so the statement follows for odd moments. For even moments, we obtain the statement by the following calculation:

$$\begin{aligned}m_{2k}^\sigma &= \frac{1}{2\pi}\int_{-2}^{2} x^{2k}\sqrt{4 - x^2}\mathrm{d}x = \frac{1}{\pi}\int_0^2 x^{2k}\sqrt{4 - x^2}\mathrm{d}x \\ &= \frac{2^{2k+1}}{\pi}\int_0^1 y^{k-1/2}(1-y)^{1/2}\mathrm{d}y = \frac{2^{2k+1}}{\pi}B(k + 1/2,\, 3/2) \\ &= \frac{2^{2k+1}}{\pi}\frac{\Gamma(k+1/2)\Gamma(3/2)}{\Gamma(k+2)} = \frac{1}{k+1}\binom{2k}{k},\end{aligned}$$

where in the second step, we used that the integrand is even, in the third step we substituted x by $2\sqrt{y}$, in the fourth step we used the definition of the beta function B, in the fifth step, we used that for all $x, y > 0$: $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, where Γ is the gamma function, and in the last step we used that for all $n \in \mathbb{N}$: $\Gamma(n) = (n-1)!$, and for all $n \in \mathbb{N}_0$: $\Gamma(n+1/2) = (2n)!\sqrt{\pi}/(n!4^n)$. □

To use the method of moments to prove weak convergence to the semicircle distribution, we need the following corollary:

Corollary 3.12. *The semicircle distribution σ is uniquely determined by its moments.*

Proof. Since the support of σ is compact, the statement follows with Lemma 3.4. □

The values of the even moments of the semicircle distribution bear a special name:

Definition 3.13. *The* Catalan numbers *are elements of the sequence of natural numbers $(\mathcal{C}_k)_{k\in\mathbb{N}_0}$, where*

$$\forall\, k \in \mathbb{N}_0 :\ \mathcal{C}_k \coloneqq \frac{(2k)!}{k!(k+1)!}.$$

Combining the results of Lemma 3.11 with the definition of the Catalan numbers, we obtain for the sequence $(m_k^\sigma)_{k\in\mathbb{N}_0}$ of the moments of the semicircle distribution:

$$m_k^\sigma = \begin{cases} \mathcal{C}_{k/2} & \text{for } k \text{ even}, \\ 0 & \text{for } k \text{ odd}. \end{cases} \tag{3.3}$$

But the Catalan numbers are not only the (even) moments of the semicircle distribution. They also appear as the solution to various combinatorial problems, see [36] or [52], for example.

3.5. The moments of the Marchenko-Pastur distribution

For sample covariance matrices, the canonical limit is not Wigner's semicircle distribution, but the Marchenko-Pastur distribution μ^y with ratio index $y \in (0, \infty)$. As a reminder to the reader, μ^y is the sum of the point mass $(1-y^{-1})\mathbb{1}_{y>1}$ in zero and a Lebesgue-continuous part given by the density f_μ (where the parameter y is suppressed) as

$$\begin{aligned} f_\mu : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto f_\mu(x) \coloneqq \frac{1}{2\pi x y}\sqrt{(y_+ - x)(x - y_-)}\mathbb{1}_{(y_-, y_+)}(x), \end{aligned}$$

where $y_+ \coloneqq (1+\sqrt{y})^2$ and $y_- \coloneqq (1-\sqrt{y})^2$. In order to apply the method of moments to prove the Marchenko-Pastur law, we need to know the moments of μ^y, which is the content of the following lemma:

Lemma 3.14. *For all $y \in (0, \infty)$ and $k \in \mathbb{N}$, it holds*

$$\langle \mu^y, x^k \rangle = \sum_{r=0}^{k-1} \frac{y^r}{r+1} \binom{k}{r} \binom{k-1}{r}.$$

Proof. The proof is rather lengthy and can be found in [5, p. 40]. □

Corollary 3.15. *For every $y > 0$, the Marchenko-Pastur distribution μ^y is uniquely determined by its moments.*

Proof. Since the support of μ^y is compact, the statement follows with Lemma 3.4. □

3.6. Application of the method of moments to RMT

So far, we have pointed out what the method of moments is and how it works in deterministic and stochastic settings. Now we want to build the bridge to random matrix theory. To this end, we need the following observation, where as before, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$:

Lemma 3.16. *Let $n \in \mathbb{N}$ and $X \in SMat_n(\mathbb{K})$, then we obtain for all $k \in \mathbb{N}$:*

$$\sum_{i=1}^{n} (\lambda_i^X)^k = \operatorname{tr} X^k = \sum_{t_1,\dots,t_k=1}^{n} X(t_1,t_2)X(t_2,t_3)\cdots X(t_k,t_1).$$

Proof. The second equality is clear. For the first equality, note that since $X \in \mathrm{SMat}_n(\mathbb{K})$, by Lemma 2.32, there exists an invertible matrix $S \in \mathrm{Mat}_n(\mathbb{K})$ so that $X = S^{-1}DS$, where $D = \operatorname{diag}(\lambda_1^X, \dots, \lambda_n^X)$. Then

$$X^k = \underbrace{S^{-1}DS \cdot S^{-1}DS \cdot \ldots \cdot S^{-1}DS}_{k \text{ factors}} = S^{-1}D^kS = S^{-1} \operatorname{diag}\left((\lambda_1^X)^k, \dots, (\lambda_n^X)^k\right) S.$$

With Lemma 2.31, we obtain

$$\operatorname{tr}(X^k) = \operatorname{tr} \operatorname{diag}\left((\lambda_1^X)^k, \dots, (\lambda_n^X)^k\right) = \sum_{i=1}^{n} (\lambda_i^X)^k.$$ □

Corollary 3.17. *Let $(X_n)_n$ be a sequence of random matrices with corresponding ESDs $(\sigma_n)_n$. Then for all $k \in \mathbb{N}$ we find*

$$\langle \sigma_n, x^k \rangle = \frac{1}{n} \operatorname{tr} X_n^k = \frac{1}{n} \sum_{t_1,\dots,t_k=1}^{n} X_n(t_1,t_2)X_n(t_2,t_3)\cdots X_n(t_k,t_1). \tag{3.4}$$

Proof. Using Lemma 3.16, we calculate:

$$\langle \sigma_n, x^k \rangle = \frac{1}{n} \sum_{i=1}^{n} (\lambda_i^{X_n})^k = \frac{1}{n} \operatorname{tr} X_n^k = \frac{1}{n} \sum_{t_1,\dots,t_k=1}^{n} X_n(t_1,t_2)X_n(t_2,t_3)\cdots X_n(t_k,t_1).$$ □

The next theorem will be of use in explorative settings where the target distribution is not known or assumed yet. This is the very first step in showing that the ESDs of random matrices converge to a probability measure. To clarify terminology that we use, if Y is a $\mathbb{K}$-valued random variable, where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, and if $p \in \mathbb{N}_0$, then we call $\mathbb{E}|Y|^p$ the p-th absolute moment of Y. Further, we say that Y has absolute moments of all orders, if $\mathbb{E}|Y|^p < \infty$ for all $p \in \mathbb{N}_0$. Note that Y is integrable iff its first absolute moment exists.

Theorem 3.18. *Let $(\sigma_n)_n$ be the empirical spectral distributions of random matrices $(X_n)_n$, whose ($\mathbb{K}$-valued) entries have absolute moments of all orders. Then if*

$$\forall\, k \in \mathbb{N} : \mathbb{E}\left\langle \sigma_n, x^k \right\rangle \xrightarrow[n\to\infty]{} m_k,$$

where $(m_k)_k$ is a sequence of real numbers that satisfy the Carleman condition (cf. Theorem 3.3), then $(\sigma_n)_n$ converges weakly in expectation to a probability measure μ on $(\mathbb{R}, \mathcal{B})$ with moments $(m_k)_k$.

Proof. This follows with Theorem 3.9, since by Corollary 3.17, for each $k \in \mathbb{N}_0$, the k-th random moment is given by

$$\left\langle \sigma_n, x^k \right\rangle = \frac{1}{n} \sum_{t_1,\ldots,t_k=1}^{n} X_n(t_1,t_2)X_n(t_2,t_3)\cdots X_n(t_k,t_1),$$

which is a real-valued random variable whose expectation is finite, see the following Lemma 3.19. □

Lemma 3.19. *Let $Y_1, \ldots, Y_k$ be $\mathbb{K}$-valued random variables such that $\mathbb{E}|Y_i|^k < \infty$ for all $i \in \{1, \ldots, k\}$, then*

$$\mathbb{E}|Y_1 Y_2 \cdots Y_k| \;\leq\; \left(\mathbb{E}|Y_1|^k\right)^{\frac{1}{k}} \cdots \left(\mathbb{E}|Y_k|^k\right)^{\frac{1}{k}} \;\leq\; \max_{i=1,\ldots,k} \mathbb{E}|Y_i|^k$$

Proof. The second inequality is clear, so we only need to show the first one, which can be regarded as a generalization of the Cauchy-Schwarz inequality. We proceed by induction. The cases $k = 1$ and $k = 2$ are already known. By Hölder's inequality,

$$E|Y_1 \cdots Y_k| \leq \left(\mathbb{E}|Y_1 \cdots Y_{k-1}|^{\frac{k}{k-1}}\right)^{\frac{k-1}{k}} \left(\mathbb{E}|Y_k|^k\right)^{\frac{1}{k}}.$$

Using the induction hypothesis, we calculate

$$\mathbb{E}|Y_1|^{\frac{k}{k-1}} \cdots |Y_{k-1}|^{\frac{k}{k-1}} \leq \left(\mathbb{E}|Y_1|^k\right)^{\frac{1}{k-1}} \cdots \left(\mathbb{E}|Y_{k-1}|^k\right)^{\frac{1}{k-1}},$$

from which the statement follows. □

We remind the reader that convergence in expectation is a necessity for stronger convergence types, see Theorem 2.29. Therefore, Theorem 3.18 is really the basis for any explorative analysis. The next theorem will be of use either after Theorem 3.18 has been applied or if a priori, one has the target distribution of the ESDs in mind, for example if one wants to show a semicircle law.

Theorem 3.20. *Let $(\sigma_n)_n$ be the empirical spectral distributions of random matrices $(X_n)_n$, whose entries have absolute moments of all orders. Denote by μ a probability measure which is uniquely determined by its moments (cf. Theorem 3.3). Then*

i) σ_n converges to μ weakly in expectation, if for all $k \in \mathbb{N}$,

$$\mathbb{E}\left\langle \sigma_n, x^k \right\rangle \xrightarrow[n\to\infty]{} m_k.$$

We assume that for all $k \in \mathbb{N}$ we find a finite decomposition

$$\left\langle \sigma_n, x^k \right\rangle = D_n^{(k,1)} + \ldots + D_n^{(k,\ell_k)}$$

such that for all $k \in \mathbb{N}$ and all $i \in [\ell_k]$, $\mathbb{E}D_n^{(k,i)}$ converges to a constant as $n \to \infty$. (This decomposition will become clear from the analysis, for example when showing that i) holds.) Then

ii) σ_n converges to μ weakly in probability, if i) holds and for all $k \in \mathbb{N}$ and $i \in [\ell_k]$,

$$\exists\, z \in \mathbb{N} : \ \mathbb{E}\left| D_n^{(k,i)} - \mathbb{E}D_n^{(k,i)} \right|^z \xrightarrow[n\to\infty]{} 0,$$

iii) σ_n converges to μ weakly almost surely, if i) holds and for all $k \in \mathbb{N}$ and $i \in [\ell_k]$,

$$\exists\, z \in \mathbb{N} : \ \mathbb{E}\left| D_n^{(k,i)} - \mathbb{E}D_n^{(k,i)} \right|^z \xrightarrow[n\to\infty]{} 0 \quad \text{summably fast.}$$

Proof. This is a direct consequence of Theorem 3.9, considering that since matrix entries have moments of all orders, Corollary 3.17 and Lemma 3.19 imply that expected random moments and all other expectations are well-defined and finite. □

Next, as an application, let us discuss the proof strategy behind Wigner's semicircle law, Theorem 2.42, where we restrict our attention to convergence in probability:

Example 3.21. *Consider the setup of Theorem 2.42. Let $(m_k^\sigma)_{k\in\mathbb{N}}$ denote the moments of the semicircle distribution, then we can use Theorem 3.20 and show that*

1. For all $k \in \mathbb{N}$:

$$\mathbb{E}\left\langle \sigma_n, x^k \right\rangle = \frac{1}{n^{1+k/2}} \sum_{t_1,\ldots,t_k=1}^{n} \mathbb{E}a(t_1,t_2)a(t_2,t_3)\cdots a(t_k,t_1) \xrightarrow[n\to\infty]{} m_k^\sigma.$$

2. For all $k \in \mathbb{N}$:

$$\mathbb{E}\left(\left\langle \sigma_n, x^k \right\rangle^2 \right) \xrightarrow[n\to\infty]{} (m_k^\sigma)^2.$$

This will imply statements i) and ii) from the preceding theorem with $z = 2$, thus the semicircle law in probability.

4. The Semicircle and MP Laws by the Moment Method

4.1. General strategy and combinatorial structures

Assume that $(\sigma_n)_n$ is a sequence of ESDs of Wigner matrices W_n as in Theorem 2.42 and $(\mu_n)_n$ is a sequence of ESDs of MP matrices V_n as in Theorem 2.46. We would like to argue that $\sigma_n \to \sigma$ and $\mu_n \to \mu^y$ weakly for some $y > 0$, and in some stochastic sense, for example in probability or almost surely. Here, σ denotes the semicircle distribution and μ^y denotes the Marchenko-Pastur distribution on the real line. To show these convergence results, we carry out the following two steps, where notationally, either $\rho_n = \sigma_n$ and $\rho = \sigma$, or $\rho_n = \mu_n$ and $\rho = \mu^y$:

1. We show that for each fixed $k \in \mathbb{N}$, the expected moments $\mathbb{E}\langle \rho_n, x^k \rangle$ of the ESDs ρ_n converge to the deterministic moments $\langle \rho, x^k \rangle$ of the limit measure ρ, as $n \to \infty$. By Theorem 3.20, this will ensure that the limit law holds in expectation.
2. For each fixed $k \in \mathbb{N}$, we find a finite decomposition of the random moments, $\langle \rho_n, x^k \rangle = D_n^{(k,1)} + \ldots + D_n^{(k,\ell_k)}$, such that for each $k \in \mathbb{N}$ and each $i \in [\ell_k]$, $D_n^{(k,i)}$ converges in expectation to a constant as $n \to \infty$. This decomposition becomes clear from the analysis, for example from the first step, and. Then we show that for each $k \in \mathbb{N}$ and $i \in [\ell_k]$, there is a $z \in \mathbb{N}$ such that
$$\mathbb{E}\left| D_n^{(k,i)} - \mathbb{E}D_n^{(k,i)} \right|^z \xrightarrow[n\to\infty]{} 0. \tag{4.1}$$
Oftentimes, but not always, $z = 2$ or $z = 4$ will suffice. If (4.1) holds (resp. holds almost surely), then this will show that the $D_n^{(k,i)}$ converge in probability (resp. almost surely) to a constant so that with the first step, we obtain that for all $k \in \mathbb{N}$,
$$\langle \rho_n, x^k \rangle = \sum_{i=1}^{\ell_k} D_n^{(k,i)} \xrightarrow[n\to\infty]{} \langle \rho, x^k \rangle$$
in probability (resp./ almost surely).

For our analysis, we introduce some combinatorial concepts.

Definition 4.1. *Let $k \in \mathbb{N}$ be arbitrary, then*

i) A coloring *is a tuple $\underline{c} \in [k]^k$ with the property that $c_1 = 1$ and*
$$\forall\, a \in \{1, \ldots, k-1\} : \; c_{a+1} \leq 1 + \max_{\ell \in [a]} c_\ell.$$
Entries in a coloring will be called colors.

ii) If $\underline{t} \in [n]^k$ is a tuple and $\underline{c}$ is a coloring, then we say that $\underline{t}$ matches the coloring *$\underline{c}$ (and write $\underline{t} \sim \underline{c}$), if*
$$\forall\, i, j \in [k] : t_i = t_j \Leftrightarrow c_i = c_j.$$
In this case, we also call $\underline{c}$ the coloring of *$\underline{t}$ and write $\underline{c} = \underline{c}(\underline{t})$.*

A coloring is used to indicate at which places in a tuple there are equal or different entries. It is clear that each tuple $\underline{t} \in [n]^k$ matches exactly one (that is, *its*) coloring, which is constructed inductively as follows. Set $c_1 \coloneqq 1$, and for $\ell \in \{1, \dots, k-1\}$, if there is no $m \in [\ell]$ with $t_{\ell+1} = t_m$, set $c_{\ell+1} = \max\{c_1, \dots, c_\ell\} + 1$, whereas if $t_{\ell+1} = t_m$ for some $m \in [\ell]$, set $c_{\ell+1} \coloneqq c_m$. As an example, the coloring of the tuple $(5, 1, 4, 13, 4)$ is given by $(1, 2, 3, 4, 3)$.

Lemma 4.2. *Let $n, k \in \mathbb{N}$ with $n \geq k$.*

i) There are at most $k!$ colorings in $[k]^k$.
ii) Let $\underline{c} \in [k]^k$ be a coloring with ℓ colors, then

$$\#\{\underline{t} \in [n]^k : \underline{t} \sim \underline{c}\} = (n)_\ell \coloneqq n \cdot (n-1) \cdots (n-\ell+1) \tag{4.2}$$

In addition, it always holds that $\underline{c} \sim \underline{c}$.
iii) For a tuple $\underline{t} \in [n]^k$ denote by $V(\underline{t}) \coloneqq \{t_1, \dots, t_k\}$. Then $\underline{c}(\underline{t})$ has $\#V(\underline{t})$ colors, hence

$$\#\{\underline{t}' \in [n]^k : \underline{t}' \sim c(\underline{t})\} = (n)_{\#V(\underline{t})}. \tag{4.3}$$

Proof. To prove *i*), note that always $c_1 = 1$ and $c_{\ell+1} \in \{c_1, \dots, c_\ell, c_\ell + 1\}$. But $\#\{c_1, \dots, c_\ell, c_\ell + 1\} \leq \ell + 1$. For *ii*), in order to construct a tuple $\underline{t} \in [n]^k$ matching the coloring $\underline{c}$ we have n choices for t_1. Then if $c_2 = c_1$ this indicates that $t_2 \overset{!}{=} t_1$ so we are left with only one choice for t_2. If $c_2 \neq c_1$, however, we have $(n-1)$ choices for t_2. Proceeding this way, if $c_m = c_a$ for some $a < m$ then $t_m \overset{!}{=} t_a$ so we are left with only one choice for t_m. Otherwise, if c_m is *new* color, we have $n - \#\{c_1, \dots, c_{m-1}\}$ choices for t_m. Now since there exactly ℓ different colors in $\underline{c}$, we will encounter a new color exactly $\ell - 1$ times. Statement *iii*) follows directly from *ii*). □

4.2. The semicircle law

Let $W_n = n^{-1/2} X_n$ be a sequence of Wigner matrices with ESDs σ_n. In order to show $\sigma_n \to \sigma$ weakly almost surely, we follow the general strategy as outlined in Section 4.1. To utilize this method, we need the moments of σ_n and σ. By Lemma 3.11, the moments of σ are given by

$$\forall\, k \in \mathbb{N} : \langle \sigma, x^k \rangle = \begin{cases} \frac{1}{\frac{k}{2}+1} \binom{k}{\frac{k}{2}} & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd,} \end{cases} \tag{4.4}$$

whereas the moments of σ_n are given by (cf. Corollary 3.17)

$$\langle \sigma_n, x^k \rangle = \frac{1}{n} \operatorname{tr}\left[\left(\frac{1}{\sqrt{n}} X_n \right)^k \right] = \frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in [n]^k} X_n(\underline{t}), \tag{4.5}$$

where for all $\underline{t} \in [n]^k$ we define

$$X_n(\underline{t}) \coloneqq X_n(t_1, t_2) X_n(t_2, t_3) \cdots X_n(t_k, t_1). \tag{4.6}$$

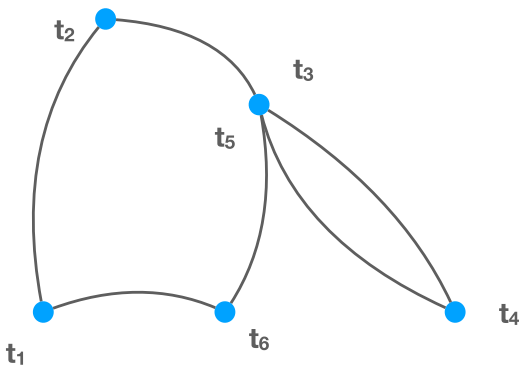


FIG 1. *Eurlerian graph* $\mathcal{G}(\underline{t})$.

Combinatorial preparations and graph theory

As we saw above in (4.5), the random moments $\langle \sigma_n, x^k \rangle$ expand into elaborate sums. In order to be able to analyze these sums, we sort them with the language of graph theory and then establish basic combinatorial facts.

Recall (4.6), then we adopt the view that each tuple $\underline{t} \in [n]^k$ spans a Eulerian graph as in Figure 1. To be precise, we obtain the (multi-)graph $\mathcal{G}(\underline{t}) = (V(\underline{t}), E(\underline{t}), \phi_{\underline{t}})$, with vertex set $V(\underline{t}) = \{t_1, \dots, t_k\}$, edge set $E(\underline{t}) = \{e_1, \dots, e_k\}$ and incidence function $\phi_{\underline{t}}(e_i) = \{t_i, t_{i+1}\}$, where $k+1 \equiv 1$. Each tuple $\underline{t}$ also denotes a Eulerian cycle of length k through its graph $\mathcal{G}(\underline{t})$ by

$$t_1, e_1, t_2, e_2, t_3, \dots, t_{k-1}, e_{k-1}, t_k, e_k, t_1. \tag{4.7}$$

Note that $\mathcal{G}(\underline{t})$ may contain loops and multi-edges. The language of graph theory allows us to express $\langle \sigma_n, x^k \rangle$ in a different way. For any tuple $\underline{t} \in [n]^k$, we define its profile

$$\rho(\underline{t}) = (\rho_1(\underline{t}), \dots, \rho_k(\underline{t})),$$

where for all $\ell \in [k]$:

$$\rho_\ell(\underline{t}) := \#\{\phi_{\underline{t}}(e) \mid e \in E(\underline{t}) \text{ is an } \ell\text{-fold edge}\}.$$

Here, an ℓ-fold edge in $E(\underline{t})$ is any element $e \in E(\underline{t})$ for which there are exactly $\ell - 1$ distinct other elements $e'_2, \dots, e'_\ell \in E(\underline{t})$ so that $\phi_{\underline{t}}(e) = \phi_{\underline{t}}(e'_j)$ for $j \in \{2, \dots, \ell\}$.

Then for all $\ell \in [k]$, the Eulerian cycle $\underline{t}$ traverses exactly $\phi_\ell(\underline{t})$ distinct ℓ-fold edges. As a result, the following trivial but useful equality holds:

$$k = \sum_{\ell=1}^{k} \ell \cdot \rho_\ell(\underline{t}). \tag{4.8}$$

Now for all $k \in \mathbb{N}$ we define the following set of profiles:

$$\Pi(k) = \left\{ \rho \in \{0, \dots, k\}^k \mid \rho \text{ profile of some } \underline{t} \in [n]^k \right\}.$$

Now we achieve a finite decomposition

$$\langle \sigma_n, x^k \rangle = \sum_{\rho \in \Pi(k)} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho)} X_n(\underline{t}), \tag{4.9}$$

where

$$\mathcal{T}^n(\rho) \coloneqq \left\{ \underline{t} \in [n]^k \mid \rho(\underline{t}) = \rho \right\}.$$

The transition from (4.5) to (4.9) allows us to identify exactly which components of the random moment contribute to the limit.

