High-dimensional Landscapes and Random Matrices

University of Vienna

Peter Mühlbacher

February 14, 2016

Abstract

Contents

Introduction

0.1 Motivation

Given some Hamiltonian $H_{N,p} : S^{N-1} \to \mathbb{R}$ given by

$$H_{N,p}(\boldsymbol{\sigma}) = \frac{1}{N^{(p-1)/2}} \sum_{i_1,\ldots,i_p=1}^{N} J_{i_1,\ldots,i_p}\sigma_{i_1}\ldots\sigma_{i_p},$$

where the $J_{i_1,\ldots,i_p}$ are independent centered standard Gaussian random variables and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$ referred to as states, one is interested in finding the expected number of critical points $Crt_N(u)$ and the restriction to some index k, denoted by $Crt_{N,k}(u)$, in some region in $(-\infty, Nu]$.

For practical applications like neural networks it may be nice to know how the critical points are distributed, but there is still the open question whether local minima that are close to the global minimum will also yield "good results". From a practical point of view this has been empirically confirmed (Choromanska et al. (2014)), however, it can also be shown theoretically like in Loh and Wainwright (2013) for a far more general class of problems.

0.2 Main Results

Theorem 0.2.1 (Large deviations for $\mathbb{E}[\mathrm{Crt}_{N,k}(u)]$)**.**

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}[Crt_{N,k}(u)] = \Theta_{k,p}(u) \tag{1}$$

and

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}[Crt_N(u)] = \Theta_p(u), \tag{2}$$

for

$$\Theta_p(u) = \begin{cases} \frac{1}{2}\log(p-1) - \frac{p-2}{4(p-1)}u^2 - I_1(u), & \text{if } u \leq -E_\infty \\ \frac{1}{2}\log(p-1) - \frac{p-2}{4(p-1)}u^2, & \text{if } -E_\infty \leq u \leq 0 \\ \frac{1}{2}\log(p-1), & \text{if } 0 \leq u \end{cases}$$

and

$$\Theta_{k,p}(u) = \begin{cases} \frac{1}{2}\log(p-1) - \frac{p-2}{4(p-1)}u^2 - (k+1)I_1(u), & \text{if } u \leq -E_\infty \\ \frac{1}{2}\log(p-1) - \frac{p-2}{p}, & \text{if } u \geq E_\infty \end{cases},$$

where $E_\infty = E_\infty(p) = 2\sqrt{\frac{p-1}{p}}$ *and* $I_1 : (-\infty, E_\infty] \to \mathbb{R}$ *is given by*

$$I_1(u) = \frac{2}{E_\infty^2}\int_u^{-E_\infty} \sqrt{z^2 - E_\infty^2}dz$$

and is the the rate function of the LDP for the smallest eigenvalue of the GOE.

Theorem 0.2.2 (Layered structure)**.** *For all* $k \geq 0$ *and* $\varepsilon > 0$ *we have*

$$\limsup_{N\to\infty} \frac{1}{N}\log \mathbb{P}\left(\left\{\sum_{i=k}^{\infty} Crt_{N,i}(-E_k - \varepsilon) > 0\right\}\right) < 0, \tag{3}$$

where $E_k = E_k(p)$ *is chosen (uniquely because of strict monotonicity) such that* $\Theta_{k,p}(-E_k) = 0$.

0.3 Outlook

According to Sagun et al. (2014) there is empirical evidence that the energy of critical points concentrates around the limiting floor value, however, up to now and to the best of my knowledge there is no proof for this theoretical statement and in particular (from a theoretical point of view) nothing is known about the speed of convergence to the expected value that is presented here.

Chapter 1

Partial Results

To be able to show the previously mentioned results we first need to prove some other theorems which require prerequisites from various other fields.

1.1 GOE

Definition 1.1. *Symmetric $N \times N$ matrices $H = H_N$ with $\mathbb{E}H_{ij} = 0$ and $\mathbb{E}H_{ij}^2 = 1 + \delta_{ij}$.*

Remark (Density in the space of matrices)**.** *Their density is given by the Gaussian measure*

$$d\mathbb{P}(H) = Z_N^{-1} \exp\left(-N\frac{1}{4}trH^2\right) \tag{1.1}$$

with normalization constant $Z_N = \int d\mathbb{P}(H) \prod_{1\leq i\leq j\leq N} dH_{ij}$ which is a shorter way of writing $Z_N^{-1} \exp\left\{-N\frac{1}{4}\left(2\sum_{i<j}^N H_{ij}^2 + \sum_{i=j}^N H_{ij}^2\right)\right\}$.

Theorem 1.1.1 (Joint probability density of eigenvalues)**.** *The joint probability density Q_N of the unordered eigenvalues $\{\lambda_i\}_{i=1}^N$ of the GOE is given by*

$$Q_N(\mathrm{d}\lambda_1, \ldots, \mathrm{d}\lambda_N) = C_N \prod_{i<j} |\lambda_i - \lambda_j| \prod_i \exp(-N\lambda_i^2/4)\, \mathrm{d}\lambda_i,$$

where, without loss of generality, we assumed the variances of the entries to be normalized to one and for some $C_N > 0$ which is uniquely determined by normalization.

Proof. This proof will follow the one given in Liu (2000).

Let H be some element of the GOE. Since it is symmetric there is a decomposition in $H = UDU^T$ with U orthogonal and $D = \text{diag}(\lambda_1, \ldots, \lambda_N)$. Thus we can write

$$H_{ij} = \sum_k \lambda_k U_{ik} U_{jk} \tag{1.2}$$

and, using the orthogonality of U: $\sum_k U_{ki} U_{kj} = \delta_{ij}$. Using those results one can infer that $\sum_{i,j} H_{ij}^2 = \sum_i \lambda_i^2$.

The key idea is to make use of the change of variables formula to go from $\mathbb{P}(\lambda_1, \ldots, \lambda_N, \alpha_1, \ldots, \alpha_{N(N-1)/2})$ to $\mathbb{P}(H)$. One should think of the $\{\alpha_i\}_{i=1}^{N(N-1)/2}$ as the parameters that determine the matrix U, which, together with the eigenvalues, uniquely determines H.