The next fundamental lemma will give an upper bound on the number of tuples $\underline{t}$ with at most $\ell \in [k]$ vertices. Notationally, we set $V(\underline{u}) \coloneqq \{u_1, \ldots, u_k\}$ for any $\underline{u} \in \mathbb{N}^k$, even if we do not interpret $\underline{u}$ as a graph. Further, if M is a set, $\#M \in \mathbb{N} \cup \{\infty\}$ denotes the number of elements in M

Lemma 4.3. *Let $n, k \in \mathbb{N}$ and $\ell \in \{1, 2, \ldots, k\}$ be arbitrary. Then*

$$\# \left\{ \underline{t} \in [n]^k \,|\, \#V(\underline{t}) \leq \ell \right\} \leq k^k \cdot n^\ell.$$

Proof. We first pick a coloring $\underline{c} \in [k]^k$ with at most ℓ colors for which we have at most k^k choices by Lemma 4.2 *i*). Since $\underline{c}$ has at most ℓ colors, the number of tuples $\underline{t}$ matching the coloring is bounded by $(n)_\ell$ by Lemma 4.2 *ii*). Therefore, we have at most $k^k (n)_\ell$ choices to pick an element from $\left\{ \underline{t} \in [n]^k \,|\, \#V(\underline{t}) \leq \ell \right\}$. □

Step 1: Convergence of expected moments

We proceed to analyze the expectation of

$$\langle \sigma_n, x^k \rangle = \sum_{\rho \in \Pi(k)} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho)} X_n(\underline{t}). \tag{4.10}$$

To this end, it suffices to analyze the expectation of each of the finitely many terms

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho)} X_n(\underline{t}) \tag{4.11}$$

for $\rho \in \Pi(k)$ separately. We make two trivial observations: If $\rho \in \Pi(k)$ with $\rho_1 > 0$, then for all $\underline{t} \in \mathcal{T}^n(\rho)$ it holds $\mathbb{E} X_n(\underline{t}) = 0$ due to independence and centeredness. Further, since $(X_n)_n$ is a Wigner scheme as in Definition 2.40, we can always apply the trivial bound

$$|\mathbb{E} X_n(\underline{t})| \leq L_k \tag{4.12}$$

for any $\underline{t} \in [n]^k$, where we also used Lemma 3.19.

For the bounds on $\#\mathcal{T}^n(\rho)$, we formulate the next lemma, which we take from [26].

Lemma 4.4. *Let $k \in \mathbb{N}$ be arbitrary. Then it holds:*

i) $\#\Pi(k) \leq 4^k$.

ii) Let $n \in \mathbb{N}$ *and* $\rho \in \Pi(k)$ *be arbitrary, then*

a) For any $\underline{t} \in \mathcal{T}^n(\rho)$ *it holds*

$$\#V(\underline{t}) \leq 1 + \rho_1 + \ldots + \rho_k - L(\underline{t}),$$

where $L(\underline{t})$ *denotes the number of loops in* $\underline{t}$*. In particular,*

$$\#\mathcal{T}^n(\rho) \leq k^k \cdot n^{1+\rho_1+\ldots+\rho_k}.$$

b) If ρ *contains an odd edge, then for any* $\underline{t} \in \mathcal{T}^n(\rho)$ *it holds*

$$\#V(\underline{t}) \leq \rho_1 + \ldots + \rho_k.$$

In particular,

$$\#\mathcal{T}^n(\rho) \leq k^k \cdot n^{\rho_1+\ldots+\rho_k}.$$

Proof. i) Each $\rho \in \Pi(k)$ is a k-tuple in which for all $\ell \in \{1, \ldots, k\}$ the entry ρ_ℓ lies in the set $\{0, 1, \ldots, \lfloor k/\ell \rfloor\}$, which follows directly from (4.8). Therefore,

$$\#\Pi(k) \leq \prod_{\ell=1}^{k} \left(\frac{k}{\ell} + 1 \right) = \frac{(2k)!}{k! \cdot k!} = \binom{2k}{k} \lesssim \frac{4^k}{\sqrt{2k\pi}} \leq 4^k,$$

where the fourth step is a well-known fact about the central binomial coefficient.
ii) It suffices to establish the upper bounds for $\#V(\underline{t})$, since the bounds on $\#\mathcal{T}^n(\rho)$ then follow directly with Lemma 4.3. Now to prove upper bounds for $\#V(\underline{t})$, the idea is to travel the Eulerian cycle generated by $\underline{t}$:

$$t_1, e_1, t_2, e_2, t_2, e_3, t_3, \ldots, t_k, e_k, t_1 \tag{4.13}$$

by picking an initial node t_i and then traversing the edges in increasing cyclic order until reaching the starting point again. On the way, we count the number of different vertices that were discovered. Whenever we pass an ℓ-fold edge, only the first instance of that edge may discover a new vertex, and only if the edge is not a loop.
a) We write $L(\underline{t}) \coloneqq L_1(\underline{t}) + \ldots + L_k(\underline{t})$ where $L_i(\underline{t})$ denotes the number of different i-fold loops in $\underline{t}$. We start our tour at t_1 and observe this very vertex. Then, as we travel along the cycle, for each $\ell \in \{1, \ldots, k\}$ we will pass $\ell \cdot (\rho_\ell - L_\ell(\underline{t}))$ *proper* ℓ-fold edges out of which only the first instance may discover a new node, and there are $\rho_\ell - L_\ell(\underline{t})$ of these first instances. Considering the initial node, we arrive at $\#V(\underline{t}) \leq 1 + \rho_1 - L_1(\underline{t}) + \ldots + \rho_k - L_k(\underline{t})$, which yields the desired inequality.
b) In presence of an odd edge, we can start the tour at a specific vertex such that the odd edge cannot contribute to the newly discovered vertices. To this end, fix an arbitrary ℓ-fold edge in $\underline{t}$ with ℓ odd. Let $e_{i_1}, \ldots, e_{i_\ell}$, $i_1 < \ldots < i_\ell$, be the instances of the ℓ-fold edges in question in the cycle (4.13). Since ℓ is odd, we must find a $k \in \{1, \ldots, \ell\}$ such that e_{i_k} and $e_{i_{k+1}}$ are traversed in the same

direction, since we are on a cycle. We then start our tour at t_{i_k} and observe this vertex. However, now none of the edges $e_{i_1}, \dots, e_{i_\ell}$ may discover a new vertex, since if our ℓ-fold edge is not a loop, the vertex $t_{i_{k+1}}$ must have been already discovered by some other edge. Therefore, the roundtrip leads to the discovery of at most $\rho_1 + \cdots + (\rho_\ell - 1) + \ldots + \rho_k$ new nodes in addition to the first node. □

We proceed to analyze (4.11) for all possible types of $\rho \in \Pi(k)$.
Case 1: $\rho_1 = 0$ and $\rho_\ell > 0$ for some $\ell \geq 3$.
Using Lemma 4.4 we obtain

$$\#\mathcal{T}^n(\rho) \leq \left\{ \begin{array}{c} k^k \cdot n^{\rho_1 + \ldots + \rho_k} \\ k^k \cdot n^{1+\rho_1+\ldots+\rho_k} \end{array} \right\} \leq k^k n^{\frac{k}{2}},$$

where the upper case is valid in presence of an odd edge (then $\rho_1 + \ldots + \rho_k \leq (k-3)/2 + 1$), and the lower case is valid if no odd edges are present (then $1 + \rho_1 + \ldots + \rho_k \leq 1 + (k-4)/2 + 1$). Therefore, by (4.12), (4.11) converges to zero in expectation.
Case 2: $\rho_1 > 0$.
Then by centeredness and independence, the expectation of the term in (4.11) is zero.
Case 3: $\rho_2 = k/2$.
Returning to the random moment in (4.10), we have seen in Cases 1 and 2 that for all $\rho \in \Pi(k)$ with $\rho_2 \neq k/2$,

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho)} X_n(\underline{t}) \xrightarrow[n\to\infty]{} 0 \qquad \text{in expectation.}$$

As a result, the only asymptotic contribution from the expectation in (4.10) may stem from cycles $\underline{t}$ containing only double edges. Their analysis is the content of this Case 3. Setting $\rho^{(k)}$ as the profile in $\Pi(k)$ with $\rho_2^{(k)} = k/2$ and $\rho_\ell^{(k)} = 0$ for all $\ell \neq 2$, then it is our goal to show

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho^{(k)})} X_n(\underline{t}) \xrightarrow[n\to\infty]{} \mathcal{C}_{\frac{k}{2}} \qquad \text{in expectation.} \tag{4.14}$$

To this end, we observe

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho^{(k)})} \mathbb{E} X_n(\underline{t}) = \frac{1}{n^{1+\frac{k}{2}}} \#\mathcal{T}^n(\rho^{(k)}). \tag{4.15}$$

Next, we note that any $\underline{t} \in \mathcal{T}^n(\rho^{(k)})$ has at most $k/2 + 1$ vertices, so we may subdivide this set further by defining

$$\mathcal{T}^n_{\leq k/2}(\rho^{(k)}) := \left\{ \underline{t} \in \mathcal{T}^n(\rho^{(k)}) : \ \#V(\underline{t}) \leq k/2 \right\},$$
$$\mathcal{T}^n_{k/2+1}(\rho^{(k)}) := \left\{ \underline{t} \in \mathcal{T}^n(\rho^{(k)}) : \ \#V(\underline{t}) = k/2 + 1 \right\}.$$

Note that by Lemma 4.3, $\#\mathcal{T}^n_{\leq k/2}(\rho^{(k)}) \leq k^k n^{k/2}$, so that (4.15) can be refined to

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t}\in\mathcal{T}^n(\rho^{(k)})} \mathbb{E}X_n(\underline{t}) = \frac{1}{n^{1+\frac{k}{2}}} \#\mathcal{T}^n_{k/2+1}(\rho^{(k)}) \ + \ o(1). \tag{4.16}$$

It is thus our task to show

$$\frac{1}{n^{1+\frac{k}{2}}} \#\mathcal{T}^n_{k/2+1} \xrightarrow[n\to\infty]{} \mathcal{C}_{\frac{k}{2}}. \tag{4.17}$$

The main tool is to count all possible colorings of tuples in $\mathcal{T}^n_{k/2+1}(\rho^{(k)})$, and then apply Lemma 4.2. It turns out that these colorings can be associated with a path difference sequence (pds) of the following form, where we may focus on even k, since otherwise, the set $\mathcal{T}^n_{k/2+1}(\rho^{(k)})$ is empty:

Definition 4.5. *A* Wigner path difference sequence *(Wigner-pds) of length* $2k$ *is a tuple* $(D_1, D_2, \ldots, D_{2k})$ *which satifies the following conditions:*

1) For all $i \in [2k]$*:* $D_i \in \{-1, +1\}$
2) $\sum_{i\in[2k]} D_i = 0$*,*
3) $\forall\, \ell \in [2k] : \sum_{i=1}^{\ell} D_i \geq 0$*.*

We denote by $\mathcal{W}(2k)$ *the set of all Wigner-pds of length* $2k$*.*

Lemma 4.6. *For all* $k \in \mathbb{N}$ *we find* $\#\mathcal{W}(2k) = \frac{1}{k+1}\binom{2k}{k} = \mathcal{C}_k$*.*

Proof. We prove the lemma with a reflection principle. To this end, due property 2), a Wigner-pds must contain as many "+1"-entries as "−1"-entries. To arrange k "+1"-entries and k "−1"-entries, we have

$$\binom{2k}{k}$$

choices. But since these choices do not in general respect condition 3) we have to subtract the number of tuples $(D_1, \ldots, D_{2k})$ that lead to a violation of 3). We show that these violating tuples are in bijective correspondence to all $(D'_1, \ldots, D'_{2k})$ with

1') $D'_i \in \{-1, +1\}$,
2') $\sum_{i\in[2k]} D'_i = -2$.

The number of these $(D'_1, \ldots, D'_k)$ is clearly given by

$$\binom{2k}{k+1}$$

so that the number of $(D_1, \ldots, D_{2k})$ that *do* satisfy 1), 2) *and* 3) is given by

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

For the bijection, let $(D_1, \dots, D_{2k})$ be arbitrary with k "+1"s and k "−1"s so that 3) is violated. Then there is an index t such that $\sum_{i=1}^{t} D_i = -1$ for the first time. Then $(D_{t+1}, \dots, D_{2k})$ is a vector which contains one more "+1" than "−1" entry. We define the vector $(D'_{t+1}, \dots, D'_{2k}) \coloneqq (-D_{t+1}, \dots, -D_{2k})$. Then $(D'_{t+1}, \dots, D'_{2k})$ contains one more "−1" than "+1". Defining $(D'_1, \dots, D'_t) \coloneqq (D_1, \dots, D_t)$ we thus have created a vector $(D'_1, \dots, D'_{2k})$ satisfying 1') and 2'). On the other hand, any vector $(D'_1, \dots, D'_{2k})$ satisfying 1) and 2) has a first hitting time t of -1. Applying exactly the same transformation as before, we will then obtain a vector $(D_1, \dots, D_{2k})$ satisfying 1) and 2), but violating 3). □

Now the clou is that all $D \in \mathcal{W}(2k)$ can be associated canonically with a specific Eulerian cycle $\underline{t}(D) \in \mathcal{T}^n_{k+1}(\rho^{(2k)})$. To see how this is done, let us first analyze simple properties a Eulerian cycle $\underline{t} \in \mathcal{T}^n_{k+1}(\rho^{(2k)})$. First, the graph $\mathcal{G}(\underline{t})$ is a *double edged tree*, that is, it consists of k distinct double edges and has $k+1$ vertices, therefore is a tree in the regular sense after eliminating one of each of the double edges (it also follows that all doubles edges are *proper*). Thus, the Eulerian cycle $\underline{t}$ crosses each edge twice, once in each direction, since a tree does not contain circles. We recall the representation of the cycle as in (4.7). Now given a $D \in \mathcal{W}(2k)$, we set $t_1 = 1$, and whenever $D_\ell = +1$, this means that a new vertex shall be discovered, so we set $t_{\ell+1} \coloneqq \max(t_1, \dots, t_\ell)+1$. On the other hand, if $D_\ell = -1$ then we shall backtrack, that is, $t_{\ell+1}$ shall be equal to one of the $t_1, \dots, t_\ell$, and so it must be equal to the t_i with $i \in \{1, \dots, \ell\}$ from which t_ℓ was visited, since otherwise, the cycle $\underline{t}$ would contain a circle. This completes the construction of $\underline{t}(D)$. It is clear by construction that $\underline{t}(D) \in \mathcal{T}^n_{k+1}(\rho^{(2k)})$. We observe that $\underline{c}(\underline{t}(D)) = \underline{t}(D)$, that is $\underline{t}(D)$ is its own coloring, since vertex numbers were always chosen as small as possible. Now if $\underline{t}' \sim \underline{c}(\underline{t}(D))$, we must have $\underline{t}' \in \mathcal{T}^n_{k+1}(\rho^{(2k)})$, since $\underline{t}'$ is then only an injective relabeling of vertices in $\underline{t}$.

We formulate the following Lemma from which (4.17) follows immediately.

Lemma 4.7. *The set $\mathcal{T}^n_{k+1}(\rho^{(2k)})$ has a decomposition as follows:*

$$\mathcal{T}^n_{k+1}(\rho^{(2k)}) = \dot{\bigcup_{D \in \mathcal{W}(2k)}} \left\{ \underline{t}' \in \mathcal{T}^n_{k+1}(\rho^{(2k)}) \mid \underline{t}' \sim \underline{c}(\underline{t}(D)) \right\} \tag{4.18}$$

In particular,

$$\#\mathcal{T}^n_{k+1}(\rho^{(2k)}) = \frac{1}{k+1} \binom{2k}{k} \cdot (n)_{k+1} \ . \tag{4.19}$$

Proof. Before the statement of Lemma 4.7, we have already argued "$\supseteq$" in (4.18). To show "$\subseteq$", let $\underline{t}' \in \mathcal{T}^n_{k+1}(\rho^{(2k)})$ be arbitrary and recall the representation of the cycle as in (4.7). We encode this cycle into a Wigner-pds $D(\underline{t}')$ and show that $\underline{t}' \sim \underline{c}(\underline{t}(D(\underline{t}'))$. To this end, start a tour at t'_1 and move along the cycle. For $\ell \in \{1, \dots, 2k\}$, if e_ℓ leads to a new vertex, we set $D_\ell = 1$ and if e_ℓ backtracks to an old vertex, we set $D_\ell = -1$. For example, we always have $D_1 = 1$, since each edge in $\underline{t}$ is proper, and $D_k = -1$, since this edge leads back to the – already seen – vertex t'_1. Let us argue that the tuple $D(\underline{t}') \coloneqq (D_1, D_2, \dots, D_{2k})$

we just constructed satisfies conditions 1), 2) and 3) as above. Condition 1) is clearly satisfied. For condition 2), note that $\underline{t}'$ has $k+1$ vertices, out of which k – all except the vertex t'_1 – were considered new while traversing $\underline{t}'$, so we must have k "+1"-entries and k "−1"-entries in $(D_1, D_2, \ldots, D_{2k})$. For condition 3) we realize that each vertex in $\underline{t}'$ is visited exactly twice by the cycle $\underline{t}'$, and that the first visit corresponds to a "+1"-entry while the second visit corresponds to a "−1"-entry in $(D_1, D_2, \ldots, D_{2k})$. Then 3) must be satisfied, since by nature of things, the "first" comes before the "second". The relation $\underline{t}' \sim \underline{c}(\underline{t}(D(\underline{t}'))$ follows with the construction of $\underline{t}(D(\underline{t}'))$ above the formulation of Lemma 4.7.

The equality (4.19) follows from Lemma 4.6, (4.18) and Lemma 4.2 *iii*), since for all $D \in \mathcal{W}(2k)$ we have $\#V(\underline{t}(D)) = k+1$ and all tuples matching the coloring $\underline{c}(\underline{t}(D))$ lie in $\mathcal{T}^n_{k+1}(\rho^{(2k)})$. □

Step 2: Decay of central moments

In Step 1, we have seen that for fixed $k \in \mathbb{N}$, the expectation of

$$\left\langle \sigma_n, x^k \right\rangle = \sum_{\rho \in \Pi(k)} \frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho)} X_n(\underline{t}) \tag{4.20}$$

converges to the k-th moment of the semicircle distribution. In particular, we have seen that each of the finitely many summands

$$\frac{1}{n^{1+\frac{k}{2}}} \sum_{\underline{t} \in \mathcal{T}^n(\rho)} X_n(\underline{t}) \tag{4.21}$$

converges to a constant in expectation. To show that the random moments in (4.20) converge almost surely to the moments of the semicircle distribution, it thus suffices – by Lemma 3.10 – to show that for all $\rho \in \Pi(k)$, the variance of each term in (4.21) decays summably fast. The variance of (4.21) is given by

$$\frac{1}{n^{k+2}} \sum_{\underline{t}, \underline{t}' \in \mathcal{T}^n(\rho^{(k)})} \left[\mathbb{E} X_n(\underline{t}) X_n(\underline{t}') - \mathbb{E} X_n(\underline{t}) \mathbb{E} X_n(\underline{t}') \right]. \tag{4.22}$$

We observe that for all $\underline{t}, \underline{t}' \in \mathcal{T}^n(\rho^{(k)})$ which are edge-disjoint, the corresponding summand in (4.22) vanishes. Thus it suffices to consider those $\underline{t}, \underline{t}' \in \mathcal{T}^n(\rho)$ which have at least one edge in common. To this end, denote for all $\ell \in [k]$:

$$\mathcal{T}^n_{c(\ell)}(\rho) \coloneqq \left\{ (\underline{t}, \underline{t}') \in (\mathcal{T}^n(\rho))^2 \,|\, \underline{t} \text{ and } \underline{t}' \text{ have exactly } \ell \text{ edges in common} \right\}.$$

Our goal now is to evaluate for each $\ell \in [k]$ the term

$$\frac{1}{n^{k+2}} \sum_{(\underline{t}, \underline{t}') \in \mathcal{T}^n_{c(\ell)}(\rho)} \left[\mathbb{E} X_n(\underline{t}) X_n(\underline{t}') - \mathbb{E} X_n(\underline{t}) \mathbb{E} X_n(\underline{t}') \right]. \tag{4.23}$$

To this end, we need to establish bounds on $\#\mathcal{T}^n_{c(\ell)}(\rho)$.

Lemma 4.8. *Let $\rho \in \Pi(k)$ and $\ell \in [k]$, then the following statements hold:*

i) For all $\underline{t}, \underline{t}' \in \mathcal{T}^n(\rho)$ with at least ℓ common edges, it holds

$$\#(V(\underline{t}) \cup V(\underline{t}')) \leq 1 + 2\sum_{i=1}^{k} \rho_i - \ell$$

In particular,

$$\#\mathcal{T}^n_{c(\ell)}(\rho) \leq (2k)^{2k} n^{1+2\sum_{i=1}^{k}\rho_i - \ell}$$

ii) If there is an $m \in [k]$ odd with $\rho_m \geq 1$, then for all $\underline{t}, \underline{t}' \in \mathcal{T}^n(\rho)$ with at least ℓ common edges, it holds

$$\#(V(\underline{t}) \cup V(\underline{t}')) \leq 2\sum_{i=1}^{k} \rho_i - \ell.$$

In particular,

$$\#\mathcal{T}^n_{c(\ell)}(\rho) \leq (2k)^{2k} n^{2\sum_{i=1}^{k}\rho_i - \ell}.$$

Proof. For statement *ii*) we assume w.l.o.g. that $\underline{t}$ has an odd edge. Since the graphs spanned by $\underline{t}$ and $\underline{t}'$ share $\ell \geq 1$ common edges, we may take a tour around the joint Eulerian cycle, starting before a common edge, traveling first all edges of $\underline{t}$ and then all edges of $\underline{t}'$. While walking the edges of $\underline{t}$, we can see at most $\rho_1 + \ldots + \rho_k$ different nodes by Lemma 4.4. Next, traveling all edges of $\underline{t}'$, at most all the single edges and first instances of m-fold edges with $m \in \{2, \ldots, k\}$ of $\underline{t}'$ may discover a new node, but only if they have not been traversed before during the walk along $\underline{t}$. Since we have ℓ common edges, we can see at most $\rho'_1 + \ldots + \rho'_k - \ell$ new nodes. We established the bounds on the number of vertices in *ii*). The second statement in *ii*) follows immediately with Lemma 4.3 *i*) by concatenating $(\underline{t}, \underline{t}') \in [n]^{2k}$. For statement *i*) we proceed exactly in the same manner: Traveling $\underline{t}$ we can see at most $1 + \rho_1 + \rho_2 + \ldots + \rho_k$ nodes by Lemma 4.4, then traveling $\underline{t}'$ we can see at most $\rho'_1 + \ldots + \rho'_k - \ell$ new nodes. Now apply Lemma 4.3 again. □

Case 1: $\rho_1 \geq 1$

In this case, the term in (4.23) simplifies and we must argue that for each $\ell \in [k]$,

$$\frac{1}{n^{k+2}} \sum_{(\underline{t},\underline{t}') \in \mathcal{T}^n_{c(\ell)}(\rho)} \mathbb{E} X_n(\underline{t}) X_n(\underline{t}') \tag{4.24}$$

decays summably fast to zero. But we note that if $\underline{t}$ and $\underline{t}'$ have $1 \leq \ell < \rho_1$ common edges, $\mathbb{E}X_n(\underline{t})X_n(\underline{t}')$ vanishes, since not all single edges can be eliminated due to overlapping. Thus, it suffices to consider those $\underline{t}, \underline{t}' \in \mathcal{T}^n(\rho)$ which have $\ell \geq \rho_1$ edges in common. Now if $\rho \in \Pi(k)$ with $\ell \geq \rho_1 \geq 1$, then

$$2\sum_{i=1}^{k} \rho_i - \ell \leq 2\left(\rho_1 + \frac{k - \rho_1}{2}\right) - \rho_1 \leq k,$$

so Lemma 4.8 *ii)* yields

$$\#\mathcal{T}^n_{c(\ell)}(\rho) \leq (2k)^{2k} n^{2\sum_{i=1}^k \rho_i - \ell} \leq (2k)^{2k} n^k.$$

Since every summand in (4.24) is bounded by L_{2k}, it follows that (4.24) is $O(n^{-2})$, thus converges to zero summably fast.