To do that we first need some information on the determinant of the Jacobian J of the change of variables: From equation 1.2 we see the linearity of H_{ij} in the eigenvalues λ_k which implies that $\partial H_{ij}/\partial\alpha$ is linear in the eigenvalues[1]. Hence, $\det J$ has to be a polynomial of degree $N(N-1)/2$ in the eigenvalues. If two eigenvalues coincide U cannot be uniquely determined anymore and thus the inverse of the transformation is not unique, meaning that $\det J = 0$. So the determinant of the Jacobian must vanish for all $\lambda_i = \lambda_j, i \neq j$, which is achieved if it contains a factor $\lambda_i - \lambda_j$. However, there are exactly $N(N-1)/2$ such factors and since that is just the degree of the polynomial it follows that we have completely accounted for J's dependence on the eigenvalues by writing

$$\det J = \prod_{i<j} (\lambda_i - \lambda_j) h(\alpha_1, \ldots, \alpha_{N(N-1)/2}).$$

Now we can write

$$\mathbb{P}(\lambda_1, \ldots, \lambda_N, \alpha_1, \ldots, \alpha_{N(N-1)/2}) =$$

$$\mathbb{P}(H) |\det J| = Z_N^{-1} \exp\left(-N \frac{1}{4} \sum_i \lambda_i^2\right) |\prod_{i<j} (\lambda_i - \lambda_j) h(\alpha_1, \ldots, \alpha_{N(N-1)/2})|.$$

Integrating out the dependence on $\{\alpha_k\}_{k=1}^{N(N-1)/2}$ yields the desired result. □

[1] To be precise, it is linear in the *vector* $(\lambda_i)_{i=1}^N$.

1.2 Gaussian Random Fields

For what follows it we shall fix some notations like in Auffinger et al. (2013):

First of all, instead of working with $H_{N,p}$ it is convenient to define

$$f(\sigma) \equiv f_{N,p}(\sigma) = N^{-1/2} H_{N,p}(N^{1/2}\sigma), \tag{1.3}$$

which has variance one on the unit sphere. We use $\langle \cdot, \cdot \rangle$ to denote the usual Euclidean scalar product, as well as the scalar product on any tangent space $T_\sigma S^{N-1}$ and $(E_i)_{1 \le i < N}$ to denote an arbitrary orthonormal frame field, that is a set of $N-1$ vector fields E_i on S^{N-1} such that $\{E_i(\sigma)\}$ is an orthonormal basis of $T_\sigma S^{N-1}$. Furthermore we write $\nabla f(\sigma) = (f_i(\sigma))_{1 \le i < N}$ for the gradient $(E_i f(\sigma))_{1 \le i < N}$, $\nabla^2 f = (f_{ij})_{1 \le i,j < N}$ for the covariant Hessian of f on S^{N-1} and $\det \nabla^2 f(\sigma)$ for the determinant of the matrix $(\nabla^2 f(E_i, E_j)(\sigma))_{1 \le i,j < N}$.

Definition 1.2 (Real valued random fields)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and T a topological space. Then a measurable mapping $f : \Omega \to \mathbb{R}^T$ is called a real valued random field.*

In what follows we will not distinguish between $f(\omega)(t)$ and $f(t)$.

In other words, a random field f is a collection $\{f(t) : t \in T\}$ where every $f(t)$ is a real valued random variable.

Definition 1.3 (Real valued Gaussian fields)**.** *Let f be a real valued random field with some parameter set T as before. Then f is a real valued Gaussian field if for every $(t_1, \dots, t_n) \in T^n$ $(1 \le n < \infty)$ the distributions $(f(t_1), \dots, f(t_n))$ are multivariate Gaussian.*

Furthermore the functions $m(t) = \mathbb{E}[f(t)]$ and $C(t, t') = \mathbb{E}[(f(t) - m(t))(f(t') - m(t'))]$ are called the mean and covariance functions of f, respectively.

Note that this definitions yields the usual definition for a multivariate Gaussian $\mathbb{R}^d$-valued random variable in case $|T| = d$ is finite.

The formulation and proofs of the following two lemmata will closely follow Auffinger et al. (2013) (and Adler and Taylor (2007) for the proofs).

Lemma 1.2.1 (Covariances of f)**.** *Let f be as defined in 1.3, then $f(\sigma), f_i(\sigma), f_{ij}(\sigma)$ are centered Gaussian random variables for all $1 \le i, j, k, l < N$ and $\sigma \in S^{N-1}$ whose joint distribution is determined by*

$$\begin{aligned}
\mathbb{E}[f(\sigma)^2] &= 1, \\
\mathbb{E}[f(\sigma)f_i(\sigma)] = \mathbb{E}[f_i(\sigma)f_{jk}(\sigma)] &= 0, \\
\mathbb{E}[f_i(\sigma)f_j(\sigma)] = -\mathbb{E}[f(\sigma)f_{ij}(\sigma)] &= p\delta_{ij}, \\
\mathbb{E}[f_{ij}(\sigma)f_{kl}(\sigma)] &= p(p-1)(\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}) + p^2\delta_{ij}\delta_{kl}. \quad (1.4)
\end{aligned}$$

Proof. Because of the rotational symmetry we may assume without loss of generality that σ is the north pole $n = (0, \dots, 0, 1)$. We define the projection $\Psi : S^{N-1} \to \mathbb{R}^{N-1}, (x_1, \dots, x_{N-1}, x_N) \mapsto (x_1, \dots, x_{N-1})$ which is a chart for some neighbourhood U of n and set $\overline{f} = f \circ \Psi^{-1}$ which is a Gaussian process on $\Psi(U)$, satisfying

$$C(x,y) = Cov(\overline{f}(x), \overline{f}(y)) = \left(\sum_{i=1}^{N-1} x_i y_i + \sqrt{(1-\sum_{i=1}^{N-1} x_i^2)(1-\sum_{i=1}^{N-1} y_i^2)} \right)^p.$$

Choosing an orthonormal frame field (E_i) such that it satisfies $E_i(n) = \partial/\partial x_i$ with respect to the chart Ψ the Christoffel symbols vanish and hence the covariant Hessian $(f_{ij}(n))$ agrees with the usual Hessian of $\overline{f}$ at 0.

Thus it suffices to prove the analogous identities for $\overline{f}(0)$, $\frac{\partial}{\partial x_i}\overline{f}(0)$, $\frac{\partial^2}{\partial x_i \partial x_j}\overline{f}(0)$. Differentiation under the integral sign yields

$$\mathbb{E}\left\{ \frac{\partial^k \overline{f}(s)}{\partial s_{i_1} \dots \partial s_{i_k}} \frac{\partial^l \overline{f}(t)}{\partial t_{i_1} \dots \partial t_{i_l}} \right\} = \frac{\partial^{k+l} C(s,t)}{\partial s_{i_1} \dots \partial s_{i_k} \partial t_{i_1} \dots \partial t_{i_l}}, \qquad (1.5)$$

which, in turn, after some simple algebra gives 1.4. □

Now we can formulate one of the central identities that allow us to use methods from random matrix theory to study the Hamiltonian of interest.

Lemma 1.2.2 (Moments under conditioning)**.** *Using the same assumptions as in the previous lemma and under the conditional distribution $\mathbb{P}[\cdot|f(\sigma) = x], x \in \mathbb{R}$ the random variables $f_{ij}(\sigma)$ are independent Gaussian variables satisfying*

$$\begin{aligned}
\mathbb{E}[f_{ij}(\sigma)|f(\sigma) = x] &= -xp\delta_{ij}, \\
\mathbb{E}[f_{ij}(\sigma)^2|f(\sigma) = x] &= (1+\delta_{ij})p(p-1). \qquad (1.6)
\end{aligned}$$

Alternatively, the random matrix $(f_{ij}(\sigma))$ (under the conditional distribution $\mathbb{P}[\cdot|f(\sigma) = x], x \in \mathbb{R}$) has the same distribution as

$$M^{N-1}\sqrt{2(N-1)p(p-1)} - xpI,$$

where M^{N-1} is a $(N-1)\times(N-1)$ GOE matrix as defined in 1.1 and I is the identity matrix.

Proof. Equations 1.6 can be seen from the last two equations of 1.4. The second statement follows by plugging in the definitions. □

1.3 LDP

Note that some theorems in this section will not be proven because their proofs are rather technical in nature and do not contribute very much to a better understanding of the subject.

Definition 1.4 (Large deviation principle)**.** *Given some separable completely metrizable topological space*[2] *X, a sequence of Borel probability measures $\{\mathbb{P}_n\}$ on X is said to satisfy a **large deviation principle** with speed $\{a_n\}$ and rate $I : X \to [0,\infty]$ if a_n goes to $+\infty$ and I is some lower semi-continuous functional such that for each Borel measurable set $E \subseteq X$ we have*

$$\limsup_n a_n^{-1} \log(\mathbb{P}_n(E)) \leq -\inf_{x\in\overline{E}} I(x)$$

and

$$\liminf_n a_n^{-1} \log(\mathbb{P}_n(E)) \geq -\inf_{x\in E^\circ} I(x).$$

*The lower semi-continuity implies that the sets $\{x \in X : I(x) \leq c\}$ are closed in X for all $c \geq 0$. If they are also compact for all $c \geq 0$, I is called a **good rate function**.*

The following definition and theorem are taken from Dembo and Zeitouni (2009):

Definition 1.5 (Exponential equivalence)**.** *Two families of probability measures $\{\mu_\varepsilon\}, \{\tilde{\mu}_\varepsilon\}$ on some metric space (Y,d) are called exponentially equivalent if there exist probability spaces $\{\Omega, \mathcal{B}_\varepsilon, \mathbb{P}_\varepsilon\}$ and two families of Y-valued random variables $\{Z_\varepsilon\}, \{\tilde{Z}_\varepsilon\}$ with joint laws $\{\mathbb{P}_\varepsilon\}$ and marginals $\{\mu_\varepsilon\}, \{\tilde{\mu}_\varepsilon\}$, respectively, such that the following holds:*

For each $\delta > 0$, the set $\{\omega : (Z_\varepsilon, \tilde{Z}_\varepsilon) \in \Gamma_\delta\}$ is $\mathcal{B}_\varepsilon$ measurable, and

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mathbb{P}_\varepsilon(\Gamma_\delta) = -\infty,$$

where $\Gamma_\delta = \{(y,\tilde{y}) : d(y,\tilde{y}) > \delta\} \subseteq Y \times Y$.

[2] Henceforth such a space will be referred to as Polish space.

For instance, if the random variables are real valued and the rate of the LDP is N (i.e. $\varepsilon = N^{-1}$), this asserts that the joint probability $\mathbb{P}_N$ of the two random variables of the area that is farther than δ away from the diagonal $x = y$ goes to zero like some $\exp(-cN^p)$ for some constants $c > 0, p > 1$. A quick sketch illustrates that the two probability measures have to be "quite" similar to satisfy the above definition – this can be made precise by the following theorem:

Theorem 1.3.1 (Rate functions of exponentially equivalent measures)**.** *If a LDP with good rate function I holds for the probability measures $\{\mu_\varepsilon\}$, which are exponentially equivalent to $\{\tilde{\mu}_\varepsilon\}$, then the same LDP holds for $\{\tilde{\mu}_\varepsilon\}$.*

Proof. In the general case this is a very technical result and hence we will refer the interested reader to Dembo and Zeitouni (2009), theorem 4.2.13. □

We are particularly interested in getting LDPs of the k-th largest eigenvalues of the GOE, but it turns out in order to get those we first need a LDP for the law of the empirical measure $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{\lambda_i}$.

Theorem 1.3.2 (LDP of Wigner's semicircle law for the GOE)**.** *Let $I(\mu) := \frac{1}{2}\left(\int x^2 \,\mathrm{d}\mu(x) - \Sigma(\mu) - \frac{3}{4} - \frac{1}{2}\log 2\right)$, for $\Sigma(\mu) := \int\int \log|x-y| \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$. Then:*

1. *I is well defined on the space of probability measures $\mathcal{M}(\mathbb{R})$ on $\mathbb{R}$ endowed with the weak topology and takes values in $[0, \infty]$.*
2. *$I(\mu) = \infty \Leftrightarrow \int x^2 \,\mathrm{d}\mu(x) = \infty$ or $\exists A \subseteq \mathbb{R} : \mu(A) > 0 \wedge \exp\{-\inf_{\nu\in\mathcal{M}(A)} \int\int \log|x-y|^{-1} \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\} = 0$, where the latter condition is usually referred to as the existence of a set of "null logarithmic capacity".*
3. *I is a good rate function.*
4. *I is a convex function.*
5. *I achieves its unique minimum value at Wigner's semicircle law.*

In particular the law of the empirical measure μ^N satisfies a LDP with good rate function I and speed N^2.

A proof can be found in Ben Arous and Guionnet (1997).

With this we can state the main result of this section. Note that without loss of generality we set the variances of the entries of the GOE to 1.

Theorem 1.3.3 (LDP for the k-th largest eigenvalue of the GOE)**.** *For each fixed $k \geq 1$, the k-th largest eigenvalue λ_{N-k+1} of the GOE satisfies a LDP with speed N and good rate function*

$$I_k(x) = kI_1(x) = \begin{cases} k\int_2^x \sqrt{\frac{z^2}{4} - 1}\,\mathrm{d}z, & \text{if } x \geq 2 \\ \infty, & \text{else.} \end{cases}$$

Proof. We will prove the two cases $x < 2$ and $x \geq 2$ separately and start with the first one:

Since $\lambda_{N-k+1} \leq x < 2$ we have that the empirical spectral measure $\mu_N((x,2]) \leq \frac{k-1}{N}$. However, Wigner's semicircle law implies that $\mu_{sc}((x,2]) = \lim_{N\to\infty} \mu_N((x,2]) > 0$. So there exists a closed set $A \in \mathcal{P}(\mathbb{R})$ such that $\mu_{sc} \notin A$, but the set of all empirical spectral distributions with k-th largest eigenvalue being smaller than x being a subset of A, i.e. $\{\tilde{\mu}_N : \lambda_{N-k+1} \leq x\} \subseteq A$. Because of the exponential tightness with speed N^2 of $\{\mu_N\}_N$ we have $Q_N(A) \leq \exp(-cN^2)$ for some $c > 0$ which concludes the case for $x < 2$.