Case 2: $\rho_1 = 0$

In this case, each summand in (4.23) is bounded by $L_{2k} + L_k^2$. Further, we obtain for all $\rho \in \Pi(k)$ with $\rho_1 = 0$ and $\ell \geq 1$ that

$$1 + 2\sum_{i=1}^k \rho_i - \ell \leq 1 + 2 \cdot \frac{k}{2} - 1 = k,$$

so that by Lemma 4.8 *i)* we find

$$\#\mathcal{T}^n_{c(\ell)}(\rho) \leq (2k)^{2k} n^{1+2\sum_{i=1}^k \rho_i - \ell} \leq (2k)^{2k} n^k,$$

so that the sum in (4.23) is $O(n^{-2})$, hence converges to zero summably fast.

4.3. The Marchenko-Pastur law

Let $(X_n)_n$ be an MP scheme as in Definition 2.43, $V_n \coloneqq n^{-1} X_n X_n^T$, and μ_n be the ESDs of V_n. Denote $y \coloneqq \lim_n p/n$. In order to show $\mu_n \to \mu^y$ weakly almost surely, we follow the general strategy as outlined in Section 4.1. To utilize this method, we need the moments of μ_n and μ^y. By Lemma 3.14, the moments of μ^y are given by

$$\forall\, k \in \mathbb{N} : \langle \mu^y, x^k \rangle = \sum_{r=0}^{k-1} \frac{y^r}{r+1} \binom{k}{r} \binom{k-1}{r}, \tag{4.25}$$

whereas the moments of μ_n are given by (cf. Corollary 3.17) and (4.5))

$$\begin{aligned}
\langle \mu_n, x^k \rangle \;=\; & \frac{1}{p} \operatorname{tr}\left[\left(\frac{1}{n} X_n X_n^T \right)^k \right] \;=\; \frac{1}{pn^k} \sum_{s \in [p]} (X_n X_n^T)^k(s,s) \\
= \; & \frac{1}{pn^k} \sum_{s_1,\ldots,s_k \in [p]} (X_n X_n^T)(s_1,s_2)(X_n X_n^T)(s_2,s_3) \ldots (X_n X_n^T)(s_k,s_1) \\
= \; & \frac{1}{pn^k} \sum_{s_1,\ldots,s_k \in [p]} \sum_{t_1,\ldots,t_k \in [n]} X_n(s_1,t_1) X_n(s_2,t_1) X_n(s_2,t_2) X_n(s_3,t_2) \ldots \\
& \times X_n(s_k,t_k) X_n(s_1,t_k) \\
= \; & \frac{1}{pn^k} \sum_{\underline{s} \in [p]^k} \sum_{\underline{t} \in [n]^k} X_n(\underline{s},\underline{t}),
\end{aligned} \tag{4.26}$$

where for all $\underline{s} \in [p]^k$ and $\underline{t} \in [n]^k$ we define

$$X_n(\underline{s},\underline{t}) \coloneqq X_n(s_1,t_1) X_n(s_2,t_1) X_n(s_2,t_2) X_n(s_3,t_2) \ldots X_n(s_k,t_k) X_n(s_1,t_k). \tag{4.27}$$

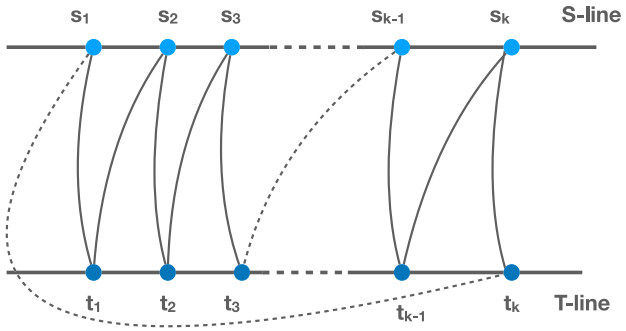


FIG 2. *Eurlerian bipartite graph* $\mathcal{G}(\underline{s},\underline{t})$.

Combinatorial preparations and graph theory

As we saw above in (4.26), the random moments $\langle \mu_n, x^k \rangle$ expand into elaborate sums. In order to be able to analyze these sums, we sort them with the language of graph theory and then establish basic combinatorial facts.

Recall (4.27), then we adopt the view that each pair $(\underline{s},\underline{t}) \in [p]^k \times [n]^k$ spans a Eulerian bipartite graph as follows:

Here, elements in the set $\{s_1,\dots,s_k\}$ resp. $\{t_1,\dots,t_k\}$ are called S-nodes resp. T-nodes. S and T-nodes are considered different even if their value is the same and are thus placed on separate lines – called S-line and T-line – which are drawn horizontally beneath each other. Then we draw an undirected edge $\{s_i,t_j\}$ between s_i and t_j, $i \in [p]$, $j \in [n]$, whenever (s_i,t_j) or (t_j,s_i) appears in (4.27), where we allow for multi-edges. This yields the (multi-)graph $\mathcal{G}(\underline{s},\underline{t}) = (V(\underline{s},\underline{t}), E(\underline{s},\underline{t}), \phi_{\underline{s},\underline{t}})$, where

$$\begin{aligned} V(\underline{s},\underline{t}) &= \{s_1,\dots,s_k\} \dot\cup \{t_1,\dots,t_k\} \qquad \text{(disjoint union)} \\ E(\underline{s},\underline{t}) &= \{d_1,\dots,d_k\} \dot\cup \{u_1,\dots,u_k\} \qquad \text{(down edges, up edges)} \\ &= \{e_1,e_2,\dots,e_{2k}\} \qquad (e_{2l-1} = d_l,\ e_{2l} = u_l,\ l = 1,\dots,k) \\ \phi_{\underline{s},\underline{t}}(d_i) &= \{s_i,t_i\} \\ \phi_{\underline{s},\underline{t}}(u_i) &= \{s_{i+1},t_i\} \end{aligned}$$

Each $(\underline{s},\underline{t})$ also denotes a Eulerian cycle of length $2k$ through its graph $\mathcal{G}(\underline{s},\underline{t})$ by

$$s_1,d_1,t_1,u_1,s_2,d_2,t_2,\dots,u_{k-1},s_k,d_k,t_k,u_k,s_1 \tag{4.28}$$

Figure 2 contains a visualisation of the graph $\mathcal{G}(\underline{s},\underline{t})$. Note that by construction, $\mathcal{G}(\underline{s},\underline{t})$ contains no loops, but may contain multi-edges. The language of graph theory allows us to express $\langle \mu_n, x^k \rangle$ in a different fashion. For any pair of tuples $(\underline{s},\underline{t}) \in [p]^k \times [n]^k$, we define its profile

$$\rho(\underline{s},\underline{t}) = (\rho_1(\underline{s},\underline{t}),\dots,\rho_{2k}(\underline{s},\underline{t})),$$

where for all $\ell \in [2k]$:

$$\rho_\ell(\underline{s},\underline{t}) = \#\{\phi_{\underline{s},\underline{t}}(e) \mid e \in E(\underline{s},\underline{t}) \text{ is an } \ell\text{-fold edge}\}.$$

Here, an ℓ-fold edge in $E(\underline{s},\underline{t})$ is any element $e \in E(\underline{s},\underline{t})$ for which there are exactly $\ell-1$ distinct other elements $e'_2,\dots,e'_\ell \in E(\underline{s},\underline{t})$ so that $\phi_{\underline{s},\underline{t}}(e) = \phi_{\underline{s},\underline{t}}(e'_j)$ for $j \in \{2,\dots,\ell\}$.

Then for all $\ell \in [2k]$, the Eulerian circuit $(\underline{s},\underline{t})$ traverses exactly $\phi_\ell(\underline{s},\underline{t})$ distinct ℓ-fold edges. As a result, the following trivial but useful equality holds:

$$2k = \sum_{\ell=1}^{2k} \ell \cdot \rho_\ell(\underline{s},\underline{t}). \tag{4.29}$$

Now for all $k \in \{1,\dots,2k\}$ we define the following set of profiles:

$$\Pi(2k) = \left\{\rho \in \{0,\dots,2k\}^{2k} \mid \rho \text{ profile of some } (\underline{s},\underline{t}) \in [p]^k \times [n]^k\right\}.$$

Now we construct the finite decomposition

$$\left\langle \mu_n, x^k \right\rangle = \sum_{\rho\in\Pi(2k)} \frac{1}{pn^k} \sum_{(\underline{s},\underline{t})\in\mathcal{T}^{p,n}(\rho)} X_n(\underline{s},\underline{t}), \tag{4.30}$$

where

$$\mathcal{T}^{p,n}(\rho) \coloneqq \left\{(\underline{s},\underline{t}) \in [p]^k \times [n]^k \mid \rho(\underline{s},\underline{t}) = \rho\right\}.$$

The transition from (4.26) to (4.30) allows us to analyze the contribution of paths that match certain profiles, and to identify the profiles the paths of which contribute to the limit.

The next fundamental lemma will give an upper bound on the number of tuple pairs $(\underline{s},\underline{t})$ with at most $\ell \in [2k]$ vertices. Note that there are always at least two vertices present, since S-nodes and T-nodes are disjoint. Notationally, we set $V(\underline{u}) \coloneqq \{u_1,\dots,u_k\}$ for any $\underline{u} \in \mathbb{N}^k$ and $V(\underline{u},\underline{v}) \coloneqq \{u_1,\dots,u_k\} \dot\cup \{v_1,\dots,v_k\}$ for any $\underline{u}$, $\underline{v} \in \mathbb{N}^k$, even if we do not view $(\underline{u},\underline{v})$ as a graph.

Lemma 4.9. *Let $p,n,k \in \mathbb{N}$, $a,b \in \{1,\dots,k\}$ and $\ell \in \{2,3,\dots,2k\}$ be arbitrary. Then*

$$\begin{aligned} &i) \quad \#\left\{(\underline{s},\underline{t}) \in [p]^k \times [n]^k \,|\, \#V(\underline{s}) = a, \#V(\underline{t}) = b\right\} \le k^{2k} \cdot p^a n^b \\ &ii) \quad \#\left\{(\underline{s},\underline{t}) \in [p]^k \times [n]^k \,|\, \#V(\underline{s},\underline{t}) \le \ell\right\} \le k^{2k+2} \cdot (p \vee n)^\ell. \end{aligned}$$

Proof. For i) we first fix the colorings for $\underline{s}$ with a colors and $\underline{t}$ with b colors, for which be have at most k^{2k} choices (Lemma 4.2). After fixing the colorings, we are left with at most p^a choices for the tuple $\underline{s} \in [p]^k$ and at most n^b choices for the tuple $\underline{t} \in [n]^k$, which yields the desired inequality. For *ii)* we first decide on the number $a \le k$ of different vertices in $\underline{s}$ and the number $b \le k$ of different vertices in $\underline{t}$ such that $a+b \le \ell$. This choice of (a,b) admits at most k^2 choices. Then with *i)*, the statement follows. □

Step 1: Convergence of expected moments

We proceed to analyze the expectation of

$$\left\langle \mu_n, x^k \right\rangle = \sum_{\rho \in \Pi(2k)} \frac{1}{pn^k} \sum_{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)} X_n(\underline{s},\underline{t}). \tag{4.31}$$

To this end, it suffices to analyze the expectation of each of the finitely many terms

$$\frac{1}{pn^k} \sum_{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)} X_n(\underline{s},\underline{t}) \tag{4.32}$$

for $\rho \in \Pi(2k)$ separately. As a first observation, note that if $\rho_1 \geq 1$, we have $\mathbb{E}X_n(\underline{s},\underline{t}) = 0$ for all $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)$ due to independence and centeredness. Further, since $(X_n)_n$ is an MP-scheme as in Definition 2.43, we can always apply the trivial bound

$$|\mathbb{E}X_n(\underline{s},\underline{t})| \leq L_{2k}, \tag{4.33}$$

where we also used Lemma 3.19.

For the bounds on $\#\mathcal{T}^{p,n}(\rho)$ we formulate the next lemma. It is a modification of similar lemmas obtained in [26].

Lemma 4.10. *Let $k \in \mathbb{N}$ be arbitrary. Then it holds:*

i) $\#\Pi(2k) \leq 16^k$.

ii) Let $p, n \in \mathbb{N}$ and $\rho \in \Pi(2k)$ be arbitrary, then

a) For any $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)$ we obtain

$$\#V(\underline{s},\underline{t}) \leq 1 + \rho_1 + \ldots + \rho_{2k}.$$

In particular,

$$\#\mathcal{T}^{p,n}(\rho) \leq k^{2k+2} \cdot (p \vee n)^{1+\rho_1+\ldots+\rho_{2k}}.$$

b) If ρ contains an odd edge, then for any $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)$ we obtain

$$\#V(\underline{s},\underline{t}) \leq \rho_1 + \ldots + \rho_{2k}.$$

In particular,

$$\#\mathcal{T}^{p,n}(\rho) \leq k^{2k+2} \cdot (p \vee n)^{\rho_1+\ldots+\rho_{2k}}.$$

Proof. i) Each $\rho \in \Pi(2k)$ is a $2k$-tuple in which for all $\ell \in \{1, \ldots, 2k\}$ the entry ρ_ℓ lies in the set $\{0, 1, \ldots, \lfloor 2k/\ell \rfloor\}$, which follows directly from (4.29). Therefore,

$$\#\Pi(2k) \leq \prod_{\ell=1}^{2k} \left(\frac{2k}{\ell} + 1 \right) = \frac{(4k)!}{(2k)! \cdot (2k)!} = \binom{2(2k)}{2k} \lesssim \frac{4^{2k}}{\sqrt{2k\pi}} \leq 16^k,$$

where the fourth step is a well-known fact about the central binomial coefficient. ii) It suffices to establish the upper bounds for $\#V(\underline{s},\underline{t})$, since the bounds on

$\#\mathcal{T}^{p,n}(\rho)$ then follow directly with Lemma 4.9 ii). Now to prove upper bounds for $\#V(\underline{s},\underline{t})$, the idea is to travel the Eulerian cycle generated by $(\underline{s},\underline{t})$:

$$s_1, e_1, t_1, e_2, s_2, e_3, t_2, \ldots, t_k, e_{2k}, s_1 \tag{4.34}$$

by picking an initial node s_i or t_i and then traversing the edges in increasing cyclic order until reaching the starting point again. On the way, we count the number of different nodes that were discovered. Whenever we pass an ℓ-fold edge, only the first instance of that edge may discover a new vertex.
a) We start our tour at s_1 and observe this very vertex. Then, as we travel along the cycle, for each $\ell \in \{1,\ldots,2k\}$ we will pass $\ell \cdot \rho_\ell$ ℓ-fold edges out of which only the first instance may discover a new node, and there are ρ_ℓ of these first instances. Considering the initial node, we arrive at $\#V(\underline{s},\underline{t}) \leq 1+\rho_1+\ldots+\rho_{2k}$, which yields the desired inequality.
b) In presence of an odd edge, we can start the tour at a specific vertex such that the odd edge cannot contribute to the newly discovered vertices. To this end, fix an ℓ-fold edge in $(\underline{s},\underline{t})$ with ℓ odd. Let $e_{i_1},\ldots,e_{i_\ell}$, $i_1 < \ldots < i_\ell$, be the instances of the ℓ-fold edge in question in the cycle (4.34). Since ℓ is odd, we must find a $k \in \{1,\ldots,\ell\}$ such that e_{i_k} and $e_{i_{k+1}}$ are both up edges or both down edges (where $\ell+1 \equiv 1$), since we are on a cycle. W.l.o.g. e_{i_k} is a down edge, thus leading to t_{i_k}. We start our tour at t_{i_k} and observe this vertex. However, now none of the edges $e_{i_1},\ldots,e_{i_\ell}$ may discover a new vertex, since the vertex $s_{i_{k+1}}$ must be discovered by some other edge. Therefore, the roundtrip leads to the discovery of at most $\rho_1+\cdots+(\rho_\ell-1)+\ldots+\rho_{2k}$ new nodes in addition to the first node. □

We proceed to analyze (4.31) for all possible types of $\rho \in \Pi(2k)$:
Case 1: $\rho_1 = 0$ and $\rho_\ell > 0$ for some $\ell \geq 3$.
Using Lemma 4.10 we obtain

$$\#\mathcal{T}^{p,n}(\rho) \leq k^{2k+2} \cdot (p \vee n)^{1+\rho_1+\ldots+\rho_{2k}} \leq k^{2k+2}(p \vee n)^k,$$

since with $\rho_\ell > 0$ for some $\ell \geq 3$ it follows

$$1+\rho_1+\ldots+\rho_{2k} \leq \left\{ \begin{array}{c} 1+\frac{2k-6}{2}+2 \\ 1+\frac{2k-4}{2}+1 \end{array} \right\} = k,$$

where the upper case is valid in presence of an odd edge (so we find at least a second odd edge), and the lower case is valid if no odd edges are present. Therefore, by (4.33), (4.32) converges to zero in expectation.
Case 2: $\rho_1 > 0$.
Then by centeredness and independence, the expectation in (4.32) is zero.
Case 3: $\rho_2 = k/2$.
Returning to the random moment in (4.31), we have seen in Cases 1 and 2 that for all $\rho \in \Pi(k)$ with $\rho_2 \neq k$,

$$\frac{1}{pn^k} \sum_{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)} X_n(\underline{s},\underline{t}) \xrightarrow[n\to\infty]{} 0 \qquad \text{in expectation.}$$

As a result, the only asymptotic contribution in (4.10) may stem from cycles $(\underline{s},\underline{t})$ containing only double edges. Their analysis is the content of this Case 3. Setting $\rho^{(k)}$ as the profile in $\Pi(2k)$ with $\rho_2^{(k)} = k$ and $\rho_\ell^{(k)} = 0$ for all $\ell \neq 2$, then it is our goal to show (cf. (4.25))

$$\frac{1}{pn^k} \sum_{(\underline{s},\underline{t})\in\mathcal{T}^{p,n}(\rho^{(k)})} X_n(\underline{s},\underline{t}) \xrightarrow[n\to\infty]{} \sum_{r=0}^{k-1} \frac{y^r}{r+1}\binom{k}{r}\binom{k-1}{r} \qquad \text{in expectation.} \tag{4.35}$$

To this end, we observe

$$\frac{1}{pn^k} \sum_{(\underline{s},\underline{t})\in\mathcal{T}^{p,n}(\rho^{(k)})} \mathbb{E}X_n(\underline{s},\underline{t}) = \frac{1}{pn^k} \#\mathcal{T}^{p,n}(\rho^{(k)}). \tag{4.36}$$

We note that any $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho^{(k)})$ has at most $k+1$ vertices, so we may subdivide this set further: We define

$$\mathcal{T}^{p,n}_{\leq k}(\rho^{(k)}) \coloneqq \left\{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho^{(k)}) : \ \#V(\underline{s},\underline{t}) \leq k\right\},$$
$$\mathcal{T}^{p,n}_{k+1}(\rho^{(k)}) \coloneqq \left\{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho^{(k)}) : \ \#V(\underline{s},\underline{t}) = k+1\right\},$$

and note that by Lemma 4.9, $\#\mathcal{T}^{p,n}_{\leq k}(\rho^{(k)}) \leq k^{2k+2}(p\vee n)^k$, so that (4.36) can be refined to

$$\frac{1}{pn^k} \sum_{(\underline{s},\underline{t})\in\mathcal{T}^{p,n}(\rho^{(k)})} \mathbb{E}X_n(\underline{s},\underline{t}) = \frac{1}{pn^k} \#\mathcal{T}^{p,n}_{k+1}(\rho^{(k)}) \ + \ o(1). \tag{4.37}$$

It is thus our task to show

$$\frac{1}{pn^k} \#\mathcal{T}^{p,n}_{k+1}(\rho^{(k)}) \xrightarrow[n\to\infty]{} \sum_{r=0}^{k-1} \frac{y^r}{r+1}\binom{k}{r}\binom{k-1}{r}. \tag{4.38}$$

To this end, for all $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}_{k+1}(\rho^{(k)})$ we track the number of vertices in $\underline{s}$ and the number vertices in $\underline{t}$ that the cycle visits. Thus, for all $a,b \in \mathbb{N}$ with $a+b = k+1$ we define

$$\mathcal{T}^{p,n}_{a,b}(\rho^{(k)}) \coloneqq \left\{(\underline{s},\underline{t}) \in [p]^k \times [n]^k \mid \rho(\underline{s},\underline{t}) = \rho^{(k)}, \#V(\underline{s}) = a, \#V(\underline{t}) = b\right\}.$$

Then we obtain a *partition*

$$\mathcal{T}^{p,n}_{k+1}(\rho^{(k)}) = \bigcup_{r=0}^{k-1} \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)}).$$

As a result, to show (4.38) it suffices to show that for all $r \in \{0,\dots,k-1\}$,

$$\frac{1}{pn^k} \#\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)}) \xrightarrow[n\to\infty]{} \frac{y^r}{r+1}\binom{k}{r}\binom{k-1}{r}. \tag{4.39}$$

It remains to evaluate $\#\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ for all $r \in \{0,\dots,k-1\}$. This is done by identifying the number of different color structures that an $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ may assume and then by multiplying this number with the number of possible colorings, which is a trivial task. The main tool to count all possible color structures is to associate with each $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ a path difference sequence (pds) of the following form:

Definition 4.11. *A Marcenko-Pastur path difference sequence (MP-pds) of length $2k$ and weight $r \in \{0,\dots,k-1\}$ is a tuple $(D_1,U_1,D_2,U_2,\dots,D_k,U_k) = (M_1,\dots,M_{2k})$ which satifies the following conditions:*

1) $D_i \in \{-1,0\}$ and $U_i \in \{0,1\}$.
2) $\sum_{i\in[k]} U_i = r$ and $\sum_{i\in[k]} D_i = -r$.
3) $\forall\, \ell \in \{1,\dots,2k\} : \sum_{i=1}^{\ell} M_i \geq 0$.

We denote by $\mathcal{M}(k,r)$ the set of all MP-pds of length $2k$ and weight r.