The second part will be split up in showing the upper and lower bounds of the equation. We will start with the upper one

$$\limsup \frac{1}{N} \log Q_N(\lambda_{N-k+1} \geq x) \leq -I_k(x). \tag{1.7}$$

However, instead of showing this directly we will use the obvious[3] upper bound

$$Q_N(\lambda_{N-k+1} \geq x) \leq Q_N(\max_{i=1}^N |\lambda_i| \geq M) + Q_N(\lambda_{N-k+1} \geq x, \max_{i=1}^N |\lambda_i| < M),$$

which, together with the estimate (see Ben Arous et al. (2001))

$$Q_N(\max_{i=1}^N |\lambda_i| \geq M) \leq \exp(-NM^2/9) \tag{1.8}$$

for M large enough and all N[4] lets us prove 1.7 by showing

$$\limsup \frac{1}{N} \log Q_N(\max_{i=1}^N |\lambda_i| \leq M, \lambda_{N-k+1} \geq x) \leq -I_k(x),$$

[3] This is basically $\mathbb{P}(E) = \mathbb{P}(\neg F) + \mathbb{P}(E \wedge F)$.

[4] It follows by integrating the elementary estimate $|x - \lambda_i|e^{-\lambda_i^2/4} \leq (|x| + |\lambda_i|)e^{-\lambda_i^2/4} \leq 2|x| \leq e^{x^2/8}$ which holds for $|x| \geq M \geq 8$ and using the fact that $Z_{N-1}/Z_N \leq e^{CN}$ for some C independent of N (see Selberg's formula in Mehta (2004)).

for all $M > x > 2$.

To that end let $\overline{Q}^N_{N-k}$ be a rescaled version of the joint law of unordered eigenvalues $\overline{Q}_{N-k}$, given by

$$\overline{Q}^N_{N-k}(\lambda \in \cdot) = \overline{Q}_{N-k}(\sqrt{1-k/N}\lambda \in \cdot),$$

where the factor $\sqrt{1-k/N} = \sqrt{\frac{N-k}{N}}$ serves the purpose of "changing the units" of an eigenvalue λ from an $(N-k)$-dimensional distribution (which is normalised by $(N-k)^{-1/2}$) to an N-dimensional one.

Similar considerations apply for the new normalisation constant

$$C^k_N = (1-k/N)^{(N-k)(N-k+1)/4}\frac{\overline{Z}_{N-k}}{\overline{Z}_N},$$

where $\overline{Z}_N$ is the normalization factor for $\overline{Q}_N$.

For $x \in \mathbb{R}$ and $\mu \in \mathcal{P}(\mathbb{R})$ we define

$$\Phi(x,\mu) = \int_{\mathbb{R}} \log|x-y|\,\mathrm{d}\mu(y) - \frac{x^2}{4},$$

which can be shown (Ben Arous et al. (2001), p.50) to be upper semi-continuous on $[-M,M]\times\mathcal{P}([-M,M])$ and continuous on $[x,y]\times\mathcal{P}([-M,M])$ for x, y, M such that $2 < M < x < y$.

Using 1.1.1 we see that

$$\begin{aligned} &Q_N(\max_{i=1}^N|\lambda_i| \le M, \lambda_{N-k+1} \ge x) = \\ &\quad N!Z_N^{-1}\int_{[x,M]^k}\int_{[-M,M]^{N-k}}\left(\prod_{1\le i<j\le N}|\lambda_i-\lambda_j|\prod_{i=1}^N\exp(-N\lambda_i^2/4)\right) \\ &\qquad \mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_{N-k}\,\mathrm{d}\lambda_{N-k+1}\ldots\mathrm{d}\lambda_N, \end{aligned}$$

where the $N!$ comes from dropping the $\mathbf{1}_{\lambda_1\le\ldots\lambda_N}$ term. This can be bounded from above by

$$\begin{aligned} C^k_N\frac{N!}{(N-k)!}\int_{[x,M]^k}\prod_{N-k<i<j\le N}|\lambda_i-\lambda_j|\int_{[-M,M]^{N-k}} e^{(N-k)\sum_{i=N-k+1}^N\Phi(\lambda_i,\mu_{N-k})} \\ \overline{Q}^N_{N-k}(\mathrm{d}\lambda_1,\ldots,\mathrm{d}\lambda_{N-k})\,\mathrm{d}\lambda_{N-k+1}\ldots\mathrm{d}\lambda_N, \end{aligned} \tag{1.9}$$

where μ_{N-k} is determined by the variables $\lambda_1,\ldots,\lambda_{N-k}$ we are integrating against.

We write $B(\rho,\delta)$ for the open ball with radius δ in $\mathcal{P}(\mathbb{R})$ with center ρ and let $B_M(\rho,\delta) = B(\rho,\delta)\cap\mathcal{P}([-M,M])$. Noting that $|\lambda_i-\lambda_j|\le 2M$ on the domain of integration and $e^{(N-k)\Phi(\lambda_i,\mu_{N-k})}\le (2M)^{N-k}$ we get the following upper bound on 1.9:

$$C_N^k\frac{N!}{(N-k)!}(2M)^{k(k-1)/2}\Big\{\Big(\int_x^M e^{(N-k)\sup_{\mu\in B_M(\mu_{sc},\delta)}\Phi(x,\mu)}\,\mathrm{d}x\Big)^k \\ +(2M)^{N-k}\overline{Q}_{N-k}^N(\mu_{N-k}\notin B(\mu_{sc},\delta))\Big\}. \quad (1.10)$$

This may be seen as a crude upper bound on some case detection whether or not μ_{N-k} is in some δ-neighbourhood of the semicircle law μ_{sc}. Intuitively the probability of the second case occuring should vanish as $N\to\infty$ and this can be made precise by showing that μ_{N-k} under $\overline{Q}_{N-k}^N$ satisfies the same LDP as under $\overline{Q}_{N-k}$ (where it is already known to vanish by 1.3.2). To do so we observe that for all Lipschitz functions $h:\mathbb{R}\to\mathbb{R}$ of norm at most 1 and $N\ge 2k$ we have the following inequality:

$$\left|(N-k)^{-1}\sum_{i=1}^{N-k}h\left(\sqrt{1-kN^{-1}}\lambda_i\right)-h(\lambda_i)\right|\le cN^{-1}\max_{i=k}^{N-k}|\lambda_i|, \quad (1.11)$$

for some $c>0$ independent of N and k.

The measure $\overline{Q}_{N-k}^N$ can be thought of describing the random variable that picks $N-k$ values as described by $\overline{Q}_{N-k}$ and then renormalizes them so that they on the same scale as the ones that would be picked by $\overline{Q}_N$ (so basically picking $N-k$ values from the distribution for N values).

Equation 1.11 lets us bound the metric d from the definition of the exponential equivalence by $cN^{-1}\max_{i=k}^{N-k}|\lambda_i|$, which in turn, using inequality 1.8 with $M=c^{-1}N$ yields the exponential equivalence of the two measures $\overline{Q}_N$ and $\overline{Q}_{N-k}^N$ because Γ_δ (as in definition 1.5) is of the order $\exp(-CN^3)$ for some $C>0$.

Hence, the second term in 1.10 is exponentially negligible for any $\delta>0$ and $M<\infty$, which, in turn, implies that

$$Q_N(\lambda_{N-k+1}\ge x,\max_{i=1}^N|\lambda_i|\le M)\le \\ \limsup_{N\to\infty}\frac{1}{N}\log C_N^k+k\lim_{\delta\searrow 0}\sup_{z\in[x,M],\mu\in B_M(\mu_{sc},\delta)}\Phi(z,\mu). \quad (1.12)$$

Since Φ is upper semi-continuous (notice that $\Phi(z,\mu) = \inf_{\eta>0} \Phi_\eta(z,\mu)$, where $\Phi_\eta(z,\mu) := \int \log(\max(|z-y|,\eta))\,\mathrm{d}\mu(y) - z^2/4$ which is continuous on $[-M,M] \times \mathcal{P}([-M,M])$) the term $\lim_{\delta\searrow 0} \sup_{z\in[x,M],\mu\in B_M(\mu_{sc},\delta)} \Phi(z,\mu)$ simplifies to $\sup_{z\in[x,M]} \Phi(z,\mu_{sc})$.

With $\mathrm{supp}(\mu_{sc}) = [-2,2]$, the derivative $D(z) := \frac{\partial}{\partial z}\Phi(z,\mu_{sc})$ exists for $z \geq 2$ and can be shown to be $-\sqrt{z^2/4-1} \leq 0$ (see Ben Arous and Guionnet (1997), proof of lemma 2.7) using several changes of variables and the residue method. Hence $\sup_{z\in[x,M]} \Phi(z,\mu_{sc}) = \Phi(x,\mu_{sc}) = -1/2 - I_1(x)$.

With Selberg's formula (Mehta (2004)) it can be shown that, in the LDP, the normalization constant $\lim_{N\to\infty} N^{-1} \log C_N^k = k/2$.

Putting those logarithmic asymptotics together completes the proof of the upper bound.

To prove the complementary lower bound we fix $y > x > r > 2$ and $\delta > 0$. Similarly to the steps 1.9 and 1.10 we get

$$
\begin{aligned}
&Q_N(\lambda_{N-k+1} \geq x) \geq \\
&\quad \overline{Q}_N(\lambda_N \in [x,y], \ldots, \lambda_{N-k+1} \in [x,y] \max_{i=1}^{N-k} |\lambda_i| \leq r) \geq \\
&\quad KC_N^k \exp\Big(k(N-k) \inf_{z\in[x,y],\mu\in B_r(\mu_{sc},\delta)} \Phi(z,\mu)\Big) \overline{Q}_{N-k}^N\big(\mu_{N-k} \in B_r(\mu_{sc},\delta)\big),
\end{aligned}
$$

for some $K = K(x,y,k) > 0$. Again, by using the LDP of μ_{N-k} under $\overline{Q}_{N-k}^N$ we see that $\lim_{N\to\infty} Q_{N-k}^N(\mu_{N-k} \notin B_r(\mu_{sc},\delta)) = 0$.

Using the behaviour of C_N^k we get

$$
\liminf_{N\to\infty} N^{-1} \log Q_N(\lambda_{N-k+1} \geq x) \geq k(2^{-1} + \inf_{z\in[x,y],\mu\in B_r(\mu_{sc},\delta)} \Phi(z,\mu)),
$$

which, after letting $\delta \searrow 0$ and $y \searrow x$ (note the continuity of Φ in the used range of parameters), yields the desired result. □

Another result that will be used later on is Varadhan's lemma which can be viewed as the LDP equivalent of Laplace's principle which states that $\lim_{\varepsilon\to 0} \varepsilon \int_A \exp(-\phi(x)/\varepsilon)\,\mathrm{d}x = -\inf_{x\in A} \phi(x)$ or more loosely $\int_A \exp(-\phi(x)/\varepsilon)\,\mathrm{d}x \approx \exp(-\inf_{x\in A} \phi(x)/\varepsilon)$.

Lemma 1.3.4 (Varadhan)**.**

state Varadhan's lemma

1.4 Morse Theory

motivate kac rice formula

Chapter 2

Putting Them Together

2.1 Intermediary Results

To prove our main results we first need some refinements for the expected values of critical values.

Theorem 2.1.1 (Express $\mathbb{E}[Crt_{N,k}(B)]$ in terms of the GOE)**.** *For all B Borel sets, N, $p \geq 2$ and $k \in \{0, \dots, N-1\}$ we have*

$$\mathbb{E}[Crt_{N,k}(B)] = 2\sqrt{\frac{2}{p}}(p-1)^{\frac{N}{2}}\mathbb{E}^N_{GOE}\left[e^{-N\frac{p-2}{2p}(\lambda^N_k)^2}\mathbf{1}\left\{\lambda^N_k \in \sqrt{\frac{p}{2(p-1)}}B\right\}\right] \tag{2.1}$$

and

$$\mathbb{E}[Crt_N(B)] = 2N\sqrt{\frac{2}{p}}(p-1)^{\frac{N}{2}}\int_{\sqrt{\frac{p}{2(p-1)}}B} exp\left\{-\frac{N(p-2)x^2}{2p}\right\}\rho_N(x)dx. \tag{2.2}$$

Proof.