Lemma 4.12. *For all $k \in \mathbb{N}$ and $r \in \{0,\dots,k-1\}$ we find $\#\mathcal{M}(k,r) = \frac{1}{r+1}\binom{k-1}{r}\binom{k}{r}$.*

Proof. We assume $r \geq 1$ since for $r = 0$ the statement is clear. We prove the lemma with a reflection principle. First note that $M_1 = D_1 = 0$ and $M_{2k} = U_k = 0$ so that we are interested in all sequences $(M_2,\dots,M_{2k-1})$ where

1) $M_i \in \{-1,0\}$ for i odd and $M_i \in \{0,1\}$ for i even.
2) $\sum_{i \text{ odd}} M_i = -r$ and $\sum_{i \text{ even}} D_i = r$.
3) $\forall\, \ell \in \{2,\dots,2k-1\} : \sum_{i=2}^{\ell} M_i \geq 0$.

To this end, we have

$$\binom{k-1}{r} \cdot \binom{k-1}{r}$$

choices to allocate r "+1"s to $k-1$ places r "-1"s to $k-1$ places. But since these choices do not in general respect condition 3) we have to subtract the number of tuples $(M_2,\dots,M_{2k-1})$ that lead to a violation of 3). We show that these violating tuples are in bijective correspondence to all $(M'_2,\dots,M'_{2k-1})$ with

1') $M'_i \in \{-1,0\}$ for i odd and $M'_i \in \{0,1\}$ for i even.
2') $\sum_{i \text{ odd}} M'_i = -(r+1)$ and $\sum_{i \text{ even}} M'_i = r-1$.

The number of these $(M'_2,\dots,M'_{2k-1})$ is clearly given by

$$\binom{k-1}{r+1} \cdot \binom{k-1}{r-1}$$

so that the number of $(M_2,\dots,M_{2k-1})$ that *do* satisfy 1), 2) *and* 3) is given by

$$\binom{k-1}{r} \cdot \binom{k-1}{r} - \binom{k-1}{r+1} \cdot \binom{k-1}{r-1} = \frac{1}{r+1}\binom{k-1}{r}\binom{k}{r}$$

For the bijection, let $(M_2,\dots,M_{2k-1})$ be arbitrary with r "+1"s and r "-1"s so that 3) is violated. Then there is an odd index t such that $\sum_{i=2}^{t} M_i = -1$

for the first time. Then $(M_{t+1}, \dots, M_{2k-1})$ is a vector of even length which contains one more "+1" than "−1" entry. We will transform this vector to a vector $(M'_{t+1}, \dots, M'_{2k-1})$ by transforming the pairs $(M_{t+1}, M_{t+2}), \dots, (M_{2k-2}, M_{2k-1})$ as follows: If the pair is $(+1, -1)$ or $(0, 0)$, we leave it unchanged. A pair $(1, 0)$ will be changed to $(0, -1)$ and a pair $(0, -1)$ will be changed to $(1, 0)$. Then $(M'_{t+1}, \dots, M'_{2k-1})$ contains one more "−1" than "+1". Defining $(M'_2, \dots, M'_t) \coloneqq (M_2, \dots, M_t)$ we thus have created a vector $(M'_2, \dots, M'_{2k-1})$ satisfying 1') and 2'). On the other hand, any vector $(M'_2, \dots, M'_{2k-1})$ satisfying 1) and 2) has a first hitting time t of -1. Applying exactly the same transformation as before, we will then obtain a vector $(M_2, \dots, M_{2k-1})$ satisfying 1) and 2), but violating 3). □

Now to each $(\underline{s}, \underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ we can associate an $M \in \mathcal{M}(k, r)$, and this association completely determines the color structure of $(\underline{s}, \underline{t})$. To see how this is done, let us first analyze simple properties of a Eulerian cycle $(\underline{s}, \underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$. First, the graph $\mathcal{G}(\underline{s}, \underline{t})$ is a *double edged tree*, that is, it consists of k distinct double edges and has $k+1$ vertices, therefore is a tree in the regular sense after eliminating one of each of the double edges. Thus, the Eulerian cycle $(\underline{s}, \underline{t})$ crosses each edge twice, once in each direction, since a tree does not have circles. Further, $(\underline{s}, \underline{t})$ starts at the S-vertex s_1 and then alternates between S- and T-vertices until reaching s_1 again. We recall the representation of the cycle as in (4.28). Now we will record two crucial pieces of information into the MP-pds M. We start a tour at s_1 and move along the cycle. Whenever a down edge d_ℓ leaves the S-vertex s_ℓ for the last time along the walk, we set $D_\ell = -1$, otherwise $D_\ell = 0$. For example, we always have $D_1 = 0$, since s_1 is the last stop of the cycle. Additionally, whenever an up edge u_ℓ visits a new S-vertex $s_{\ell+1}$, which has not been visited before, we set $U_\ell = 1$ and otherwise $U_\ell = 0$. For example, we will always have $U_k = 0$, since this edge leads to the starting point s_1 again.

Let us argue that the tuple $(D_1, U_1, D_2, \dots, U_k)$ we just constructed satisfies conditions 1), 2) and 3) as above. Condition 1) is clearly satisfied. For condition 2), note that $(\underline{s}, \underline{t})$ has $r + 1$ S-nodes, out of which r – all except the vertex s_1 – were considered new, so that $\sum U_i = r$. Since the last edge u_k leads back to the vertex s_1, we must have left each of the r new S-vertices for a last time while on the cycle, so $\sum D_i = -r$. For condition 3) we realize only the r new nodes are left for a last time along the cycle, and before they can be left a last time (leading to a summand -1) they must have been discovered (leading to a summand $+1$). Thus, condition 3) holds.

Let us now see that each MP-pds $M \in \mathcal{M}(k, r)$ completely determines the color structure of an $(\underline{s}, \underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ by constructing a canonical $(\underline{s}, \underline{t})$ (that is, one with lowest vertex numbers possible) from M, and showing that we have only one choice for this construction. We set $s_1 = 1 = t_1$. Then whenever $U_\ell = +1$, this means that a new S-node is discovered, so we set $s_{\ell+1} \coloneqq \max(s_1, \dots, s_\ell) + 1$. On the other hand, if $U_\ell = 0$ then this means that $s_{\ell+1}$ shall be equal to one of the $s_1, \dots, s_\ell$, and so it must be equal to the s_i with $i \in \{1, \dots, \ell\}$ maximal from which t_ℓ was visited, since otherwise, the cycle

$(\underline{s},\underline{t})$ would contain a circle.

Now for $\ell \geq 2$, whenever $D_\ell = 0$, this means that s_ℓ is *not* visited the last time. But then t_ℓ must be different from $t_1,\dots,t_{\ell-1}$ since otherwise the cycle $(\underline{s},\underline{t})$ would contain a circle. Therefore, for $\ell \geq 2$, if $D_\ell = 0$ we set $t_\ell := \max(t_1,\dots,t_{\ell-1}) + 1$. Otherwise, if $D_\ell = -1$, this means that s_ℓ was visited for the last time by the cycle. But then t_ℓ must be equal to some element in $\{t_1,\dots,t_{\ell-1}\}$, since if t_ℓ were new, the edge $\{s_\ell, t_\ell\}$ would be new and there would then have to be a second edge traveling back from t_ℓ to s_ℓ, which would entail yet another visit of s_ℓ. So if $D_\ell = -1$, t_ℓ must be equal to some vertex in $\{t_1,\dots,t_{\ell-1}\}$, and then it must be equal to the last vertex with the highest index number in the set, from which s_ℓ was visited, since otherwise, again, the cycle $\{s_\ell, t_\ell\}$ would contain a circle.

As we saw, an $(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ is compatible with exactly one $M \in \mathcal{M}(k,r)$, and we then write $(\underline{s},\underline{t}) \sim M$. On the other hand, given an $M \in \mathcal{M}(k,r)$ we could create exactly one canonical $(\underline{s}^*,\underline{t}^*) \in \mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ compatible with M, which determines the color structure. All other $(\underline{s},\underline{t})$ compatible with M are then obtained by picking different vertex names for the $r+1$ vertices in $\underline{s}$ and $k-r$ vertices in $\underline{t}$, which yields a total of $(p)_{r+1}\cdot(n)_{k-r}$ tuples in $\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ compatible with each $M \in \mathcal{M}(k,r)$, where for any $\ell \leq m \in \mathbb{N}$, we set $(m)_\ell := m\cdot(m-1)\cdots(m-\ell+1)$. This analysis yields the following lemma:

Lemma 4.13. *The set $\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)})$ has a decomposition as follows:*

$$\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)}) = \dot{\bigcup_{M\in\mathcal{M}(k,r)}} \left\{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r} \mid (\underline{s},\underline{t}) \sim M\right\} \tag{4.40}$$

Further, for all $M \in \mathcal{M}(k,r)$,

$$\#\left\{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}_{r+1,k-r} \mid (\underline{s},\underline{t}) \sim M\right\} = (p)_{r+1}(n)_{k-r}\ , \tag{4.41}$$

such that by Lemma 4.12, (4.40) and (4.41), we obtain

$$\#\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)}) = \frac{1}{r+1}\binom{k-1}{r}\binom{k}{r}\cdot(p)_{r+1}(n)_{k-r}\ .$$

Proof. See the discussion before Lemma 4.13. □

Now since

$$(p)_{r+1}(n)_{k-r} = p\cdot\underbrace{\frac{(p-1)_r}{n^r}}_{\to y^r}\cdot\underbrace{n^r\cdot(n)_{k-r}}_{\sim n^k}\ ,$$

we find by Lemma 4.13 that

$$\frac{1}{pn^k}\#\mathcal{T}^{p,n}_{r+1,k-r}(\rho^{(k)}) \xrightarrow[n\to\infty]{} \frac{y^r}{r+1}\binom{k}{r}\binom{k-1}{r},$$

which is (4.39). Therefore, we have shown (4.38) which entails (4.35).

Step 2: Decay of central moments

In Step 1 we have seen that for fixed $k \in \mathbb{N}$, the expectation of

$$\left\langle \mu_n, x^k \right\rangle = \sum_{\rho \in \Pi(2k)} \frac{1}{pn^k} \sum_{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)} X_n(\underline{s},\underline{t}) \tag{4.42}$$

converges to the k-th moment of the MP distribution. In particular, we have seen that each of the finitely many summands

$$\frac{1}{pn^k} \sum_{(\underline{s},\underline{t}) \in \mathcal{T}^{p,n}(\rho)} X_n(\underline{s},\underline{t}) \tag{4.43}$$

converges to a constant in expectation. To show that the random moments in (4.42) converge almost surely to the moments of the MP distribution, it thus suffices – by Lemma 3.10 – to show that for all $\rho \in \Pi(2k)$, the variance of (4.43) decays summably fast. The variance of (4.43) is given by

$$\frac{1}{p^2 n^{2k}} \sum_{(\underline{s},\underline{t}),(\underline{s}',\underline{t}') \in \mathcal{T}^{p,n}(\rho)} \left[\mathbb{E} X_n(\underline{s},\underline{t}) X_n(\underline{s}',\underline{t}') - \mathbb{E} X_n(\underline{s},\underline{t}) \mathbb{E} X_n(\underline{s}',\underline{t}') \right]. \tag{4.44}$$

We see that whenever the Eulerian cycles $(\underline{s},\underline{t})$ and $(\underline{s}',\underline{t}')$ are edge-disjoint, the term in (4.44) vanishes due to independence. Therefore, it suffices to consider those cycles $(\underline{s},\underline{t})$ and $(\underline{s}',\underline{t}')$ which have at least one edge in common. To this end, denote for all $\ell \in \{1,\dots,2k\}$:

$$\mathcal{T}^{p,n}_{c(\ell)}(\rho) \coloneqq \big\{ ((\underline{s},\underline{t}),(\underline{s}',\underline{t}')) \in (\mathcal{T}^{p,n}(\rho))^2 \,\big|\, (\underline{s},\underline{t}) \text{ and } (\underline{s}',\underline{t}') \text{ have exactly } \ell \text{ edges in common} \big\}.$$

Then it is now our goal to show that for each $\rho \in \Pi(2k)$,

$$\frac{1}{p^2 n^{2k}} \sum_{((\underline{s},\underline{t}),(\underline{s}',\underline{t}')) \in \mathcal{T}^{p,n}_{c(\ell)}(\rho)} \left[\mathbb{E} X_n(\underline{s},\underline{t}) X_n(\underline{s}',\underline{t}') - \mathbb{E} X_n(\underline{s},\underline{t}) \mathbb{E} X_n(\underline{s}',\underline{t}') \right] \tag{4.45}$$

converges to zero summably fast. Before proceeding, we need to establish bounds on $\#\mathcal{T}^{p,n}_{c(\ell)}(\rho)$.

Lemma 4.14. *Let $\rho \in \Pi(2k)$ and $\ell \in [2k]$, then the following statements hold:*

i) For all $(\underline{s},\underline{t}),(\underline{s}',\underline{t}') \in \mathcal{T}^{p,n}(\rho)$ with at least ℓ common edges, it holds

$$\#(V(\underline{s},\underline{t}) \cup V(\underline{s}',\underline{t}')) \leq 1 + 2\sum_{i=1}^{2k} \rho_i - \ell$$

In particular,

$$\#\mathcal{T}^{p,n}_{c(\ell)}(\rho,\rho') \leq (2k)^{4k+2} (n \vee p)^{1+2\sum_{i=1}^{2k}\rho_i - \ell}$$

ii) If there is an $\ell \in [2k]$ odd with $\rho_\ell \geq 1$, then for all $(\underline{s},\underline{t}),(\underline{s}',\underline{t}') \in \mathcal{T}^{p,n}(\rho)$ with at least ℓ common edges, it holds

$$\#(V(\underline{s},\underline{t}) \cup V(\underline{s}',\underline{t}')) \leq 2\sum_{i=1}^{2k} \rho_i - \ell.$$

In particular,

$$\#\mathcal{T}^{p,n}_{c(\ell)}(\rho,\rho') \leq (2k)^{4k+2}(n \vee p)^{2\sum_{i=1}^{2k}\rho_i - \ell}.$$

Proof. For statement *ii)* we assume w.l.o.g. that $(\underline{s},\underline{t})$ has an odd edge. Since the graphs spanned by $(\underline{s},\underline{t})$ and $(\underline{s}',\underline{t}')$ share $l \geq 1$ common edges, we may take a tour around the joint Eulerian cycle, starting before a common edge, traveling first all edges of $(\underline{s},\underline{t})$ and then all edges of $(\underline{s}',\underline{t}')$. While walking the edges of $(\underline{s},\underline{t})$, we can see at most $\rho_1 + \ldots + \rho_{2k}$ different nodes by Lemma 4.10. Next, traveling all edges of $(\underline{s}',\underline{t}')$, at most all the single edges and first instances of m-fold edges with $m \in \{2,\ldots,2k\}$ of $(\underline{s}',\underline{t}')$ may discover a new node, but only if they have not been traversed before during the walk along $(\underline{s},\underline{t})$. Since we have ℓ common edges, we can see at most $\rho_1' + \ldots + \rho_{2k}' - \ell$ new nodes. We established the bounds on the number of vertices in *ii)*. The second statement in *ii)* follows immediately with Lemma 4.9 *ii)* by concatenating $(\underline{s},\underline{s}') \in [p]^{2k}$ and $(\underline{t},\underline{t}') \in [n]^{2k}$. For statement *i)* we proceed exactly in the same manner: Traveling $(\underline{s},\underline{t})$ we can see at most $1+\rho_1+\rho_2+\ldots+\rho_{2k}$ nodes by Lemma 4.10, then traveling $(\underline{s}',\underline{t}')$ we can see at most $\rho_1'+\ldots+\rho_{2k}'-\ell$ new nodes. Now apply Lemma 4.9 *ii)* again. □

Returning to (4.45), we distinguish the following cases:

Case 1: $\rho_1 \geq 1$

In this case, the term in (4.45) simplifies and we must argue that for each $\ell \in [2k]$,

$$\frac{1}{p^2 n^{2k}} \sum_{((\underline{s},\underline{t}),(\underline{s}',\underline{t}')) \in \mathcal{T}^{p,n}_{c(\ell)}(\rho)} \mathbb{E}X_n(\underline{s},\underline{t})X_n(\underline{s}',\underline{t}') \tag{4.46}$$

converges to zero summably fast. If $(\underline{s},\underline{t})$ and $(\underline{s}',\underline{t}')$ have $1 \leq \ell < \rho_1$ common edges, $\mathbb{E}X_n(\underline{s},\underline{t})X_n(\underline{s}',\underline{t}')$ vanishes, since not all single edges can be eliminated due to overlapping. Thus, it suffices to consider those $(\underline{s},\underline{t}),(\underline{s}',\underline{t}') \in \mathcal{T}^{p,n}(\rho)$ which have $\ell \geq \rho_1$ edges in common. Then

$$2\sum_{i=1}^{2k}\rho_i - \ell \leq 2\left(\rho_1 + \frac{2k-\rho_1}{2}\right) - \rho_1 \leq 2k,$$

so Lemma 4.14 *ii)* yields

$$\#\mathcal{T}^{p,n}_{c(\ell)}(\rho) \leq (2k)^{4k+2}(n \vee p)^{2\sum_{i=1}^{2k}\rho_i - \ell} \leq (2k)^{4k+2}(n \vee p)^{2k},$$

Since every summand in (4.46) is bounded by L_{4k}, it follows that (4.46) is $O(n^{-2})$, thus converges to zero summably fast.

Case 2: $\rho_1 = 0$
In this case, each summand in (4.45) is bounded by $L_{4k} + L_{2k}^2$. Further, we obtain for all $\rho \in \Pi(2k)$ with $\rho_1 = 0$ and $\ell \geq 1$ that

$$1 + 2\sum_{i=1}^{2k} \rho_i - \ell \leq 1 + 2 \cdot \frac{2k}{2} - 1 = 2k$$

so that by Lemma 4.14 i),

$$\#\mathcal{T}_{c(\ell)}^{p,n}(\rho) \leq (2k)^{4k+2}(n \vee p)^{1+2\sum_{i=1}^{2k}\rho_i - \ell} \leq (2k)^{4k+2}(n \vee p)^{2k},$$

so that the sum in (4.45) is $O(n^{-2})$, hence converges to zero summably fast.

5. The Stieltjes transform method

5.1. Motivation and basic properties

In order to analyze properties of random variables and their distributions, it is a common technique to use transforms of these distributions which make analysis more accessible due to their favorable algebraic structure. For example, a common and short proof of the central limit theorem is conducted by using the Fourier transform of the random variables involved, owing to the property that Fourier transforms handle convolutions particularly well, and the central limit theorem is about a sum of independent random variables.

In random matrix theory, however, when analyzing empirical spectral distributions of diverse matrix ensembles, it is desirable to use a tool for analysis that relates the behavior of the empirical spectral distribution back to the level of the entries of the matrices. For example, using the method of moments, one sees in equation (3.4) that the moments of the ESD σ_n of a random matrix X_n can be calculated through:

$$\forall\, k \in \mathbb{N}: \quad \langle \sigma_n, x^k \rangle = \frac{1}{n} \operatorname{tr}(X_n^k) = \frac{1}{n} \sum_{i_1,\ldots,i_k=1}^{n} X_n(i_1, i_2) X_n(i_2, i_3) \cdots X_n(i_k, i_1).$$

In other words, instead trying to work with an ESD directly, we can analyze its moments which allows us to work on the level of the matrix entries.

A tool that combines both worlds, that is, that provides the structure of a transform with favorable algebraic properties and that allows us to work on the level of the matrix entries is the so called Stieltjes transform:

Definition 5.1. *Let μ be a finite measure on $(\mathbb{R}, \mathcal{B})$. Then we define the Stieltjes transform S_μ of μ as the map*

$$\begin{aligned} S_\mu : \mathbb{C}\backslash\mathbb{R} &\longrightarrow \mathbb{C} \\ z &\longmapsto \int_{\mathbb{R}} \frac{1}{x - z} \mu(dx). \end{aligned}$$

We note that the Stieltjes transform is defined via a measure-theoretical integral over a complex-valued function. We assume the reader to be acquainted with measure-theoretical integration of real-valued functions on measure spaces and give a very short introduction to complex-valued integration in the form of one definition and two lemmata.

Definition 5.2. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $f : (\Omega, \mathcal{A}) \to \mathbb{C}$ measurable, then f is called* μ-integrable*, if the real-valued functions* $\operatorname{Re} f$ *and* $\operatorname{Im} f$ *both are μ- integrable. In this case, we define*

$$\int_\Omega f \, d\mu \coloneqq \int_\Omega \operatorname{Re} f \, d\mu + i \int_\Omega \operatorname{Im} f \, d\mu.$$

We will denote the space of $\mathbb{C}$-valued integrable functions as $\mathcal{L}_1(\mu, \mathbb{C})$.

It is worth noting the following lemma about the properties of the integral:

Lemma 5.3. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, then the following statements hold:*

1. *The map $\mathcal{L}_1(\mu, \mathbb{C}) \to \mathbb{C}$, $f \mapsto \int f \, d\mu$ is $\mathbb{C}$-linear.*
2. *$\forall f \in \mathcal{L}_1(\mu, \mathbb{C}) : \overline{\int f \, d\mu} = \int \overline{f} \, d\mu$.*
3. *$\forall f \in \mathcal{L}_1(\mu, \mathbb{C}) : \left| \int f \, d\mu \right| \leq \int |f| \, d\mu$.*

Proof. 1) follows by elementary calculations and 2) holds by the definition of the integral. To see 3), let $z \in \mathbb{C}$ with $|z| = 1$, such that $z \int f \mathrm{d}\mu = \left| \int f \mathrm{d}\mu \right|$, then it follows

$$\left| \int f \mathrm{d}\mu \right| = z \int f \mathrm{d}\mu = \int \operatorname{Re}(zf) \mathrm{d}\mu + i \underbrace{\int \operatorname{Im}(zf) \mathrm{d}\mu}_{=0} \leq \int |zf| \mathrm{d}\mu = \int |f| \mathrm{d}\mu. \qquad \square$$

Lemma 5.4 (Lebesgue's Dominated Convergence Theorem)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $(f_n)_n, f : \Omega \to \mathbb{C}$ be measurable with $f_n \to f$ μ-almost everywhere. If there exists a μ-integrable $g : \Omega \to \mathbb{R}$ with $|f_n| \leq g$ μ-almost everywhere for all n, then f is μ-integrable and it holds*

$$\lim_{n \to \infty} \int |f - f_n| d\mu = 0,$$

so that in particular

$$\lim_{n \to \infty} \int f_n \, d\mu = \int f \, d\mu.$$

Proof. Certainly, $|\operatorname{Re} f_n|, |\operatorname{Im} f_n| \leq |f_n| \leq |g|$ and $\operatorname{Re} f_n \to \operatorname{Re} f$, $\operatorname{Im} f_n \to \operatorname{Im} f$ μ-almost everywhere. Also, $|f - f_n| \leq |\operatorname{Re} f - \operatorname{Re} f_n| + |\operatorname{Im} f - \operatorname{Im} f_n|$. Now for real-valued measurable functions, the theorem is assumed to be known. See [35, p. 158] for a reference. $\square$

The following lemma studies the Stieltjes transform $S_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z} \mu(\mathrm{d}x)$. Note that we do not have to consider the trivial case where $\mu \equiv 0$, since in this case, $S_\mu \equiv 0$. Notationally, we set $\mathbb{C}_+ \coloneqq \{z \in \mathbb{C} \mid \operatorname{Im}(z) > 0\}$.