First of all let us show that we can apply lemma **??**. To do so we use the same notation as in section 1.2. Because of the rotational symmetry and since S^{N-1} can be covered by a finite number of copies of an open neighbourhood of some point it suffices to investigate only one point. We will choose the north-pole n and see that $(f_i(n), f_{ij}(n))$ is not degenerate. Since the covariances are continuous this is also true for some neighbourhood U. Hence the conditions of lemma **??** are satisfied. Again due to the rotational symmetry the integrand does not depend on σ, so we get

prove that Gaussian + nondegenerate implies Morse function

formulate it!

$$\mathbb{E}Crt_{N,k}(B) =$$
$$\mathrm{vol}(S^{N-1})\mathbb{E}[|\det \nabla^2 f(n)|\mathbf{1}\{i(\nabla^2 f(\sigma)) = k, f(n) \in \sqrt{N}B\}|\nabla f(n) = 0]d\mathbb{P}(\nabla f(n) = 0)$$

To compute this expectation we condition on $f(n)$

$$\mathbb{E}[|\det \nabla^2 f(n)|\mathbf{1}\{i(\nabla^2 f(\sigma)) = k, f(n) \in \sqrt{N}B\}|\nabla f(n) = 0] =$$
$$\mathbb{E}\left[\mathbb{E}[|\det \nabla^2 f(n)|\mathbf{1}\{i(\nabla^2 f(\sigma)) = k, f(n) \in \sqrt{N}B\}|f(n)]\right]$$

By lemma the following equality for the interior expectation

$$\mathbb{E}[|\det \nabla^2 f(n)|\mathbf{1}\{i(\nabla^2 f(\sigma)) = k, f(n) \in \sqrt{N}B\}|f(n)] =$$
$$(2(N-1)p(p-1))^{(N-1)/2}\mathbb{E}_{GOE}^{N-1}\Big[|\det(M^{N-1} - \sqrt{p/(2(N-1)(p-1))}f(n)I)|$$
$$\times \mathbf{1}\{i(M^{N-1} - \sqrt{p/(2(N-1)(p-1))}f(n)I) = k, f(n) \in \sqrt{N}B\}\Big].$$

Substituting back, writing t^2 for $p/(2(N-1)(p-1))$ as well as G for $\sqrt{Np/(2(N-1)(p-1))}B$ and setting X to be a real valued Gaussian random variable with zero mean and variance t^2 we get

$$\mathbb{E}\Big[|\det(M^{N-1} - XI)\mathbf{1}\{i(M^{N-1} - XI) = k, X \in G\}\Big] =$$
$$(2\pi t^2)^{-1/2}\int_G \exp(-\frac{x^2}{2t^2})\mathbb{E}_{GOE}^{N-1}\Big[|\det(M - xI)|\mathbf{1}\{i(M - xI) = k\}\Big]\,\mathrm{d}x. \tag{2.3}$$

Now observe that the event $\{i(M - xI) = k\}$ is equal to the event $\{A_k^N(x)\}$, where $A_k^N(x) = \{\lambda^{N-1} : \lambda_0^{N-1} \leq \cdots \leq \lambda_{k-1}^{N-1} < x \leq \lambda_k^{N-1} \leq \cdots \leq \lambda_{N-2}^{N-1}\}$.

Using 1.1.1 we get

$$\mathbb{E}_{GOE}^{N-1}[|\det(M - xI)|\mathbf{1}\{i(M - xI) = k\}] = \int_{A_k^N(x)}\prod_{i=1}^{N-2}|\lambda_i^{N-1} - x|Q_{N-1}(\mathrm{d}\lambda^{N-1}).$$

The definition of $A_k^N(x)$ and the determinant in the equation above suggests considering x as the $k+1$-th smallest eigenvalue of a $N \times N$ GOE

matrix. Performing the corresponding rescaling, that is, change of variables given by $\lambda_i^{N-1} = \sqrt{N/(N-1)}\lambda_i^N$ for $i \in \{0, \dots, k-1\}$, $\lambda_i^{N-1} = \sqrt{N(/(N-1)}\lambda_{i+1}^N$ for $i \in \{k, \dots, N-2\}$ and $x = \sqrt{N/(N-1)\lambda_k^N}$, writing out Q_N and substituting back in equation 2.3 we obtain

$$\frac{Z_N}{Z_{N-1}\sqrt{2\pi t^2}}(N/(N-1))^{(N+2)(N+1)/4}$$
$$\times \mathbb{E}_{GOE}^N\Big[\exp\Big(\frac{N(\lambda_k^N)^2}{2} - \frac{\frac{N}{N-1}(\lambda_k^N)^2}{2t^2}\Big)\mathbf{1}\{\lambda_k^N \in \sqrt{(N-1)/N)}G\}\Big],$$

where the constants Z_N is given by

$$\frac{1}{N!}(2\sqrt{2})^N N^{-N(N+1)/4}\prod_{i=1}^{N}\Gamma(1+i/2)$$

which can be computed from Selberg's integral (cf. Mehta (2004)).

Some straightforward algebra then yields the first claim.

The second one is obtained by summing over $k \in \{0, \dots, N-1\}$.

□

Appendix A

Additional Theorems

The following section will follow Tao (2010) quite closely while trying to be as self-contained as possible.

A.1 Wigner's Semicircle Law

Definition A.1 (Wigner matrix)**.** *A Wigner matrix M is a complex, Hermitian matrix with independent and identically distributed entries M_{ij} for $i \geq j$ and with mean 0 and variance 1 for $i > j$. The diagonal entries M_{ii} have bounded mean and variance.*

If M_n is an n-dimensional Wigner matrix we know that the operator norm $\|M_n\|_{OP}$ is typically of size $O(\sqrt{n})$, so it is natural to define the empirical spectral distribution (ESD) as follows:

Definition A.2 (ESD)**.**

$$\mu_{\frac{1}{\sqrt{n}}M_n} := \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(M_n)/\sqrt{n}},$$

where $\lambda_1(M_n) \leq \cdots \leq \lambda_n(M_n)$ are the ordered, real eigenvalues of M_n.

Since we are considering random matrices the ESDs will be random as well and thus it is interesting to ask if there is a measure on the real line μ such that it is the weak limit $\mu_{M_n/\sqrt{n}} \rightharpoonup \mu$ of $\mu_{M_n/\sqrt{n}}$, that is $\int_{\mathbb{R}} \varphi d\mu_{M_n/\sqrt{n}}$ converges in almost surely against $\int_{\mathbb{R}} \varphi d\mu$ for all $\varphi \in C_c(\mathbb{R})$. This can also be derived from the more general definition of convergence in probability or almost surely, but we will not do that here.

Surprisingly such a limit μ exists and is even deterministic.

Theorem A.1.1 (Wigner's semicircle law)**.** *Let M_n be the top left $n \times n$ minors of an infinite Wigner matrix, then the ESDs $\mu_{M_n/\sqrt{n}}$ converge in probability*[1] *to the Wigner semicircle distribution given by*

$$\mu_{sc} := \begin{cases} \frac{1}{2\pi}\sqrt{4-|x|^2}dx, & \text{if } |x| \leq 2 \\ 0, & \text{else} \end{cases} =: \frac{1}{2\pi}\sqrt{4-x^2}_+ \,\mathrm{d}x.$$

A rough outline of the proof is given by this list of intermediary results that will be shown:

1. Show that without loss of generality we can set the diagonal elements to zero and bound all other entries by some constant. Additionally some classic concentration of measure results will be shown.

2. To show that $\mu_n \rightharpoonup \mu$ almost surely, it suffices to show that the respective Stieltjes transforms converge almost surely, pointwise in the upper half plane, i.e. $\mu_n \rightharpoonup \mu \Leftrightarrow \forall z \in \mathbb{C} : Im(z) > 0 s_{\mu_n}(z) \to s_\mu(z)$ almost surely.

3. The Stieltjes transform $s_n := s_{\mu_{M_n/\sqrt{n}}}$ is "stable in n", i.e. $s_n(z) = s_{n-1}(z) + O(\frac{1}{n})$, where O can depend on z and even $s_n(z) - \mathbb{E}s_n(z) \to 0$ almost surely.

4. Derive the semicircle law by deriving the recursion $\mathbb{E}s_n(z) = -\frac{1}{z+\mathbb{E}s_n(z)} + o(1)$, where, again, $o(1)$ will depend on z and "inverting" the Stieltjes transform.

Remark. *Note that instead of step 4 one could have plugged in the semicircle distribution and simplified the proof by just checking that this is indeed the limit. This is not done here because we want to see how the Stieltjes transform method can be used to derive such a conclusion without knowing about it beforehand.*

Also, there are other proofs (e.g. Boutet de Monvel and Khorunzhy (2015)) specifically for the GOE/GUE (instead of the more general Wigner matrices) which exploit their symmetries to shorten the proof considerably.

A.1.1 Preliminary Reductions

Lemma A.1.2. *For the matrices M_n as given in A.1.1 it can be assumed without loss of generality that the diagonal entries are zero and the absolute values $|[M_n]_{ij}|$ are bounded by some constant $C > 0$ which does not depend on i, j or n.*

[1] Even almost sure convergence can be shown, but we will not do so in this proof.

Proof. For every n define the normalized random variable $\overline{X}_n$ by setting the diagonal elements to zero and bounding every entry by zero[2], i.e.

$$[\overline{X}_n]_{ij} := \begin{cases} [M_n]_{ij}/\sqrt{n}\mathbf{1}_{|[M_n]_{ij}|\leq C} - \mathbb{E}\left[[M_n]_{ij}\mathbf{1}_{|[M_n]_{ij}|\leq C}\right], & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}.$$

Now we want to show that convergence of $\mu_{\overline{X}_n}$ to μ implies convergence of μ_{X_n} to the same limit μ. To do so it suffices to show that $|\int \varphi \, \mathrm{d}\mu_{X_n} - \int \varphi \, \mathrm{d}\mu_{\overline{X}_n}| \to 0$ for every smooth φ with compact support. In particular every such φ is a Lipschitz function with Lipschitz-constant $\|\varphi'\|_{L^\infty}$. By denoting the eigenvalues of $X_n, \overline{X}_n$ as $\lambda_i, \overline{\lambda}_i$ and invoking the Wielandt-Hoffmann inequality we get the following bound:

$$|\int \varphi \, \mathrm{d}\mu_{X_n} - \int \varphi \, \mathrm{d}\mu_{\overline{X}_n}| \leq \|\varphi'\|_{L^\infty} \frac{1}{n} \sum_{i=1}^{n} |\lambda_i - \overline{\lambda}_i| \leq$$

$$\|\varphi'\|_{L^\infty} \left[\frac{1}{n} \sum_{i=1}^{n} |\lambda_i - \overline{\lambda}_i|^2\right]^{1/2} \leq \|\varphi'\|_{L^\infty} \left[\frac{1}{n} tr\left((X_n - \overline{X}_n)^2\right)\right]^{1/2}.$$

Now we need to show that for every $\varepsilon, \delta > 0$ we can find a C such that $\mathbb{P}\left(\frac{1}{n} tr\left((X_n - \overline{X}_n)^2\right) > \varepsilon\right) < \delta$ for large enough n.

Using the definitions one sees that

$$\frac{1}{n} tr\left((X_n - \overline{X}_n)^2\right) \leq \frac{1}{n^2} \sum_{i \neq j} \sqrt{n} \left[X_n - \overline{X}_n\right]_{ij}^2 + \frac{1}{n} \sum_i [X_n]_{ii}^2,$$

where the second term vanishes almost surely as $n \to \infty$ according to the strong law of large numbers. So we can use this to bound the probability of the event $\frac{1}{n} tr\left((X_n - \overline{X}_n)^2\right) > \varepsilon$ by

$$\mathbb{P}\left(\frac{1}{n} tr\left((X_n - \overline{X}_n)^2\right)\right) \leq \frac{1}{n^2} \sum_{i \neq j} \mathbb{P}\left(\sqrt{n}[X_n - \overline{X}_n]_{ij}^2 > \varepsilon\right).$$

To see that this goes to zero for large C we apply Markov's inequality, yielding

[2]Since the distribution changes we also need to recenter them by subtracting its mean value.

$$\mathbb{P}(\sqrt{n}[X_n - \overline{X}_n]_{ij}^2 > \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}\left[[\sqrt{n}X_n - \overline{X}_n]_{ij}^2\right] \leq$$
$$\frac{\sqrt{n}}{\varepsilon}\left(\mathbb{E}[[X_n]_{ij}^2 \mathbf{1}_{\sqrt{n}|[X_n]_{ij}|>C}] + \mathbb{E}[[X_n]_{ij}\mathbf{1}_{\sqrt{n}|[X_n]_{ij}|>C}]^2\right).$$

Since the entries $\sqrt{n}[X_n]_{ij} = [M_n]_{ij}$ have finite variance the right hand side will go to zero as $C \to \infty$ which proves the claim. □

Theorem A.1.3 (Talagrand's concentration inequality)**.** *Let $K > 0$ and $X_1, \ldots, X_n$ be independent complex variables with $|X_i| < K$ for all $1 \leq i \leq n$. Identifying $\mathbb{C}$ with $\mathbb{R}^2$, let $F : \mathbb{C}^n \to \mathbb{R}$ a 1-Lipschitz, convex function. Then for every $\lambda > 0$ one has*

$$\mathbb{P}(|F(X) - MF(X)| \geq \lambda K) \leq C\exp^{-c\lambda^2}$$

and

$$\mathbb{P}(|F(X) - \mathbb{E}F(X)| \geq \lambda K) \leq C\exp^{-c\lambda^2}$$

for some absolute constants $c, C > 0$, where $MF(X)$ is the median of $F(X)$.