Lemma 5.5. *Let μ be a finite measure on $(\mathbb{R}, \mathcal{B})$ with $\mu(\mathbb{R}) > 0$ and S_μ be its Stieltjes transform. Further, let $E \in \mathbb{R}$, $\eta \in \mathbb{R}\backslash\{0\}$ and $z \coloneqq E + i\eta$, then we obtain:*

i) For any $x \in \mathbb{R}$ we find: $\frac{1}{x-z} = \frac{x-E}{(x-E)^2+\eta^2} + i\frac{\eta}{(x-E)^2+\eta^2}$.
ii) $\operatorname{Re} S_\mu(z) = \int \frac{x-E}{(x-E)^2+\eta^2}\mu(dx)$ and $\operatorname{Im} S_\mu(z) = \int \frac{\eta}{(x-E)^2+\eta^2}\mu(dx)$.
iii) $\operatorname{Im}(z) \gtrless 0 \Leftrightarrow \operatorname{Im} S_\mu(z) \gtrless 0$.
iv) $S_\mu(\overline{z}) = \overline{S_\mu(z)}$.
v) S_μ is uniquely determined by its restriction $S_\mu : \mathbb{C}_+ \to \mathbb{C}_+$.
vi) $|S_\mu(z)| \leq \frac{\mu(\mathbb{R})}{|\operatorname{Im}(z)|}$
vii) S_μ is holomorphic.
viii) In particular, S_μ is continuous, can be represented by a power series around any $z_0 \in \mathbb{C}\backslash\mathbb{R}$, and is infinitely often differentiable.

Proof. Statement i) is obvious, ii) follows from i) by definition of the complex-valued integral, iii) follows directly from ii) and so does iv) in combination with the construction of the integral. Statement v) follows directly from iii) and iv), and vi) follows from

$$\left|\frac{1}{x-z}\right| = \frac{1}{|x-z|} \leq \frac{1}{|\operatorname{Im}(x-z)|} = \frac{1}{|\operatorname{Im}(z)|}.$$

To show statement vii), let $(z_n)_n$ and $z \in \mathbb{C}\backslash\mathbb{R}$ with $z_n \to z$, but $z_n \neq z$ be arbitrary, then:

$$\begin{aligned}\frac{S_\mu(z_n) - S_\mu(z)}{z_n - z} &= \frac{1}{z_n - z}\int \frac{1}{x - z_n} - \frac{1}{x-z}\mu(\mathrm{d}x) \\ &= \frac{1}{z_n - z}\int \frac{z_n - z}{(x-z_n)(x-z)}\mu(\mathrm{d}x) \xrightarrow[n\to\infty]{} \int \frac{1}{(x-z)^2}\mu(\mathrm{d}x)\end{aligned}$$

by dominated convergence, since for some $C > 0$ and all $n \in \mathbb{N}$,

$$\left|\frac{1}{(x-z_n)(x-z)}\right| \leq \frac{1}{|\operatorname{Im}(z_n)||\operatorname{Im}(z)|} \leq C,$$

for convergent sequences are bounded. □

Theorem 5.6 (Retrieval of Measure)**.** *For any bounded interval $I \subseteq \mathbb{R}$ with end points $\alpha < \beta$, we obtain the following:*

$$\mu((\alpha,\beta)) + \frac{1}{2}(\mu(\{\alpha\}) + \mu(\{\beta\})) = \lim_{\eta \searrow 0}\frac{1}{\pi}\int_I \operatorname{Im} S_\mu(E+i\eta)\lambda(dE).$$

Proof. Let I be an interval with end points $\alpha < \beta$ and $\eta > 0$. Then we obtain via Fubini:

$$\begin{aligned}\frac{1}{\pi}\int_I \operatorname{Im} S_\mu(E+i\eta)\lambda(\mathrm{d}E) &= \frac{1}{\pi}\int_I\int_{\mathbb{R}} \frac{\eta}{(x-E)^2+\eta^2}\mu(\mathrm{d}x)\lambda(\mathrm{d}E) \\ &= \frac{1}{\pi}\int_{\mathbb{R}}\int_I \frac{\eta}{(x-E)^2+\eta^2}\lambda(\mathrm{d}E)\mu(\mathrm{d}x)\end{aligned}$$

$$= \frac{1}{\pi} \int_{\mathbb{R}} \int_{\alpha}^{\beta} \frac{\eta}{(x-E)^2+\eta^2} \mathrm{d}E \mu(\mathrm{d}x).$$

Now since

$$\begin{aligned} \int_{\alpha}^{\beta} \frac{\eta}{(x-E)^2+\eta^2} \mathrm{d}E &= \frac{1}{\eta} \int_{\alpha}^{\beta} \frac{1}{(\frac{E-x}{\eta})^2+1} \mathrm{d}E \\ &= \int_{\frac{\alpha-x}{\eta}}^{\frac{\beta-x}{\eta}} \frac{1}{E^2+1} \mathrm{d}E \\ &= \arctan\left(\frac{\beta-x}{\eta}\right) - \arctan\left(\frac{\alpha-x}{\eta}\right), \end{aligned}$$

and $\arctan : \mathbb{R} \to (-\frac{\pi}{2}, +\frac{\pi}{2})$ is strictly increasing with $\lim_{x\to\pm\infty} \arctan(x) = \pm\frac{\pi}{2}$, we obtain

$$\lim_{\eta \searrow 0} \left[\arctan\left(\frac{\beta-x}{\eta}\right) - \arctan\left(\frac{\alpha-x}{\eta}\right)\right] = \begin{cases} \pi & \text{if } x \in (\alpha,\beta) \\ 0 & \text{if } x \notin [\alpha,\beta] \\ \frac{\pi}{2} & \text{if } x = \alpha \vee x = \beta. \end{cases}$$

Thus, by dominated convergence we find

$$\begin{aligned} &\lim_{\eta \searrow 0} \frac{1}{\pi} \int_I \operatorname{Im} S_\mu(E+i\eta) \mathbb{\Lambda}(\mathrm{d}E) \\ &= \lim_{\eta \searrow 0} \frac{1}{\pi} \int_{\mathbb{R}} \arctan\left(\frac{\beta-x}{\eta}\right) - \arctan\left(\frac{\alpha-x}{\eta}\right) \mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}} \mathbb{1}_{(\alpha,\beta)}(x) + \frac{1}{2}\mathbb{1}_{\{\alpha,\beta\}}(x) \mu(\mathrm{d}x) \\ &= \mu((\alpha,\beta)) + \frac{1}{2}(\mu(\{\alpha\}) + \mu(\{\beta\})) \end{aligned}$$

□

The previous theorem and the following corollary are similar to Theorem 2.4.3 in [4]. As usual, for a subset I of a topological space, we denote by ∂I its boundary, which is a concept we assume to be known to the reader.

Corollary 5.7. *For any bounded interval $I \subseteq \mathbb{R}$ with $\mu(\partial I) = 0$, we find:*

$$\mu(I) = \lim_{\eta \searrow 0} \frac{1}{\pi} \int_I \operatorname{Im} S_\mu(E+i\eta) \mathbb{\Lambda}(dE).$$

Thus, any finite measure μ on $(\mathbb{R}, \mathcal{B})$ is uniquely determined by S_μ. In other words, $\mu \mapsto S_\mu$ is injective.

Proof. The convergence statement follows from Theorem 5.6. In particular, if μ and ν are finite measures on $(\mathbb{R}, \mathcal{B})$ with $S_\mu = S_\nu$, then for any bounded interval $I \subseteq \mathbb{R}$ with $\mu(\partial I) = 0 = \nu(\partial I)$, we have $\mu(I) = \nu(I)$. Therefore, $\mu = \nu$ by Lemma 2.6. □

The last theorem and its corollary suggest that for any finite measure μ on $(\mathbb{R}, \mathcal{B})$ and $\eta > 0$ small, $E \mapsto \frac{1}{\pi} \operatorname{Im} S_\mu(E + i\eta)$ acts as a Lebesgue density for (a measure approximating) μ. In particular, even measures that do not possess a Lebesgue density (for example, all empirical measures) can be approximated in this way by using the Stieltjes transform. In Section 5.3 we will see how this can be made precise.

5.2. The Stieltjes transform and weak convergence

For any finite measure μ, S_μ carries all the information of μ (cf. Corollary 5.7). Therefore, it is not surprising that this tool can be used particularly well to analyze weak convergence of probability measures. The following theorem generalizes Theorem 2.4.4 in [4].

Theorem 5.8 (Convergence Theorem)**.** *Let $Z \subseteq \mathbb{C}\backslash\mathbb{R}$ be a subset that has an accumulation point in $\mathbb{C}\backslash\mathbb{R}$ (which is not necessarily an element of Z itself). Then the following statements hold:*

1. *Let $(\mu_n)_n$ in $\mathcal{M}_1(\mathbb{R})$, such that for all $z \in Z$ we find that $S(z) := \lim_{n\to\infty} S_{\mu_n}(z)$ exists. Then there is a sub-probability measure μ with $\mu_n \to \mu$ vaguely and $S_\mu = S$.*
2. *Let $(\mu_n)_n$ and μ in $\mathcal{M}_1(\mathbb{R})$, then we find:*

$$\mu_n \to \mu \text{ weakly } \Leftrightarrow \ S_{\mu_n}(z) \to S_\mu(z) \text{ for all } z \in Z.$$

3. *Let $(\mu_n)_n$ be random probability measures and μ be a deterministic probability measure, then:*
   a) *$\mu_n \to \mu$ weakly in expectation $\Leftrightarrow \mathbb{E}S_{\mu_n}(z) \to S_\mu(z)$ for all $z \in Z$.*
   b) *$\mu_n \to \mu$ weakly in probability $\Leftrightarrow S_{\mu_n}(z) \to S_\mu(z)$ in probability for all $z \in Z$.*
   c) *$\mu_n \to \mu$ weakly almost surely $\Leftrightarrow$ [$S_{\mu_n}(z) \to S_\mu(z)$ almost surely] for all $z \in Z$.*

Proof. 1. Let $(\mu_n)_{n\in J}$ be an arbitrary subsequence of $(\mu_n)_{n\in\mathbb{N}}$. Due to Lemma 2.15, there exists a subsequence $(\mu_n)_{n\in I}$, $I \subseteq J$, such that $\mu_n \to \mu$ vaguely for $n \in I$ and a sub-probability measure μ. Since $x \mapsto \frac{1}{x-z}$ vanishes at $\pm\infty$, it follows $S_{\mu_n}(z) \to S_\mu(z)$ for $n \in I$ for all $z \in Z$ (cf. Lemma 2.10). Therefore, $S(z) = S_\mu(z)$ for all $z \in Z$. If ν is another subsequential limit of $(\mu_n)_{n\in J}$, we find by the same argument that $S_\mu(z) = S(z) = S_\nu(z)$ for all $z \in Z$. This implies $S_\mu = S_\nu$, since Stieltjes transforms are holomorphic. Therefore, $\mu = \nu$ by Theorem 5.6. By Lemma 2.9, we find $\mu_n \to \mu$ vaguely for $n \in \mathbb{N}$.
2. Since $x \mapsto \frac{1}{x-z}$ is continuous, "$\Rightarrow$" is obvious. To show "$\Leftarrow$", statement 1 yields that $\mu_n \to \mu$ vaguely, thus $\mu_n \to \mu$ weakly, since all measures involved are probability measures (cf. Lemma 2.14).
3.a) This follows directly from statement 2, considering

$$\mathbb{E}S_{\mu_n}(z) = \mathbb{E}\int \frac{1}{x-z}\mu_n(\mathrm{d}x) = \int \frac{1}{x-z}\mathbb{E}\mu_n(\mathrm{d}x) = S_{\mathbb{E}\mu_n}(z),$$

where we used Theorem 2.20.
3.c) If $\mu_n \to \mu$ weakly on a measurable set A with $\mathbb{P}(A) = 1$, then we have on A that for all $z \in Z$ we find $S_{\mu_n}(z) \to S_\mu(z)$ (by statement 2). This shows "⇒", and to show "⇐", fix a sequence $(z_k)_k$ in Z that converges to some $z \in \mathbb{C}\backslash\mathbb{R}$. For all $k \in \mathbb{N}$ we find a measurable set A_k with $\mathbb{P}(A_k) = 1$ on which $S_{\mu_n}(z_k) \to S_\mu(z_k)$ as $n \to \infty$. Then $A := \cap_{k\in\mathbb{N}} A_k$ is measurable with $\mathbb{P}(A) = 1$, and on A we find that for all $z \in Z' := \{z_k | k \in \mathbb{N}\}$ we have $S_{\mu_n}(z) \to S_\mu(z)$. Since Z' has an accumulation point in $\mathbb{C}\backslash\mathbb{R}$, we find on the set A that $\mu_n \to \mu$ weakly by statement 2.
3.b) The direction "⇒" is trivial since $x \mapsto \frac{x-E}{(x-E)^2+\eta^2}$ and $x \mapsto \frac{\eta}{(x-E)^2+\eta^2}$ are bounded and continuous (cf. Theorem 2.25). For "⇐" we let $f \in \mathcal{C}_b(\mathbb{R})$ be arbitrary. Then we need to show that $\langle \mu_n, f\rangle \to \langle \mu, f\rangle$ in probability. Let $J \subseteq \mathbb{N}$ be a subsequence, then by Lemma 2.28, we find a subsequence $I \subseteq J$ and a measurable set N with $\mathbb{P}(N) = 0$, such that for $(z_k)_k$ fixed as in the proof of 3.c):

$$\forall\, \omega \in \Omega\backslash N : \forall\, k \in \mathbb{N} : S_{\mu_n(\omega)}(z_k) \xrightarrow[n\in I]{} S_{\mu(\omega)}(z_k).$$

Therefore, it follows with statement 3.c) that $\mu_n \xrightarrow[n\in I]{} \mu$ almost surely, so in particular $\langle \mu_n, f\rangle \xrightarrow[n\in I]{} \langle \mu, f\rangle$ almost surely. Then $\langle \mu_n, f\rangle \xrightarrow[n\in\mathbb{N}]{} \langle \mu, f\rangle$ in probability by Lemma 2.27. □

We refer the reader to Remark 2.26 for an explanation on the use of brackets [...] in Theorem 5.8 3. *c*). Also, due to Lemma 5.5 *iv*), it would have been enough to restrict one's attention to the set $\{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$ in Theorem 5.8.

5.3. The imaginary part of the Stieltjes transform

In Corollary 5.7 we saw that if $\mu \in \mathcal{M}_1(\mathbb{R})$, then for a small $\eta > 0$, the function $E \mapsto \frac{1}{\pi} \mathrm{Im}\, S_\mu(E + i\eta)$ should be the Lebesgue density of a probability measure on $(\mathbb{R}, \mathcal{B})$ that approximates μ well. But so far, we do not even know whether $E \mapsto \frac{1}{\pi} \mathrm{Im}\, S_\mu(E+i\eta)$ yields a density of a *probability measure* at all. How can this intuition be portrayed in the right context, and is there a connection to the weak convergence results of Section 5.2? This section aims to shed light onto these aspects. First, we will rigorously delve into convolution of probability measures, for a reference see [12] or [3]. Second, we will introduce kernel density estimators, which motivate further the use of the Stieltjes transform when analyzing ESDs of random matrices. We begin by making the following definition:

Definition 5.9. *Let μ and ν be probability measures on $(\mathbb{R}, \mathcal{B})$ and $f, g : \mathbb{R} \to \mathbb{R}$ Lebesgue-density functions (i.e. $h \geq 0$ and $\int h d\lambda = 1$, $h \in \{f, g\}$).*

i) The convolution of the probability measures μ and ν is defined as $\mu * \nu := (\mu \otimes \nu)^+$. Here, $\mu \otimes \nu$ is the product measure on $(\mathbb{R}^2, \mathcal{B}^2)$, $+ : \mathbb{R}^2 \to \mathbb{R}$ is the addition map, and $(\mu \otimes \nu)^+$ is the push-forward of the product measure under the addition map.

ii) The convolution of the density f and the probability measure ν is defined as the function $f * \nu : \mathbb{R} \to \mathbb{R}$ with

$$\forall\, x \in \mathbb{R} :\ (f * \nu)(x) \coloneqq \int_{\mathbb{R}} f(x-y)\nu(dy).$$

iii) The convolution of the densities f and g is the function $f * g : \mathbb{R} \to \mathbb{R}$ with

$$\forall\, x \in \mathbb{R} :\ (f * g)(x) \coloneqq \int_{\mathbb{R}} f(x-y)g(y)\lambda\!\!\!\lambda(dy).$$

Note that in ii) and iii) above, the definitions of the convolution are to be understood for $\lambda\!\!\!\lambda$-almost all $x \in \mathbb{R}$, since the respective integrals are well-defined only for $\lambda\!\!\!\lambda$-almost all $x \in \mathbb{R}$, which can be observed via Fubini/Tonelli. The convolutions are understood to equal zero on the respective sets of measure zero.

Lemma 5.10. *In the situation of Definition 5.9, we make the following observations (point x here is with respect to point x in Definition 5.9, $x \in \{i), ii), iii)\}$):*

i) The convolution is a commutative binary operation on the space of probability measures. The neutral element is given by δ_0, the Dirac measure in 0. Further, the following formula holds:

$$\forall\, B \in \mathcal{B} : (\mu * \nu)(B) = \int_{\mathbb{R}} \mu(B-y)\nu(dy).$$

ii) $f * \nu$ is a Lebesgue-density for the convolution $(f\lambda\!\!\!\lambda) * \nu$, that is, $(f\lambda\!\!\!\lambda) * \nu = (f * \nu)\lambda\!\!\!\lambda$.

iii) $f * g$ is a Lebesgue-density for the convolution $(f\lambda\!\!\!\lambda) * (g\lambda\!\!\!\lambda)$, that is, $(f\lambda\!\!\!\lambda) * (g\lambda\!\!\!\lambda) = (f * g)\lambda\!\!\!\lambda$.

Proof. The proof follows from elementary considerations and is left to the reader, see also [12]. □

The following lemma will capture a very important property of the convolution:

Lemma 5.11. *The convolution of probability measures on $(\mathbb{R}, \mathcal{B})$ is continous with respect to weak convergence. That is, if $(\mu_n)_n$, $(\nu_n)_n$, μ and ν are probability measures on $(\mathbb{R}, \mathcal{B})$ with $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly, then $\mu_n * \nu_n \to \mu * \nu$ weakly.*

Proof. With [11, p. 23] it follows that $\mu_n \otimes \nu_n \to \mu \otimes \nu$. Now if $f \in \mathcal{C}_b(\mathbb{R})$ is arbitrary, then we also have that $(x, y) \mapsto f(x+y)$ is a continuous and bounded function on $\mathbb{R}^2$, so

$$\int_{\mathbb{R}} f \mathrm{d}(\mu_n * \nu_n) = \int_{\mathbb{R}^2} (f \circ +) \mathrm{d}(\mu_n \otimes \nu_n) \xrightarrow[n\to\infty]{} \int_{\mathbb{R}^2} (f \circ +) \mathrm{d}(\mu \otimes \nu) = \int_{\mathbb{R}} f \mathrm{d}(\mu * \nu).$$

□

Now, we will bring the Stieltjes transform into play:

Definition 5.12. *For all $\eta > 0$, we define the Cauchy kernel $P_\eta : \mathbb{R} \to \mathbb{R}$ as the function with*

$$\forall\, x \in \mathbb{R} : P_\eta(x) := \frac{1}{\pi} \frac{\eta}{x^2 + \eta^2},$$

which is the λ-density function of the Cauchy distribution with scale parameter η.

We will collect a quick lemma before proceeding:

Lemma 5.13. *As $\eta \searrow 0$, we find $(P_\eta \lambda) \to \delta_0$ weakly.*

Proof. The characteristic function of the measure $P_\eta \lambda$ is given by $t \mapsto e^{-\eta|t|}$, see [35, p. 337] or [45, p. 208]. Fixing $t \in \mathbb{R}$ and letting $\eta \to 0$ will yield the statement, since e^0 is the characteristic function of δ_0. □

Now, as we see, for any probability measure μ on $(\mathbb{R}, B)$, we have

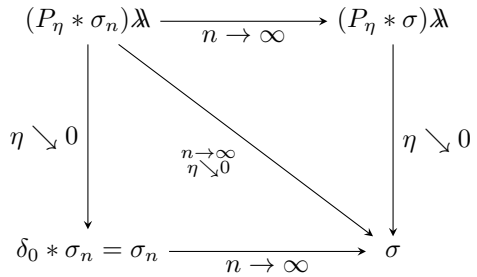
$$\frac{1}{\pi} \operatorname{Im} S_\mu(E + i\eta) = \int_{\mathbb{R}} \frac{1}{\pi} \frac{\eta}{(E-x)^2 + \eta^2} \mu(\mathrm{d}x) = (P_\eta * \mu)(E)$$

Therefore, $1/\pi \operatorname{Im} S_\mu(\cdot + i\eta)$ is the convolution of the density P_η with μ and thus a Lebesgue-density for the probability measure $(P_\eta \lambda) * \mu$. In particular, as $\eta \searrow 0$ we have that

$$\frac{1}{\pi} \operatorname{Im} S_\mu(\cdot + i\eta)\lambda = (P_\eta \lambda) * \mu \longrightarrow \delta_0 * \mu = \mu \quad \text{weakly.}$$

This immediately proves Corollary 5.7 again (using the Portmanteau theorem). But due to continuity of the convolution, we can say much more:

Assume that $(\sigma_n)_n$ is a sequence of ESDs of random matrices, so that σ_n converges almost surely to the semicircle distribution σ. We assume this convergence takes place on a measurable set A with $\mathbb{P}(A) = 1$. Then we find on A that the following commutative diagram holds, where all arrows indicate weak convergence:



In particular, the diagonal arrow says that we obtain weak convergence $(P_{\eta_n} * \sigma_n)\lambda \to \sigma$ as $n \to \infty$ for any sequence $\eta_n \searrow 0$. This is an interesting result, but it does not tell us if also densities align. More concretely, write $\sigma = f_\sigma \lambda$, then from $(P_\eta * \sigma_n)\lambda \to f_\sigma \lambda$ weakly we cannot infer that also $P_\eta * \sigma_n \to f_\sigma$ in some sense, for example in $\| \cdot \|_\infty$ over a specified compact interval. This is desirable

since it allows conclusion about *local* estimation of σ_n by σ. If $\eta = \eta_n$ drops too quickly to zero as $n \to \infty$, then $(P_{\eta_n} * \sigma_n)$ will have steep peaks at each eigenvalue, thus will not approximate the density of the semicircle distribution uniformly. This "problem" is typical for kernel density estimators in general (see [9], in particular their Section 11.2.1, or [45], especially their Remark 11.2.10), which we will introduce next.

Definition 5.14. *A kernel K is a Lebesgue-probability-density function $\mathbb{R} \to \mathbb{R}$, that is, K is non-negative and*

$$\int_{\mathbb{R}} K(y) \lambda(dy) = 1.$$

Further, if K is a kernel and $h > 0$, we define K_h as the kernel with $K_h(x) = \frac{1}{h} K(\frac{x}{h})$ for all $x \in \mathbb{R}$ and call K_h the kernel K at bandwidth h. In particular, $K = K_1$.

In above definition, it is clear that K_h is a kernel if K is a kernel and $h > 0$. An example of a kernel is the Cauchy kernel P_1 from Definition 5.12, which yields the standard Cauchy distribution. We have for all $x \in \mathbb{R}$ and $\eta > 0$:

$$P_1(x) = \frac{1}{\pi} \frac{1}{x^2 + 1} \quad \text{and} \quad P_\eta(x) = \frac{1}{\pi \eta} \frac{1}{\left(\frac{x}{\eta}\right)^2 + 1} = \frac{1}{\pi} \frac{\eta}{x^2 + \eta^2}.$$

Now given a vector $v = (v_1, \ldots, v_n)$ of real-valued observations, we are interested in constructing a Lebesgue-density that describes the experiment of drawing uniformly at random from these observations, in other words that approximates the empirical probability measure

$$\nu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{v_i}. \tag{5.1}$$

This can be done with help of a kernel K, which is oftentimes chosen to be unimodal and symmetric around 0, just as the Cauchy kernel P_1.

Definition 5.15. *The* kernel density estimator with kernel K and bandwidth $h > 0$ *for an empirical measure ν as in (5.1) is the Lebesgue-density given by the convolution $K_h * \nu$, thus*

$$\begin{aligned} K_h * \nu : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto (K_h * \nu)(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - v_i) = \frac{1}{nh} \sum_{i=1}^{n} K_1\left(\frac{x - v_i}{h}\right) \end{aligned}$$

Heuristically speaking, the concept works in the following way: The center of the kernel is placed upon each observation, whose influence (i.e. probability mass of $1/n$) is smoothed over its neighborhood. The size of this neighborhood is governed by the bandwidth h: A small h will restrain the probability mass

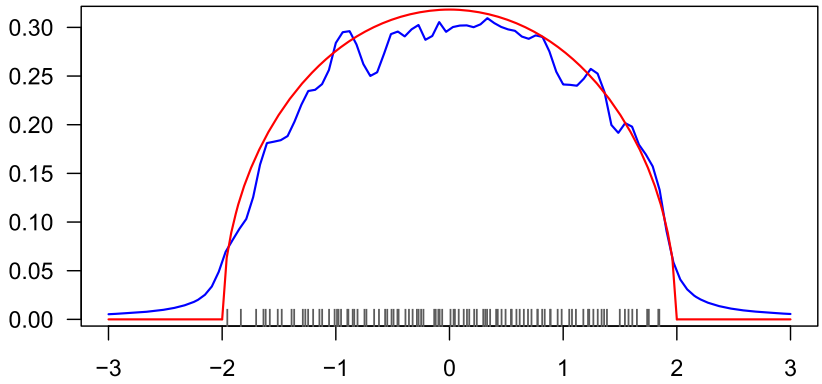


FIG 3. *Red line:* f_σ. *Blue line:* $\frac{1}{\pi}\operatorname{Im} S_{\sigma_{100}}(\cdot + i\eta_1) = P_{\eta_1} * \sigma_{100}$. *Grey bars: eigenvalue locations.*

of $1/n$ to be closer to its observation, whereas a larger h will result in a wider spread of probability mass. Therefore, a smaller h will result in a peaky density function (with steep peaks at the observation), whereas a larger h will result in a smoother density function.

We now assume we are given an empirical spectral distribution σ_N from an $n \times n$ random matrix X_n. The kernel density estimator at location $E \in \mathbb{R}$ for σ_n with kernel P_1 at bandwidth $\eta > 0$ is then given by

$$(P_\eta * \sigma_n)(E) = \frac{1}{n\eta} \sum_{i=1}^{n} \frac{1}{\pi} \frac{1}{\left(\frac{E-\lambda_i^{X_n}}{\eta}\right)^2 + 1} = \frac{1}{\pi} \operatorname{Im} S_{\sigma_n}(E + i\eta).$$

This gives the imaginary part of the Stieltjes transform the new role of a kernel density estimator for the empirical spectral distribution. Let us conduct a simulation study for $n = 100$. Let A_{100} be a symmetric 100×100 random matrix with independent Rademacher distributed variables in the upper half triangle, including the main diagonal. Let $X_{100} := \frac{1}{\sqrt{100}} A_{100}$. Denote by σ_{100} the empirical spectral distribution of X_{100}. Further, we define the bandwidths $\eta_1 := n^{-1/2} = 1/10$ and $\eta_2 := n^{-1} = 1/100$. With respect to the commutative diagram after Lemma 5.13 and the discussion below it, let us analyze how well $P_{\eta_1} * \sigma_{100}$ and $P_{\eta_2} * \sigma_{100}$ can be approximated by the density of the semicircle distribution, f_σ, in Figures 3 and 4, which are based on the same simulation outcome.

As we see, considering that we are in the case of a very low $n = 100$, we already obtain a decent approximation by the semicircle density in Figure 3. Reducing the scale from η_1 to η_2 we obtain the result in Figure 4. There we observe that for the smaller bandwidth parameter η_2, we do not obtain a useful approximation by the semicircle density anymore. Indeed, the scale n^{-1} is too fast to obtain uniform convergence of the estimated density to the target density, whereas a scale of $n^{\gamma-1}$ for any $\gamma \in (0, 1)$ would be sufficient. Nevertheless, Figure 4 displays nicely how the kernel density estimator works: A closer look –

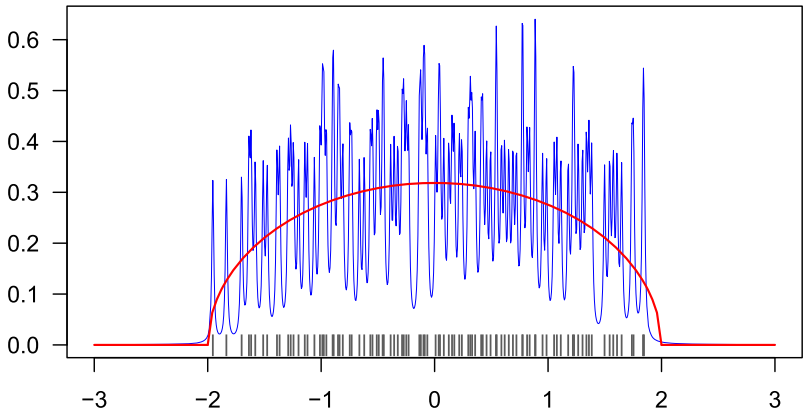


FIG 4. *Red line:* f_σ. *Blue line:* $\frac{1}{\pi}\operatorname{Im} S_{\sigma_{100}}(\cdot + i\eta_2) = P_{\eta_2} * \sigma_{100}$. *Grey bars: Eigenvalue locations.*

in particular to the edges of the bulk – shows how the probability mass of each individual eigenvalue is spread around its neighborhood.

5.4. The Stieltjes transform of ESDs of random matrices

As we motivated the Stieltjes transform in the beginning of this chapter, it is possible to relate the Stieltjes transform of an ESD of a random matrix to the entries of the random matrix. We will now see how this is done. Notationally, as the Stieltjes transform of the semicircle distribution received the special letter $s \coloneqq S_\sigma$, the Stieltjes transform of an ESD σ_n of an $n \times n$ random matrix X_n is denoted by $s_n \coloneqq S_{\sigma_n}$. The following theorem summarizes the findings of this section (see also [5, p. 470–472]).

Theorem 5.16. *Let X_n be a random $n \times n$ matrix with ESD σ_n.*

i) For all $z \in \mathbb{C}\backslash\mathbb{R}$ we find:

$$s_n(z) = S_{\sigma_n}(z) = \frac{1}{n}\operatorname{tr}(X_n - z)^{-1} = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{X_n(k,k) - z - x_k^*(X_n^{(k)} - z)^{-1}x_k}.$$

ii) For $z = E + i\eta$, where $E \in \mathbb{R}$ and $\eta > 0$, we obtain for all $k \in \{1, \dots, n\}$:

$$\left|\operatorname{tr}(X_n - z)^{-1} - \operatorname{tr}(X_n^{(k)} - z)^{-1}\right| \leq \frac{1}{\eta}.$$

Here, $X_n^{(k)}$ denotes the k-th principal minor of X_n (thus an $(n-1) \times (n-1)$ matrix) and x_k the k-th column of X_n without the k-th entry (thus an $(n-1)$-vector).

Proof. *i)* The first equality is just a notational convention. For the second equality, let $\overline{\lambda}_1, \ldots, \lambda_n$ be the eigenvalues of X_n, then by the spectral theorem for normal operators, $\frac{1}{\lambda_1 - z}, \ldots, \frac{1}{\lambda_n - z}$ are the eigenvalues of $(X_n - z)^{-1}$. Since for normal matrices, the trace yields the sum of the eigenvalues, we conclude

$$S_{\sigma_n}(z) = \int_{\mathbb{R}} \frac{1}{x - z} \sigma_n(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i - z} = \frac{1}{n} \operatorname{tr}(X_n - z)^{-1}.$$

It remains to prove the last equality of statement *i)* of Theorem 5.16: X_n and all $X_n^{(k)}$ are self-adjoint, thus $X_n - z$ and $X_n^{(k)} - z = (X_n - z)^{(k)}$ are invertible for all k. We also know that the k-th column (resp. row) of X_n without the k-th entry is also the k-th column (resp. row) of $(X_n - z)$ without the k-th entry. Therefore, the statement follows directly with Lemma 5.18 below.

ii) By Lemma 5.19 below, we know that

$$\begin{aligned} \left| \operatorname{tr}(X_n - z)^{-1} - \operatorname{tr}(X_n^{(k)} - z)^{-1} \right| &= \left| \frac{1 + x_k^* (X_n^{(k)} - z)^{-2} x_k}{X_n(k,k) - z - x_k^* (X_n^{(k)} - z)^{-1} x_k} \right| \\ &\leq \frac{1 + |x_k^* (X_n^{(k)} - z)^{-2} x_k|}{|-\eta - \operatorname{Im}(x_k^* (X_n^{(k)} - z)^{-1} x_k)|} \end{aligned} \tag{5.2}$$

where x_k denotes the k-th column of X_n without the k-th entry. We also used that $X_n(k,k) \in \mathbb{R}$, since X_n is self-adjoint. We proceed by inspecting the numerator and the denominator separately. For the numerator, Let U be unitary such that $U X_n^{(k)} U^* = \operatorname{diag}(\lambda_1, \ldots, \lambda_{n-1}) =: D$, where $\lambda_1, \ldots, \lambda_{n-1}$ are the eigenvalues of $X_n^{(k)}$. Since $X_n^{(k)}$ is self-adjoint, these eigenvalues are real, and such a U actually exists. Set $x_k^* U^* = (y_1, \ldots, y_{n-1})$, then we get

$$\begin{aligned} x_k^* (X_n^{(k)} - z)^{-2} x_k &= x_k^* (U^* (D - z) U)^{-2} x_k = x_k^* U^* (D - z)^{-2} U x_k \\ &= \sum_{\ell=1}^{n-1} \frac{|y_\ell|^2}{(\lambda_\ell - z)^2} \underset{|\ldots|}{\leq} \sum_{\ell=1}^{n-1} \frac{|y_\ell|^2}{(\lambda_\ell - E)^2 + \eta^2} \\ &= x_k^* [(X_n^{(k)} - E I_{n-1})^2 + v^2 I_{n-1}]^{-1} x_k, \end{aligned}$$

where the last equality follows with

$$\begin{aligned} [(X_n^{(k)} - E I_{n-1})^2 + v^2 I_{n-1}]^{-1} &= [U^* [(D - E I_{n-1})^2 + \eta^2 I_{n-1}] U]^{-1} \\ &= U^* [(D - E I_{n-1})^2 + \eta^2 I_{n-1}]^{-1} U. \end{aligned}$$

With the exact arguments we just used, we further obtain for the denominator in (5.2) that

$$\begin{aligned} x_k^* (X_n^{(k)} - z)^{-1} x_k &= \sum_{\ell=1}^{n-1} \frac{|y_\ell|^2}{\lambda_\ell - z} \\ &= \sum_{\ell=1}^{n-1} \frac{|y_\ell|^2}{(\lambda_\ell - E)^2 + \eta^2} (\lambda_\ell - E) + i \sum_{\ell=1}^{n-1} \frac{|y_\ell|^2}{(\lambda_\ell - E)^2 + \eta^2} \eta, \end{aligned}$$

so that

$$-\eta - \operatorname{Im}(x_k^*(X_n^{(k)} - z)^{-1}x_k) = -\eta(1 + x_k^*[(X_n^{(k)} - EI_{n-1})^2 + v^2 I_{n-1}]^{-1}x_k). \square$$

Note that Theorem 5.16 *i)* also allows us to work with the Stieltjes transform $S_{\mathbb{E}\sigma_n}$ of the expected ESD $\mathbb{E}\sigma_n$, since as in the proof of Theorem 5.8 we have $S_{\mathbb{E}\sigma_n} = \mathbb{E}S_{\sigma_n} = \mathbb{E}s_n$.

The remainder of this section will be devoted to the proof of Theorem 5.16, for which we follow the roadmap as in [5]. In the following Lemma, the Schur complement is defined and studied (see also [62]).

Lemma 5.17. *Let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

be a quadratic block matrix with A_{11} invertible. Then the Schur complement of A_{11} in A is defined as

$$B \coloneqq A_{22} - A_{21}A_{11}^{-1}A_{12}$$

and has the following properties, where I resp. 0 are identity matrices resp. 0-matrices of appropriate dimension:

i) We obtain the Schur complement formula

$$\begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & B \end{pmatrix}.$$

ii) We find the Schur complement determinant formula

$$\det(A) = \det(A_{11})\det(B) = \det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12})$$

iii) If A is invertible, so is $B = A_{22} - A_{21}A_{11}^{-1}A_{12}$.

iv) In case A is invertible, we find the Schur complement inversion formula

$$\begin{aligned} A^{-1} &= \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \\ &= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}B^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{pmatrix}. \end{aligned}$$

Proof. Statement *i)* requires mere verification by multiplication of the matrices, *ii)* follows directly from *i)* and *iii)* follows directly from *ii)*. The first equality of statement *iv)* follows directly by inverting the Schur complement formula and multiplying from the left and right with the appropriate matrices. The second equality is again verified through simple multiplication of the matrices. $\square$

Lemma 5.18. *Let A be an invertible $n \times n$ matrix. If $A^{(k)}$ is invertible for some $k \in \{1, \dots, n\}$, then*

$$A^{-1}(k,k) = \frac{1}{A(k,k) - r_k A^{(k)-1} c_k},$$

where r_k is the k-th row of A without the k-th entry and c_k is the k-th column of A without the k-th entry.

Proof. We first prove the statement for $k = n$. We write

$$A = \begin{pmatrix} A^{(n)} & c_n \\ r_n & A(n,n) \end{pmatrix}$$

and set $B \coloneqq A(n,n) - r_n A^{(n)-1} c_n$. Then by the Schur complement inversion formula, $A^{-1}(n,n) = B^{-1}$, which shows the statement for $k = n$. Next, we assume $k < n$. Then define a permutation matrix column-wise as

$$V \coloneqq (e_1|e_2|\dots|\widehat{e_k}|\dots|e_n|e_k)$$

where the e_i are the standard n-dimensional basis vectors, and the hat over e_k indicates that this vector is left out. In other words, V is obtained through the identity matrix by erasing its k-th column e_k and appending it at the end of the matrix. We obtain immediately that $V^T = V = V^{-1}$. Then AV is the matrix A with erased and then appended k-th column and VA is the matrix A with erased and then appended k-th row. Therefore, $(VAV)^{(n)} = A^{(k)}$ and by the case above

$$A^{-1}(k,k) = (VA^{-1}V)(n,n) = (VAV)^{-1}(n,n) = \frac{1}{(VAV)(n,n) - r'_n (VAV)^{(n)-1} c'_n},$$

where r'_n denotes the n-th row of VAV and c'_n denotes the n-th column of VAV, both without their n-th entry. But $r'_n = r_k$, $c'_n = c_k$ and $(VAV)(n,n) = A(k,k)$. □

Lemma 5.19. *Let A be an invertible $n \times n$ matrix and $k \in \{1,\dots,n\}$, such that $A^{(k)}$ is invertible. Then we obtain:*

$$\operatorname{tr} A^{-1} - \operatorname{tr} A^{(k)-1} = \frac{1 + r_k A^{(k)-2} c_k}{A(k,k) - r_k A^{(k)-1} c_k},$$

where r_k denotes the k-th row of A without the k-th entry and c_k denotes the k-th column of A without the k-th entry.

Proof. We first prove the statement for $k = n$. The Schur complement inversion formula for

$$A = \begin{pmatrix} A^{(n)} & c_n \\ r_n & A(n,n) \end{pmatrix}$$

yields with $B \coloneqq A(n,n) - r_n A^{(n)-1} c_n \in \mathbb{C}$, that

$$A^{-1} = \begin{pmatrix} A^{(n)-1} + A^{(n)-1} c_n B^{-1} r_n A^{(n)-1} & -A^{(n)-1} c_n B^{-1} \\ -B^{-1} r_n A^{(n)-1} & B^{-1} \end{pmatrix}.$$

Therefore, since the trace is linear and only depends on the diagonal block matrices, we find

$$\operatorname{tr} A^{-1} - \operatorname{tr} A^{(n)-1} \;=\; \operatorname{tr} \begin{pmatrix} A^{(n)-1} c_n B^{-1} r_n A^{(n)-1} & 0 \\ 0 & B^{-1} \end{pmatrix}$$

$$= \frac{1}{B} \operatorname{tr} \begin{pmatrix} A^{(n)-1} c_n r_n A^{(n)-1} & 0 \\ 0 & 1 \end{pmatrix} = \frac{1}{B} \left(1 + r_n A^{(n)-2} c_n \right),$$

where we used the cyclic property of the trace in the last step. This concludes the statement for $k = n$. Now if $k < n$, let V be the permutation matrix as in the proof of Lemma 5.18, then since $A^{(k)} = (VAV)^{(n)}$, we obtain with first part that

$$\begin{aligned} \operatorname{tr} A^{-1} - \operatorname{tr} A^{(k)-1} &= \operatorname{tr} VA^{-1}V - \operatorname{tr}(VAV)^{(n)-1} \\ &= \operatorname{tr}(VAV)^{-1} - \operatorname{tr}(VAV)^{(n)-1} \\ &= \frac{1 + r'_n (VAV)^{(n)-2} c'_n}{(VAV)(n,n) - r'_n (VAV)^{(n)-2} c'_n} \end{aligned}$$

where r'_n (resp. c'_n) is the n-th row (resp. column) of VAV without the n-th entry. This concludes the statement, since $r'_n = r_k$, $c'_n = c_k$, and $(VAV)(n,n) = A(k,k)$. □

6. The semicircle and MP laws by the Stieltjes transform method

6.1. General strategy and quadradic form estimates

In this very short section we introduce a general strategy behind the proofs of limit laws in random matrix theory utilizing Stieltjes transforms. We also introduce some versatile quadratic form estimates which allow us to carry out smooth proofs of the semicircle law and Marchenko-Pastur law in the following sections. Assume that $(\sigma_n)_n$ is a sequence of ESDs of Wigner matrices and $(\mu_n)_n$ is a sequence of ESDs of MP matrices. We would like to argue that $\sigma_n \to \sigma$ or $\mu_n \to \mu^y$ weakly for some $y > 0$, and in some stochastic sense, for example in probability or almost surely. To this end, we carry out the following three steps, where notationally, either $\rho_n = \sigma_n$ and $\rho = \sigma$, or $\rho_n = \mu_n$ and $\rho = \mu^y$:

1. We show that the Stieltjes transform S_ρ of the limit measure ρ satisfies a self-consistent quadratic equation and that the solutions can be separated so that if some S_ν solves the equation for some probability measure ν on $\mathbb{R}$ (if $\rho = \sigma$) or on $\mathbb{R}_+$ (if $\rho = \mu^y$), then necessarily $S_\nu = S_\rho$.
2. Applying the Schur complement formula, the Stieltjes transforms of the ESDs ρ_n can be written as a sum of inverses of complex numbers. We decompose each summand into a part pertaining to the self-consistent equation (the wanted part w) and an error term (the remainder r), using

$$\frac{1}{w + r} = \frac{1}{w} - \frac{r}{w(w + r)}. \tag{6.1}$$

   We establish that if the error term converges to zero in probability resp. almost surely, the limit law holds in probability resp. almost surely.

3. We establish that the error term converges to zero almost surely by employing quadratic form estimates in combination with an estimate on the difference of Stieltjes transforms of the ESD of a random matrix and its minors. Quadratic form estimates are elementary yet very powerful. They belong to the main ingredients to prove some of the most fruitful results in contemporary random matrix theory – namely the so-called local laws, see [8] or [22].

For the third step, we use quadratic form estimates which can later be applied to Wigner and Marchenko-Pastur matrices. In the following, for $p \geq 1$ the norm $\|\cdot\|_p$ shall denote the $\mathcal{L}_p(\mathbb{P})$-seminorm, so for any random variable $Y : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow \mathbb{C}$, $\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}$.

Theorem 6.1 (Marcinkiewicz-Zygmund Inequality)**.** *If $Y_1, \ldots, Y_n$ are independent, centered and complex-valued random variables with existing absolute moments, then for every $p \geq 1$ there exists a positive constant A_p which depends only on p, such that*

$$\left\| \sum_{i=1}^{n} Y_i \right\|_p \leq A_p \left\| \left(\sum_{i=1}^{n} |Y_i|^2 \right)^{\frac{1}{2}} \right\|_p$$

Proof. In [16, p. 386], the statement is proved for independent real-valued random variables. The statement is extended to the complex valued case in [5, p. 33]. □

We now formulate an important lemma, which is mainly based on Theorem 6.1.

Lemma 6.2. *Let $Y_1, \ldots, Y_n$ be independent, centered and complex-valued random variables which are uniformly $\|\cdot\|_p$-bounded for all $p \geq 2$. Then it holds for any complex numbers $(b_i)_{i\in[n]}$ and $(a_{i,j})_{i,j\in[n]}$*

$$i) \quad \forall\, p \geq 2 : \left\| \sum_{i=1}^{n} b_i Y_i \right\|_p \leq A_p \left(\sum_{i=1}^{n} |b_i|^2 \right)^{\frac{1}{2}},$$

$$ii) \quad \forall\, p \geq 2 : \left\| \sum_{i\neq j=1}^{n} a_{i,j} Y_i Y_j \right\|_p \leq A_p \left(\sum_{i\neq j=1}^{n} |a_{i,j}|^2 \right)^{\frac{1}{2}},$$

where A_p is a constant depending only on p and the uniform $\|\cdot\|_p$-bound.

Proof. The proofs of the two statements can be found in [8]. □

The following theorem establishes deviation bounds for the expressions in Lemma 6.2, when the constants b_i and $a_{i,j}$ are replaced by functions of random variables.

Theorem 6.3. *Let for all $n \in \mathbb{N}$, Y and W be n-dependent objects ($Y = Y^{(n)}, W = W^{(n)}$) that satisfy the following for all $n \in \mathbb{N}$:*

- $W = W^{(n)}$ *is a finite index set.*
- $Y_W = (Y_i)_{i\in W} = (Y_i^{(n)})_{i\in W^{(n)}} = Y^{(n)}_{W^{(n)}}$ *is a family of independent, real-valued and centered random variables, so that for all* $p \geq 2$*, the family* $(Y_i^{(n)} : i \in W^{(n)}, n \in \mathbb{N})$ *is uniformly* $\mathcal{L}_p$*-bounded.*

Further, denote for all subsets $K \subseteq W$ *by* $\mathcal{F}_W(\mathbb{R}^K)$ *the set of tuples* $C = (C_i)_{i\in W}$*, where for each* $i \in W$*,* $C_i : \mathbb{R}^K \to \mathbb{C}$ *is a complex-valued measurable function. Analogously, define for all subsets* $K \subseteq W$ *by* $\mathcal{F}_{W\times W}(\mathbb{R}^K)$ *the set of tuples* $C = (C_{i,j})_{i,j\in W}$*, where for all* $i, j \in W$*,* $C_{i,j} : \mathbb{R}^K \to \mathbb{C}$ *is a complex-valued measurable function. Then we obtain the following probability bounds:*

i) For all $\varepsilon, D > 0$ *there is a constant* $C_{\varepsilon,D} \geq 0$*, such that for all* $n \in \mathbb{N}$*, all disjoint subsets* $I, K \subseteq W$ *and all function tuples* $B \in \mathcal{F}_W(\mathbb{R}^K)$ *the following holds:*

$$\mathbb{P}\left(\left|\sum_{i\in I} B_i[Y_K]Y_i\right| \;>\; n^{\varepsilon} \cdot \sqrt{\sum_{i\in I}|B_i[Y_K]|^2}\right) \leq \frac{C_{\varepsilon,D}}{n^D}.$$

ii) For all $\varepsilon, D > 0$ *there is a constant* $C_{\varepsilon,D} \geq 0$*, such that for all* $n \in \mathbb{N}$*, all disjoint subsets* $I, K \subseteq W$*, and all function tuples* $A \in \mathcal{F}_{W\times W}(\mathbb{R}^K)$ *the following holds:*

$$\mathbb{P}\left(\left|\sum_{i,j\in I, i\neq j} Y_i A_{i,j}[Y_K]Y_j\right| \;>\; n^{\varepsilon} \cdot \sqrt{\sum_{i,j\in I, i\neq j}|A_{i,j}[Y_K]|^2}\right) \leq \frac{C_{\varepsilon,D}}{n^D}.$$

Proof. We only prove *ii)*, since *i)* can be proved analogously. Let $\varepsilon, D > 0$ be arbitrary and choose $p \in \mathbb{N}$ with $p \geq 2$ so large that $p\varepsilon > D$. Then we pick an $n \in \mathbb{N}$, disjoint subsets $I, K \subseteq W^{(n)}$ and a function tuple $A \in \mathcal{F}_{W\times W}(\mathbb{R}^K)$. To avoid division by zero, we define the set:

$$\mathcal{A}_2 \coloneqq \left\{ y_K \in \mathbb{R}^K \,|\, \sum_{i,j\in I, i\neq j} |A_{i,j}[y_K]|^2 > 0 \right\}.$$

Then we conduct the following calculation (explanations are found below the calculation; the sums over "$i \neq j$" are over all $i, j \in I$ with $i \neq j$):

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i\neq j} Y_i A_{i,j}[Y_K]Y_j\right| > n^{\varepsilon}\left(\sum_{i\neq j}|A_{i,j}[Y_K]|^2\right)^{\frac{1}{2}}\right) \\
&= \mathbb{P}\left(\left|\frac{\sum_{i\neq j} Y_i A_{i,j}[Y_K]Y_j}{\left(\sum_{i\neq j}|A_{i,j}[Y_K]|^2\right)^{\frac{1}{2}}}\right|^p \mathbb{1}_{\mathcal{A}_2}(Y_K) > n^{p\varepsilon}\right)
\end{aligned}$$

$$\leq \frac{1}{n^{p\varepsilon}}\mathbb{E}\left|\frac{\sum_{i\neq j} Y_i A_{i,j}[Y_K]Y_j}{\left(\sum_{i\neq j}|A_{i,j}[Y_K]|^2\right)^{\frac{1}{2}}}\right|^p \mathbb{1}_{\mathcal{A}_2}(Y_K)$$

$$= \frac{1}{n^{p\varepsilon}}\int_{\mathbb{R}^K}\int_{\mathbb{R}^I}\left|\frac{\sum_{i\neq j} y_i A_{i,j}[y_K]y_j}{\left(\sum_{i\neq j}|A_{i,j}[y_K]|^2\right)^{\frac{1}{2}}}\right|^p \mathrm{d}\mathbb{P}^{Y_I}(y_I)\mathbb{1}_{\mathcal{A}_2}(y_K)\mathrm{d}\mathbb{P}^{Y_K}(y_K)$$

$$\leq \frac{(A_p)^p}{n^{p\varepsilon}} \leq \frac{(A_p)^p}{n^D}$$

where the first step follows from the fact that for

$$\left|\sum_{i\neq j} Y_i A_{i,j}[Y_K]Y_j\right| > n^{\varepsilon}\left(\sum_{i\neq j}|A_{i,j}[Y_K]|^2\right)^{\frac{1}{2}}$$

to hold, not all $A_{i,j}[Y_K]$ may vanish, in the second step we used Markov's inequality, in the third step we used Fubini, in the fourth step we applied Lemma 6.2 and in the last step we used the choice of p in the beginning of the proof. Note that $(A_p)^p$ denotes a constant which depends only on p, which in turn depends only on the choices of ε and D. In particular, this constant does not depend on the choice of $n \in \mathbb{N}$, the sets I and K or the function tuple A. This shows *ii*). □

6.2. The semicircle law

We follow the general strategy outlined in Section 6.1.

Step 1: Self-consistent equation and separation of solutions

The first step of the proof consists of the following lemma:

Lemma 6.4. *The Stieltjes transform of the semicircle distribution σ is given by*

$$\forall\, z \in \mathbb{C}_+ : S_\sigma(z) = \frac{-z+\sqrt{z^2-4}}{2},$$

where we use the convention that $\sqrt{\cdot}$ denotes the complex square root with non-negative imaginary part. Consider the following equation in $m \in \mathbb{C}$, where $z \in \mathbb{C}_+$ is fixed:

$$m = \frac{1}{-z-m} \tag{6.2}$$

Then the following statements hold:

i) The solutions to (6.2) are given by

$$m_{+,-} = \frac{-z\pm\sqrt{z^2-4}}{2}.$$

ii) $S_\sigma(z)$ *is the positive branch of the solution in* (6.2)*, that is,* $S_\sigma(z) = m_+$.
iii) For the denominator in (6.2) *it holds* $\operatorname{Im}(-z - m_-) \geq -\frac{1}{2}\operatorname{Im}(z)$
iv) If $\nu \in \mathcal{M}_1(\mathbb{R})$*, then for all* $z \in \mathbb{C}_+$ *it holds*

$$\operatorname{Im}(-z - S_\nu(z)) \leq -\operatorname{Im}(z).$$

In particular, if $S_\nu(z)$ *satisfies* (6.2)*, we must have* $S_\nu(z) = S_\sigma(z)$.

Proof. The Stieltjes transform of the semicircle distribution is derived in Lemma 2.11 in [5]. Statements *i)* and *ii)* can be shown directly by solving the quadratic equation (6.2). Statement *iii)* follows since

$$\operatorname{Im}\left(-z - \frac{-z - \sqrt{z^2-4}}{2}\right) = -\operatorname{Im}(z) + \frac{\operatorname{Im}(z)}{2} + \frac{\operatorname{Im}\sqrt{z^2-4}}{2} \geq -\frac{\operatorname{Im}(z)}{2},$$

since we defined the complex square root $\sqrt{\cdot}$ to be the square root with non-negative imaginary part. Statement *iv)* follows trivially since $\operatorname{Im} S_\nu(z) \geq 0$. □

Step 2: Derivation of the error term

By Theorem 5.16 *i)*, the Stieltjes transform s_n of a Wigner matrix $\frac{1}{\sqrt{n}}X_n$ is given by

$$s_n(z) = \frac{1}{n}\sum_{k\in[n]} \frac{1}{\frac{1}{\sqrt{n}}X_n(k,k) - z - \frac{1}{n}x_k^T\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}x_k}, \tag{6.3}$$

where $X_n^{(k)}$ denotes the k-th principle minor of X_n and x_k the k-th column of X_n without the k-th entry. The desired denominator in each summand of (6.3) is

$$-z - s_n(z),$$

stemming from the self-consistent equation (6.2). In the k-th summand for $k \in \{1,\dots,n\}$, we obtain the remainder term

$$\Omega_n^{(k)}(z) := \frac{1}{\sqrt{n}}X_n(k,k) + s_n(z) - \frac{1}{n}x_k^T\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}x_k. \tag{6.4}$$

Using (6.1), we conclude

$$s_n(z) = \frac{1}{-z - s_n(z)} - \delta_n(z) \tag{6.5}$$

with

$$\delta_n(z) = \frac{1}{n}\sum_{k\in[n]} \frac{\Omega_n^{(k)}(z)}{(-z - s_n(z))(-z - s_n(z) + \Omega_n^{(k)}(z))}.$$

The error term $\delta_n(z)$ can be bounded in absolute terms as follows: By Lemma 6.4, we find

$$\forall\, n \in \mathbb{N} : \operatorname{Im}(-z - s_n(z)) \leq -\operatorname{Im}(z).$$

If we assume

$$\max_k |\Omega_n^{(k)}(z)| \leq \frac{1}{2} \operatorname{Im}(z) \tag{6.6}$$

we may therefore conclude

$$\begin{aligned} |\delta_n(z)| &= \left| \frac{1}{n} \sum_{k\in[n]} \frac{\Omega_n^{(k)}(z)}{(-z - s_n(z))(-z - s_n(z) + \Omega_n^{(k)}(z))} \right| \\ &\leq \frac{1}{n} \sum_{k\in[n]} \frac{|\Omega_n^{(k)}|}{\operatorname{Im}(z)^2/2} \;\leq\; \frac{2}{\operatorname{Im}(z)^2} \max_k |\Omega_n^{(k)}|. \end{aligned} \tag{6.7}$$

The following lemma puts our findings into perspective:

Theorem 6.5. *In above situation, the following statements hold for any fixed $z \in \mathbb{C}_+$:*

i) If in (6.5), $\delta_n(z) \to 0$ in probability resp. almost surely, then $s_n(z) \to s(z)$ in probability resp. almost surely.

ii) If $\max_{k\in[n]} |\Omega_n^{(k)}(z)| \xrightarrow[n\to\infty]{} 0$ in probability resp. almost surely, then $\delta_n(z) \to 0$ in probability resp. almost surely.

Proof. Statement *ii)* follows with (6.7), using (6.6). We proceed to show statement *i)* in the almost sure sense. Fix $z \in \mathbb{C}_+$. Let A be a measurable set with $\mathbb{P}(A) = 1$, on which $\delta_n(z) \to 0$. Let $\omega \in A$ be arbitrary, and denote by $s_n^\omega(z)$ the realization of $s_n(z)$ at ω. To show that $s_n^\omega(z)$ converges to $s(z)$, we show that any subsequence of $s_n^\omega(z)$ contains another subsequence that converges to $s(z)$. To this end, let $J \subseteq \mathbb{N}$ be a subsequence. Then $(s_n^\omega(z))_{n\in J}$ is a bounded sequence of complex numbers (with absolute bound $\operatorname{Im}(z)^{-1} > 0$), therefore has a convergent subsequence $(s_n^\omega(z))_{n\in I}$, $I \subseteq J$, with some limit $t \in \mathbb{C}$ (Bolzano-Weierstrass). Considering (6.5), t satisfies

$$t = \frac{1}{-z - t}.$$

Since $\operatorname{Im}(-z - s_n^\omega(z)) \leq -\operatorname{Im}(z)$ by Lemma 6.4, we find $\operatorname{Im}(-z - t) \leq -\operatorname{Im}(z)$, so $t = s(z)$ by Lemma 6.4. We have seen that any subsequence of $(s_n^\omega(z))_{n\in\mathbb{N}}$ has a subsequence which converges to $s(z)$. Therefore, $s_n(z) \to s(z)$ on A, that is, almost surely. Statement *i)* in probability follows from the almost sure version we just proved, using Lemma 2.27: To show that $s_n(z) \to s(z)$ in probability, it suffices to show that for any subsequence $I \subseteq \mathbb{N}$ there is a subsequence $J \subseteq I$ such that $s_n(z) \to s(z)$ for $n \in J$. So let $I \subseteq \mathbb{N}$ be an arbitrary subsequence. Since $\delta_n(z) \to 0$ in probability, there is a subsequence $J \subseteq I$ with $\delta_n(z) \to 0$ almost surely for $n \in J$. But then $s_n(z) \to s(z)$ for $n \in J$ almost surely as we just proved above. This completes the argument. □

Step 3: Analysis of the error term

By Theorem 6.5, it suffices to show that $\max_{k\in[n]} |\Omega_n^{(k)}(z)| \xrightarrow[n\to\infty]{} 0$ in probability or almost surely, where $\Omega_n^{(k)}(z)$ was defined in (6.4). In this subsection, we will show almost sure convergence. Note that

$$\begin{aligned}\Omega_n^{(k)}(z) = &\frac{1}{\sqrt{n}} X_n(k,k) \\ &- \frac{1}{n}\sum_{i\neq j}^{n} x_k(i)\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}(i,j)x_k(j) \\ &- \frac{1}{n}\sum_{i=1}^{n}(x_k(i)^2-1)\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}(i,i) \\ &- \frac{1}{n}\operatorname{tr}\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1} + s_n(z) \\ =: &\, A(n,k) + B(n,k,z) + C(n,k,z) + D(n,k,z). \end{aligned} \tag{6.8}$$

We will analyze these four terms separately and show that their maxima over $k \in \{1,\dots,n\}$ converge to zero almost surely. For B and C we will use Theorem 6.3, whereas for D we will use Theorem 5.16 *ii*).

Lemma 6.6. *In* (6.8), $\max_{k\in[n]} |A(n,k)| \to 0$ *almost surely as* $n \to \infty$.

Proof. Let C_8 be an upper bound of $\mathbb{E}(X_n(i,j))^8$ for all n, i, j, then we find

$$\forall\, n \in \mathbb{N} : \forall\, k \in [n] :\ \mathbb{P}\left(\left|\frac{1}{\sqrt{n}}X_n(k,k)\right| > \frac{1}{n^{\frac{1}{8}}}\right) = \mathbb{P}\left(|X_n(k,k)|^8 > \frac{n^4}{n}\right) \leq \frac{C_8}{n^3}.$$

Therefore, taking the union bound, we obtain for all $n \in \mathbb{N}$:

$$\mathbb{P}\left(\max_{k\in[n]}\left|\frac{1}{\sqrt{n}}X_n(k,k)\right| > \frac{1}{n^{\frac{1}{8}}}\right) \leq \sum_{k\in[n]} \mathbb{P}\left(\left|\frac{1}{\sqrt{n}}X_n(k,k)\right| > \frac{1}{n^{\frac{1}{8}}}\right) \leq \frac{nC_8}{n^3}$$

which converges to zero summably fast. This concludes the proof by Borel-Cantelli. □

For $B(n,k,z)$ we define the terms

$$S(n,k,z) \coloneqq \sum_{i\neq j}^{n} x_k(i)\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}(i,j)x_k(j)$$

and

$$R(n,k,z) \coloneqq \sqrt{\sum_{i\neq j}^{n}\left|\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}(i,j)\right|^2}$$

to employ Theorem 6.3 *ii*). To bound $R(n,k,z)$, we use the following trivial lemma:

Lemma 6.7. *Let X be an $n \times n$ random matrix and $z \in \mathbb{C}_+$. Then*

$$\sqrt{\sum_{i,j\in[n]} |(X-z)^{-1}(i,j)|^2} \leq \frac{\sqrt{n}}{\operatorname{Im}(z)}.$$

Proof. For $n \times n$ matrices X, the general inequality $\|X\|_F \leq \sqrt{n}\|X\|_{op}$ holds, where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_{op}$ is the operator norm. On the other hand, for any *self-adjoint* $n \times n$ matrix X, $\|(X-z)^{-1}\|_{op} \leq \operatorname{Im}(z)^{-1}$. Combining these facts, we obtain

$$\sqrt{\sum_{i,j\in[n]} |(X-z)^{-1}(i,j)|^2} = \|(X-z)^{-1}\|_F \leq \sqrt{n}\|(X-z)^{-1}\|_{op} \leq \frac{\sqrt{n}}{\operatorname{Im}(z)}. \qquad \square$$

Lemma 6.8. *In (6.8), $\max_{k\in[n]} |B(n,k,z)| \to 0$ almost surely as $n \to \infty$.*

Proof. Using the terms $S(n,k,z)$ and $R(n,k,z)$ defined above, we find by Lemma 6.7 that

$$|R(n,k,z)| \leq \frac{\sqrt{n}}{\operatorname{Im}(z)},$$

Therefore, using Theorem 6.3 *ii)* with $\varepsilon = 1/4$ and $D = 3$, we obtain a constant $C_{\frac{1}{4},3} \geq 0$, such that for all $n \in \mathbb{N}$ and all $k \in [n]$:

$$\mathbb{P}\left(|B(n,k,z)| > \frac{n^{\frac{1}{4}}\sqrt{n}}{n\operatorname{Im}(z)}\right) \leq \mathbb{P}\left(|S(n,k,z)| > n^{\frac{1}{4}}R(n,k,z)\right) \leq \frac{C_{\frac{1}{4},3}}{n^3}.$$

Applying the union bound as in the proof of Lemma 6.6 concludes the statement. $\square$

For $C(n,k,z)$, we define the terms

$$S'(n,k,z) \coloneqq \sum_{i\in[n]} (x_k(i)^2 - 1)\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}(i,i)$$

and

$$R'(n,k,z) \coloneqq \sqrt{\sum_{i\in[n]} \left|\left(\frac{1}{\sqrt{n}}X_n^{(k)} - z\right)^{-1}(i,i)\right|^2}$$

to employ Theorem 6.3 *i)*.

Lemma 6.9. *In (6.8), $\max_{k\in[n]} |C(n,k,z)| \to 0$ almost surely as $n \to \infty$.*

Proof. Using the terms $S'(n,k,z)$ and $R'(n,k,z)$ defined above, we find by Lemma 6.7 that

$$|R'(n,k,z)| \leq \frac{\sqrt{n}}{\operatorname{Im}(z)},$$

Therefore, using Theorem 6.3 *i)* with $\varepsilon = 1/4$ and $D = 3$, we obtain a constant $C_{\frac{1}{4},3} \geq 0$, such that for all $n \in \mathbb{N}$ and all $k \in [n]$:

$$\mathbb{P}\left(|C(n,k,z)| > \frac{n^{\frac{1}{4}}\sqrt{n}}{n}\right) \leq \mathbb{P}\left(|S'(n,k,z)| > n^{\frac{1}{4}} R'(n,k,z)\right) \leq \frac{C_{\frac{1}{4},3}}{n^3}.$$

Applying the union bound as in the proof of Lemma 6.6 concludes the statement. □

Lemma 6.10. *In* (6.8), $\max_{k\in[n]} |D(n,k,z)| \to 0$ *almost surely as* $n \to \infty$.

Proof. With Theorem 5.16 *ii)*, we obtain for any $n \in \mathbb{N}$ and $k \in [n]$ that $|D(n,k,z)| \leq (n \operatorname{Im}(z))^{-1}$, so that

$$\max_{k\in[n]} |D(n,k,z)| \leq \frac{1}{n \operatorname{Im}(z)} \xrightarrow[n\to\infty]{} 0 \quad \text{almost surely.}$$ □

Theorem 6.11. *In above situation, we find for any fixed* $z \in \mathbb{C}_+$ *that*

$$\max_{k\in[n]} |\Omega_n^{(k)}(z)| \xrightarrow[n\to\infty]{} 0 \quad \textit{almost surely.}$$

Proof. This follows directly by the decomposition (6.8) with Lemma 6.6, Lemma 6.8, Lemma 6.9 and Lemma 6.10. □

6.3. The Marchenko-Pastur law

Again, we follow the general strategy outlined in Section 6.1.

Step 1: Self-consistent equation and separation of solutions

The first step of the proof consists of the following lemma:

Lemma 6.12. *Fix* $y > 0$. *The Stieltjes transform of the Marchenko-Pastur distribution* μ^y *is given by*

$$\forall\ z \in \mathbb{C}_+ :\ S_{\mu^y}(z) = \frac{1 - y - z + \sqrt{(z-1-y)^2 - 4y}}{2yz}.$$

Consider the equation in $m \in \mathbb{C}$, *where* $z \in \mathbb{C}_+$ *is fixed:*

$$m = \frac{1}{1 - z - y - yzm} \tag{6.9}$$

Then the following statements hold:

i) The solutions to (6.9) *are given by*

$$m_{+,-} = \frac{1 - y - z \pm \sqrt{(1-y-z)^2 - 4yz}}{2yz}.$$

ii) $S_{\mu^y}(z)$ *is the positive branch of the solutions to* (6.9)*, that is,* $S_{\mu^y}(z) = m_+$.
iii) For the denominator in (6.9) *it holds* $\operatorname{Im}(1 - z - y - yzm_-) \geq -\frac{1}{2}\operatorname{Im}(z)$.
iv) If $\nu \in \mathcal{M}_1([0,\infty))$*, then for all* $z \in \mathbb{C}_+$ *we find*

$$\operatorname{Im}(1 - z - y - yzS_\nu(z)) \leq -\operatorname{Im}(z)$$

In particular, if $S_\nu(z)$ *satisfies* (6.9)*, we must have* $S_\nu(z) = S_{\mu^y}(z)$.

Proof. The Stieltjes transform S_{μ^y} is derived in Lemma 3.11 in [5]. Statement *i)* is verified by solving the quadratic equation (6.9). For *ii)*, we calculate $(1 - y - z)^2 - 4yz = z^2 - 2yz + y^2 - 2y - 2z + 1 = (z - y - 1)^2 - 4y$. For *iii)*, we calculate

$$\begin{aligned}
&\operatorname{Im}(1 - z - y - yzm_-) \\
&= \operatorname{Im}\left(1 - z - y - yz\frac{1 - y - z - \sqrt{(1 - y - z)^2 - 4yz}}{2yz}\right) \\
&= \operatorname{Im}\left(\frac{1 - y - z + \sqrt{(1 - y - z)^2 - 4yz}}{2}\right) \\
&= \frac{1}{2}\left(-\operatorname{Im}(z) + \underbrace{\operatorname{Im}\sqrt{(1 - y - z)^2 - 4yz}}_{\geq 0 \text{ per definition of } \sqrt{\cdot}}\right) \geq -\frac{\operatorname{Im}(z)}{2}.
\end{aligned}$$

For *iv)*, note that with $z = E + i\eta$, where $E \in \mathbb{R}$ and $\eta > 0$, we find

$$\begin{aligned}
\operatorname{Im}(zS_\nu(z)) &= \operatorname{Re}(z)\operatorname{Im} S_\nu(z) + \operatorname{Im}(z)\operatorname{Re} S_\nu(z) \\
&= E\int \frac{\eta}{(x - E)^2 + \eta^2}\nu(\mathrm{d}x) + \eta\int \frac{x - E}{(x - E)^2 + \eta^2}\nu(\mathrm{d}x) \\
&= \eta\int \frac{x}{(x - E)^2 + \eta^2}\nu(\mathrm{d}x).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{Im}(1 - z - y - yzS_\nu(z)) &= -\eta - y\eta\int_{[0,\infty)} \frac{x}{(x - E)^2 + \eta^2}\nu(\mathrm{d}x) \\
&= -\eta\left(1 + y\int_{[0,\infty)} \frac{x}{(x - E)^2 + \eta^2}\nu(\mathrm{d}x)\right) \\
&\leq -\eta.
\end{aligned}$$

The very last statement follows with part *iii)*. □

Step 2: Derivation of the error term

By Theorem 5.16 *i*), the Stieltjes transform s_n of an MP matrix $\frac{1}{n}X_nX_n^T$ is given by

$$s_n(z) = \frac{1}{p}\sum_{k=1}^{p} \frac{1}{\frac{1}{n}\alpha_k^T\alpha_k - z - \frac{1}{n^2}\alpha_k^T X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T} - z\right)^{-1} X_n^{(k)}\alpha_k}, \quad (6.10)$$

where α_k^T is the k-th row of X_n (note that α_k also depends on n, which we drop from the notation), $X_n^{(k)}$ is X_n with k-th row removed (thus a $(p-1)\times n$-matrix). The desired denominator in each summand of (6.10) is

$$1 - z - y_n - y_n z s_n(z),$$

stemming from the self-consistent equation (6.9), where $y_n \coloneqq p/n$ and it is assumed that there exists a $y \in (0,\infty)$ such that $y_n \to y$. (It is favorable to work with y_n instead of y, since this leads to a cancellation within the error term $\Omega_n^{(k)}(z)$ we define below, see the proof of Lemma 6.18 below.) In the k-th summand for $k \in \{1,\dots,p\}$, we obtain the remainder term

$$\Omega_n^{(k)}(z) \coloneqq \frac{1}{n}\alpha_k^T\alpha_k - 1 - \frac{1}{n^2}\alpha_k^T X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T} - z\right)^{-1} X_n^{(k)}\alpha_k + y_n + y_n z s_n(z). \quad (6.11)$$

Using (6.1), we conclude

$$s_n(z) = \frac{1}{1 - z - y_n - y_n z s_n(z)} - \delta_n(z) \quad (6.12)$$

with

$$\delta_n(z) = \frac{1}{p}\sum_{k=1}^{p} \frac{\Omega_n^{(k)}(z)}{(1 - z - y_n - y_n z s_n(z))(1 - z - y_n - y_n z s_n(z) + \Omega_n^{(k)}(z))}.$$

The error term $\delta_n(z)$ can be bounded in absolute terms as follows: By Lemma 6.12, we find

$$\forall\, n \in \mathbb{N} : \operatorname{Im}(1 - z - y_n - y_n z s_n(z)) \leq -\operatorname{Im}(z).$$

If we assume

$$\max_{k\in[p]} |\Omega_n^{(k)}(z)| \leq \frac{1}{2}\operatorname{Im}(z) \quad (6.13)$$

we may therefore conclude

$$\begin{aligned} |\delta_n(z)| &= \left|\frac{1}{p}\sum_{k=1}^{p} \frac{\Omega_n^{(k)}(z)}{(1 - z - y_n - y_n z s_n(z))(1 - z - y_n - y_n z s_n(z) + \Omega_n^{(k)}(z))}\right| \\ &\leq \frac{1}{p}\sum_{k=1}^{p} \frac{|\Omega_n^{(k)}|}{\operatorname{Im}(z)^2/2} \;\leq\; \frac{2}{\operatorname{Im}(z)^2}\max_{k\in[p]}|\Omega_n^{(k)}|. \end{aligned} \quad (6.14)$$

The following lemma puts our findings into perspective:

Theorem 6.13. *In above situation, the following statements hold for any fixed* $z \in \mathbb{C}_+$*:*

i) If in (6.12)*,* $\delta_n(z) \to 0$ *in probability resp. almost surely, then* $s_n(z) \to s(z)$ *in probability resp. almost surely.*

ii) If $\max_{k\in[p]} |\Omega_n^{(k)}(z)| \xrightarrow[n\to\infty]{} 0$ *in probability resp. almost surely, then* $\delta_n(z) \to 0$ *in probability resp. almost surely.*

Proof. Statement *ii)* follows with (6.14), using (6.13). We proceed to show statement *i)* in the almost sure sense. Fix $z \in \mathbb{C}_+$. Let A be a measurable set with $\mathbb{P}(A) = 1$, on which $\delta_n(z) \to 0$. Let $\omega \in A$ be arbitrary, and denote by $s_n^\omega(z)$ the realization of $s_n(z)$ at ω. To show that $s_n^\omega(z)$ converges to $s(z)$, we show that any subsequence of $s_n^\omega(z)$ contains another subsequence that converges to $s(z)$. To this end, let $J \subseteq \mathbb{N}$ be a subsequence. Then $(s_n^\omega(z))_{n\in J}$ is a bounded sequence of complex numbers (with absolute bound $\operatorname{Im}(z)^{-1} > 0$), therefore has a convergent subsequence $(s_n^\omega(z))_{n\in I}$, $I \subseteq J$, with some limit $t \in \mathbb{C}$ (Bolzano-Weierstrass). Also, as $n \to \infty$ we find $y_n \to y$. Therefore, considering (6.12), t satisfies

$$t = \frac{1}{1 - z - y - yzt}.$$

For all realizations of the ESDs μ_n of $\frac{1}{n}X_nX_n^T$, $\mu_n([0,\infty)) = 1$, since the matrix has only non-negative spectrum. Therefore, by Lemma 6.12,

$$\forall\, n \in I : \operatorname{Im}(1 - z - y - yzs_n^\omega(z)) \le -\operatorname{Im}(z),$$

so also $\operatorname{Im}(1 - z - y - yzt) \le -\operatorname{Im}(z)$, hence $t = s(z)$ by Lemma 6.12. We have seen that any subsequence of $(s_n^\omega(z))_{n\in\mathbb{N}}$ has a subsequence which converges to $s(z)$. Therefore, $s_n(z) \to s(z)$ on A, that is, almost surely. The statement about convergence in probability can be proved verbatim as in the proof of Theorem 6.5. □

Step 3: Analysis of the error term

By Theorem 6.13, it suffices to show that $\max_{k\in[p]} |\Omega_n^{(k)}(z)| \xrightarrow[n\to\infty]{} 0$ in probability or almost surely, where $\Omega_n^{(k)}(z)$ was defined in (6.11). In this subsection, we show almost sure convergence. Note that

$$\begin{aligned}\Omega_n^{(k)}(z) = \frac{1}{n}\alpha_k^T\alpha_k - 1 \\ - \frac{1}{n^2}\sum_{i\neq j}^{n} \alpha_k(i)\left[X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T} - z\right)^{-1}X_n^{(k)}\right](i,j)\alpha_k(j) \\ - \frac{1}{n^2}\sum_{i=1}^{n}(\alpha_k(i)^2 - 1)\left[X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T} - z\right)^{-1}X_n^{(k)}\right](i,i) \\ - \frac{1}{n^2}\operatorname{tr}\left[X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T} - z\right)^{-1}X_n^{(k)}\right]\end{aligned}$$

$$
\begin{aligned}
&+ y_n + y_n z s_n(z) + y_n + y_n z s_n(z) \\
&=: A(n,k) + B(n,k,z) + C(n,k,z) + D(n,k,z).
\end{aligned} \tag{6.15}
$$

We will analyze these four terms separately and show that their maxima over $k \in \{1, \dots, p\}$ converge to zero almost surely. For A, B and C we will use Theorem 6.3, whereas for D we will use Theorem 5.16 *ii*).

Lemma 6.14. *In* (6.15), $\max_{k\in[p]} |A(n,k)| \to 0$ *almost surely as* $n \to \infty$.

Proof. We employ Theorem 6.3 *i*) with $B_i \equiv 1$, $\varepsilon = 1/4$ and $D = 3$ to obtain a constant $C_{\frac{1}{4},3} \geq 0$ such that

$$
\forall\, n \in \mathbb{N} : \forall\, k \in [p] :\ \mathbb{P}\left(\left| \frac{1}{n} \sum_{i\in[n]} (X_n(k,i)^2 - 1) \right| > \frac{n^{1/4}\sqrt{n}}{n} \right) \leq \frac{C_{\frac{1}{4},3}}{n^3}.
$$

Therefore, taking the union bound, we obtain for all $n \in \mathbb{N}$:

$$
\mathbb{P}\left(\max_{k\in[p]} \left| \frac{1}{n} \sum_{i\in[n]} (X_n(k,i)^2 - 1) \right| > \frac{n^{1/4}\sqrt{n}}{n} \right) \leq \frac{p C_{1/4,3}}{n^3}
$$

which converges to zero summably fast. This concludes the proof by Borel-Cantelli. □

For $B(n,k,z)$, we define the terms

$$
S(n,k,z) \coloneqq \sum_{i\neq j}^{n} \alpha_k(i) \left[X_n^{(k)T} \left(\frac{1}{n} X_n^{(k)} X_n^{(k)T} - z \right)^{-1} X_n^{(k)} \right] (i,j) \alpha_k(j)
$$

and

$$
R(n,k,z) \coloneqq \sqrt{ \sum_{i\neq j}^{n} \left| \left[X_n^{(k)T} \left(\frac{1}{n} X_n^{(k)} X_n^{(k)T} - z \right)^{-1} X_n^{(k)} \right] (i,j) \right|^2 }
$$

to employ Theorem 6.3 *ii*). To bound $R(n,k,z)$, we formulate the following lemma, which is taken from [27]:

Lemma 6.15. *Let* X *be a* $p \times n$ *matrix with real-valued entries,* $z \in \mathbb{C}_+$. *Define*

$$
F(X) \coloneqq X^T \left(\frac{1}{n} X X^T - z \right)^{-1} X. \tag{6.16}
$$

Then we obtain the following bound:

$$
\sqrt{ \sum_{i,j\in[n]} |F_{ij}(X)|^2 } \leq n\sqrt{p} \left(1 + \frac{|z|}{\operatorname{Im}(z)} \right)
$$

Proof. We recall that

a) $\text{Spectrum}(X^T(XX^T-z)^{-1}X)\cup\{0\} = \text{Spectrum}((XX^T-z)^{-1}XX^T)\cup\{0\}$,
b) $(XX^T-z)^{-1}XX^T = I + z(XX^T-z)^{-1}$,

and that $\|\cdot\|_F \leq \sqrt{m}\|\cdot\|_{op}$ for $m\times m$ matrices, where $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_{op}$ denotes the operator norm. Therefore,

$$\begin{aligned}
&\sqrt{\sum_{i,j\in[n]}|F_{ij}(X)|^2} = n\left\|\frac{1}{n}X^T\left(\frac{1}{n}XX^T-z\right)^{-1}X\right\|_F \\
&= n\left\|\left(\frac{1}{n}XX^T-z\right)^{-1}\left(\frac{1}{n}XX^T\right)\right\|_F = n\left\|I_p + z\left(\frac{1}{n}XX^T-z\right)^{-1}\right\|_F \\
&\leq n\sqrt{p}\left\|I_p + z\left(\frac{1}{n}XX^T-z\right)^{-1}\right\|_{op} \leq n\sqrt{p}\left(1+\frac{|z|}{\text{Im}(z)}\right). \qquad \square
\end{aligned}$$

Lemma 6.16. *In* (6.15), $\max_{k\in[p]}|B(n,k,z)|\to 0$ *almost surely as* $n\to\infty$.

Proof. Using the terms $S(n,k,z)$ and $R(n,k,z)$ defined above, we find by Lemma 6.15 that

$$|R(n,k,z)| \leq n\sqrt{p}c(z),$$

where $c(z) = 1+|z|/\text{Im}(z)$. Therefore, using Theorem 6.3 *ii*) with $\varepsilon = 1/8$ and $D=3$, we obtain a constant $C_{\frac{1}{8},3}\geq 0$, such that for all $n\in\mathbb{N}$ and all $k\in[p]$:

$$\mathbb{P}\left(|B(n,k,z)| > \frac{n^{\frac{1}{8}}n\sqrt{p}}{n^2}c(z)\right) \leq \mathbb{P}\left(|S(n,k,z)| > n^{\frac{1}{8}}R(n,k,z)\right) \leq \frac{C_{\frac{1}{8},3}}{n^3}.$$

Using the union bound as in the proof of Lemma 6.14 concludes the statement. $\square$

For $C(n,k,z)$, we define the terms

$$S'(n,k,z) \coloneqq \sum_{i\in[n]}(\alpha_k(i)^2-1)\left[X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T}-z\right)^{-1}X_n^{(k)}\right](i,i)$$

and

$$R'(n,k,z) \coloneqq \sqrt{\sum_{i\in[n]}\left|\left[X_n^{(k)T}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T}-z\right)^{-1}X_n^{(k)}\right](i,i)\right|^2}$$

to employ Theorem 6.3 *i*).

Lemma 6.17. *In* (6.15), $\max_{k\in[p]}|C(n,k,z)|\to 0$ *almost surely as* $n\to\infty$.

Proof. Using the terms $S'(n,k,z)$ and $R'(n,k,z)$ defined above, we find by Lemma 6.15 that

$$|R'(n,k,z)| \leq n\sqrt{p}c(z),$$

where $c(z) = 1 + |z|/\operatorname{Im}(z)$. Therefore, using Theorem 6.3 *i)* with $\varepsilon = 1/8$ and $D = 3$, we obtain a constant $C_{\frac{1}{8},3} \geq 0$, such that for all $n \in \mathbb{N}$ and all $k \in [p]$:

$$\mathbb{P}\left(|C(n,k,z)| > \frac{n^{\frac{1}{8}}\sqrt{n}}{n^2}c(z)\right) \leq \mathbb{P}\left(|S'(n,k,z)| > n^{\frac{1}{8}}R'(n,k,z)\right) \leq \frac{C_{\frac{1}{8},3}}{n^3}.$$

Using the union bound as in the proof of Lemma 6.14 concludes the statement. □

Lemma 6.18. *In* (6.15), $\max_{k\in[p]} |D(n,k,z)| \to 0$ *almost surely as* $n \to \infty$.

Proof. With the observations a) and b) in the proof of Lemma 6.15 and setting $X \coloneqq X_n^{(k)}$, we calculate

$$\begin{aligned}
&-\frac{1}{n^2}\operatorname{tr}\left[X^T\left(\frac{1}{n}XX^T - z\right)^{-1}X\right] = -\frac{1}{n^2}\operatorname{tr}\left[\left(\frac{1}{n}XX^T - z\right)^{-1}XX^T\right] \\
&= -\frac{1}{n}\operatorname{tr}\left[I_{p-1} + z\left(\frac{1}{n}XX^T - z\right)^{-1}\right] = -\frac{p}{n} + \frac{1}{n} - \frac{z}{n}\operatorname{tr}\left(\frac{1}{n}XX^T - z\right)^{-1}
\end{aligned}$$

Hence, using that $y_n = p/n$ and with Theorem 5.16 *ii)* (Note that our construction of $X_n^{(k)}$ differs from that in the theorem), we obtain

$$\begin{aligned}
|D(n,k,z)| = \Bigg| &-\frac{p}{n} + \frac{1}{n} - \frac{z}{n}\operatorname{tr}\left(\frac{1}{n}X_n^{(k)}X_n^{(k)T} - z\right)^{-1} \\
&+ y_n + y_n z\frac{1}{p}\operatorname{tr}\left(\frac{1}{n}X_nX_n^T - z\right)^{-1}\Bigg| \\
\leq\ &\frac{1}{n} + \frac{|z|}{n\operatorname{Im}(z)}.
\end{aligned}$$

Since this bound holds uniformly for all $k \in \{1,\dots,p\}$, it follows that

$$\max_{k\in[p]}|B(n,k)| \leq \frac{1}{n} + \frac{|z|}{n\operatorname{Im}(z)} \xrightarrow[n\to\infty]{} 0 \quad \text{surely.}$$ □

Theorem 6.19. *In above situation, we find for any fixed* $z \in \mathbb{C}_+$ *that*

$$\max_{k\in[p]}|\Omega_n^{(k)}(z)| \xrightarrow[n\to\infty]{} 0 \quad \textit{almost surely.}$$

Proof. This follows directly by the decomposition (6.15) with Lemma 6.14, Lemma 6.16, Lemma 6.17 and Lemma 6.18. □

Acknowledgments

We are grateful for the very detailed feedback from the referee and the associate editor, which helped improve this paper in many places.

References

[1] N. I. Akhiezer. *The classical moment problem*. Oliver and Boyd, 1965.

[2] Nikita Alexeev, Friedrich Götze, and Alexander Tikhomirov. Asymptotic distribution of singular values of powers of random matrices. *Lithuanian Mathematical Journal*, 50:121–132, 2010. MR2653641

[3] Gerold Alsmeyer. *Wahrscheinlichkeitstheorie*. Number 30 in Skripten zur Mathematischen Statistik. 5th edition, 2007. (Münster).

[4] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2010. MR2760897

[5] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010. MR2567175

[6] Zhidong Bai and Yanqing Yin. Necessary and Sufficient Conditions for Almost Sure Convergence of the Largest Eigenvalue of a Wigner Matrix. *The Annals of Probability*, 16(4):1729–1741, 1988. MR0958213

[7] Florent Benaych-Georges and Antti Knowles. Lectures on the local semicircle law for wigner matrices, 2018. URL https://arxiv.org/abs/1601.04055. MR3792624

[8] Florent Benaych-Georges and Antti Knowles. Lectures on the local semicircle law for wigner matrices. June 2019. URL http://www.unige.ch/~knowles/SCL.pdf. MR3792624

[9] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics*, volume 2. Taylor and Francis, 2016. MR0443141

[10] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, 3rd edition, 1995. MR1324786

[11] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, 2nd edition, 1999. MR0233396

[12] Vladimir Bogachev. *Measure Theory*, volume 1. Springer, 2007. MR2267655

[13] Charles Bordenave and Djalil Chafaï. Around the circular law. *Probability Surveys*, (9):1–89, 2012. MR2908617

[14] Wlodzimierz Bryc, Amir Dembo, and Tiefeng Jiang. Spectral measure of large random hankel, markov and toeplitz matrices. *The Annals of Probability*, 34(1):1–38, 2006. MR2206341

[15] Riccardo Catalano, Michael Fleermann, and Werner Kirsch. Random band and block matrices with correlated entries, 2022. URL https://arxiv.org/abs/2202.04707.

[16] Yuan Shih Chow and Henry Teicher. *Probability Theory*. Springer, 3rd edition, 1997. MR1476912

[17] John B. Conway. *A Course in Functional Analysis*. Springer, 2nd edition, 1997. MR1070713

[18] Rick Durrett. *Probability*. Cambridge University Press, 5th edition, 2019. MR2722836

[19] Jürgen Elstrodt. *Maß- und Integrationstheorie*. Springer, 6th edition, 2009. MR2257838

[20] Ryszard Engelking. *General Topology*. Heldermann, 1989. MR1039321

[21] László Erdős. The matrix dyson equation and its applications for random

matrices, 2019. URL https://arxiv.org/abs/1903.10060. MR3971154
[22] László Erdős and Horng-Tzer Yau. *A Dynamical Approach to Random Matrix Theory*. American Mathematical Society, 2017. MR3699468
[23] László Erdős, Benjamin Schlein, and Horng-Tzer Yau. Local semicircle law and complete delocalization for wigner random matrices. *Communications in Mathematical Physics*, 287:641–655, 2009. MR2481753
[24] Gerd Fischer. *Lineare Algebra*. SpringerSpektrum, 2014.
[25] Michael Fleermann. The empirical spectral distribution of symmetric random matrices with correlated entries. an asymptotic analysis employing the method of moments. Master's thesis, University of Münster, Germany, 2015.
[26] Michael Fleermann. *Global and Local Semicircle Laws for Random Matrices with Correlated Entries*. PhD thesis, FernUniversität in Hagen, Germany, 2019.
[27] Michael Fleermann and Johannes Heiny. High-dimensional sample covariance matrices with curie-weiss entries. *ALEA, Latin American Journal of Probability and Mathematical Statistics*, 17:857–876, 2020. MR4169603
[28] Michael Fleermann and Johannes Heiny. Large sample covariance matrices of gaussian observations with uniform correlation decay, 2022. URL https://arxiv.org/abs/2203.04057.
[29] Stuart Geman. A Limit Theorem for the Norm of Random Matrices. *The Annals of Probability*, 8(2):252–261, 1980. MR0566592
[30] Friedrich Götze, Alexey Naumov, Alexander Tikhomirov, and Dmitry Timushev. On the local semicircular law for wigner ensembles. *Bernoulli*, 24(3):2358–2400, 2018. MR3757532
[31] Friedrich Götze, Alexey Naumov, and Alexander Tikhomirov. Local semicircle law under fourth moment condition. *Journal of Theoretical Probability*, April 2019. MR4125959
[32] Martin Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Vieweg+Teubner, 3rd edition, 2009. MR2503883
[33] Todd Kemp. Math 247a: Introduction to random matrix theory. December 2016. URL http://www.math.ucsd.edu/~tkemp/247A.Notes.pdf.
[34] Werner Kirsch. A survey on the method of moments. October 2015. URL https://www.fernuni-hagen.de/stochastik/docs/pub/momente.pdf.
[35] Achim Klenke. *Probability Theory*. Springer, 3rd edition, 2020. MR3112259
[36] Thomas Koshy. *Catalan Numbers with Applications*. Oxford University Press, 2009. MR2526440
[37] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967. MR0208649
[38] Madan Lal Mehta. *Random Matrices*. Elsevier Academic Press, 2004. MR2129906
[39] James Mingo and Roland Speicher. *Free Probability and Random Matrices*. Elsevier Academic Press, 2004. MR2009835
[40] K.R. Parthasarathy. *Probability measures on metric spaces*. Academic Press, 1967. MR0226684

[41] Leonid Pastur and Mariya Shcherbina. *Eigenvalue Distribution of Large Random Matrices*. American Mathematical Society, 2011. MR2808038
[42] Marc Potters and Jean-Philippe Bouchard. *A First Course in Random Matrix Theory*. Cambridge University Press, 2021.
[43] Michael Reed and Barry Simon. *Fourier Analysis, Self-Adjointness*. Academic Press, 1975. MR0493420
[44] Steve Roman. *Advanced Linear Algebra*. Springer, 3rd edition, 2008. MR2344656
[45] Ludger Rüschendorf. *Mathematische Statistik*. Springer Spektrum, 1st edition, 2014.
[46] Kevin Schnelli and Yuanyuan Xu. Convergence rate to the tracy–widom laws for the largest eigenvalue of sample covariance matrices, 2021. URL `https://arxiv.org/abs/2108.02728`. MR4551561
[47] Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 2003. MR2002723
[48] Satish Shirali and Harkrishan L. Vasudeva. *Metric Spaces*. Springer, 2006. MR2161427
[49] J. A. Shohat and J. D. Tamarkin. *The problem of moments*. American Mathematical Society, 1943. MR0008438
[50] Alexander Soshnikov. Universality at the edge of the spectrum in wigner random matrices. *Communications in Mathematical Physics*, 207:697–733, 1999. MR1727234
[51] Roland Speicher. Lecture notes on random matrices, 2020. URL `https://arxiv.org/abs/2009.05157`. MR3331748
[52] Richard P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press, 2012. MR2868112
[53] Terence Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012. MR2906465
[54] Terence Tao and Van Vu. Random matrices: Universality of ESDs and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010. MR2722794
[55] Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206:127–204, 2011. MR2784665
[56] Konstantin Tikhomirov. Singularity of random Bernoulli matrices. *Annals of Mathematics*, 191(2):593–634, 2020. MR4076632
[57] Craig A. Tracy and Harold Widom. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994. MR1257246
[58] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics*, 62(3):548–564, 1955. MR0077805
[59] Eugene P. Wigner. On the distribution of roots of certain symmetric matrices. *The Annals of Mathematics*, 67(2):225–327, 1958. MR0095527
[60] Stephen Willard. *General Topology*. Addison-Wesley, 1970. MR0264581
[61] Hermann Witting and Ulrich Müller-Funk. *Mathematische Statistik II*. Teubner, 1st edition, 1995. MR1363716
[62] Fuzhen Zhang. *Matrix Theory*. Springer, 2nd edition, 2011. MR2857760