For the proof refer to `https://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/#boof0`.

Theorem A.1.4 (Weilandt-Hoffmann inequality)**.** *For Hermitian*[3] *A, B, where $\|B\|_F^2 := tr(B^2)^{\frac{1}{2}}$ is the Frobenius norm, we have*

$$\sum_{j=1}^{n} |\lambda_j(A+B) - \lambda_j(A)|^2 \leq \|B\|_F^2.$$

Proof. We fix B to be $B - A$ and show the equivalent statement

$$\sum_{i=1}^{n} |\lambda_i(A) - \lambda_i(B)|^2 \leq tr\left((A-B)^2\right)$$

and denote by D_M the diagonal matrix $\mathrm{diag}(\lambda_1(M), \ldots, \lambda_n(M))$ and U the matrix that diagonalises $B = UD_BU^*$ in the basis of A, such that $tr(AB) = tr(D_AUD_BU^*) = \sum_{i,j} \lambda_i(A)\lambda_j(B)|[U]_{ij}|^2$.

Now we have to get some bounds on the "worst case" scenario, that is when $tr(AB)$ is big. To do so we define $v_{ij} := |[U]_{ij}|^2$ and notice

[3]The original Weilandt-Hoffmann inequality holds for normal operators, but we will restrict ourselves to Hermitian ones for simplicity's sake here.

that the above $tr(AB)$ is linear in v_{ij}. Due to the orthogonality of U one has $\sum_i v_{ij} = \sum_j v_{ij} = 1$ so we can bound $tr(AB)$ from above by $\sup_{v_{ij}\geq 0, \sum_i v_{ij} = \sum_j v_{ij} = 1} \sum_{i,j} \lambda_i(A)\lambda_j(B)v_{ij}$.

This is an optimization problem of a linear functional over a convex set of doubly stochastic matrices. Hence the maximum is attained at the extreme points which are exactly the permutations. Of all the permutations the identity gives the maximal value which, after writing out the first inequality proves the theorem. □

Theorem A.1.5 (Chernoff inequality)**.** *Let $X_1, \ldots, X_n$ be independent scalar random variables with $|X_i| \leq K$ almost surely, mean μ_i and variance σ_i^2. Then for any $\lambda > 0$ there are absolute constants $c, C > 0$ such that*

$$\mathbb{P}(|S_n - \lambda| \geq \lambda\sigma) \leq C \max(e^{-c\lambda^2}, e^{-c\lambda\sigma/K}),$$

where $\lambda := \sum_i \lambda_i$ and $\sigma^2 := \sum_i \sigma_i^2$.

Proof. We begin with some preliminary reductions, namely that it is sufficient to assume X_i to be real valued by taking real and imaginary parts. Furthermore without loss of generality we set $K = 1$ by dividing X_i through K, center X_i (i.e. set μ_i to zero by subtracting its mean), and, by symmetry, only show the upper tail estimate $\mathbb{P}(S_n \geq \lambda\sigma) \leq C \max(e^{-c\lambda^2}, e^{-c\lambda\sigma})$ (with different constants c, C).

To do so we compute the exponential moments $\mathbb{E}\exp(tS_n)$ for some $0 \leq t \leq 1$. By independence of the X_i we have $\mathbb{E}\exp(tS_n) = \prod_{i=1}^n \mathbb{E}\exp(tX_i)$. The $\exp(tX_i)$ can be expanded into a Taylor series $1+tX_i+O(t^2X_i^2\exp(O(t)))$ which, after taking expectation and noting that $|X_i| \leq 1$, yields

$$\mathbb{E}\exp(tX_i) = 1 + O(t^2\sigma_i^2\exp(O(t))) = \exp(O(t^2\sigma_i^2))$$

and hence $\mathbb{E}\exp(tS_n) = \exp(O(t^2\sigma^2))$. By Markov's inequality we conclude with

$$\mathbb{P}(S_n \geq \lambda\sigma) \leq \exp(O(t^2\sigma^2) - t\lambda\sigma),$$

which proves the claim after minimising the right hand side in $t \in [0, 1]$. □

Theorem A.1.6 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values in ranges $R_1, \ldots, R_n$ and let $F : R_1 \times \cdots \times R_n \to \mathbb{C}$, such that for every $1 \leq i \leq n$ we have $|F(x_1, \ldots, x_i, \ldots, x_n) - F(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$. Then for any $\lambda > 0$ one has*

$$\mathbb{P}(|F(X) - \mathbb{E}F(X)| > \lambda\sigma) \leq C\exp^{-c\lambda^2},$$

for some absolute[4] *constants* $c, C > 0$ *and* $\sigma = \sum_{i=1}^n c_i^2$.

Proof. Similar as above we may assume that F is real, it suffices to show the upper tail estimate $\mathbb{P}(F(X) - \mathbb{E}F(X) \geq \lambda\sigma^2) \leq Ce^{-c\lambda^2}$ and try to bound the exponential moment $\mathbb{E}e^{tF(X)}$.

To get a better idea how the exponential moment behaves we write it in such a way that we can use the fact that F does not fluctuate "too much", i.e. we consider the conditional expectation $\mathbb{E}(e^{tF(X)}|X_1, \ldots, X_{n-1})$ for $X_1, \ldots, X_{n-1}$ fixed and write this as

$$\mathbb{E}(e^{tF(X)}|X_1, \ldots, X_{n-1}) = \mathbb{E}(e^{tY}|X_1, \ldots, X_{n-1})e^{t\mathbb{E}(F(X)|X_1,\ldots,X_{n-1})},$$

where $Y := F(X) - \mathbb{E}(F(X)|X_1, \ldots, X_{n-1})$.

Now we want to control the first term on the right hand side. Since tY has mean zero and variance $t^2c_n^2$ we can (just in the proof of Chernoff's inequality) expand to get a Taylor series and take expectations to get

$$\mathbb{E}(e^{tY}|X_1, \ldots, X_{n-1}) \leq e^{O(t^2c_n^2)}.$$

Integrating out the conditioning (note that X_i are independent) we get the following bound on the exponential moment

$$\mathbb{E}e^{tF(X)} \leq e^{O(t^2c_n^2)}\mathbb{E}e^{t\mathbb{E}(F(X)|X_1,\ldots,X_{n-1})}.$$

Basically we reduced the problem by one dimension and noting that $\mathbb{E}(F(X)|X_1, \ldots, X_{n-1})$ is a function $F_{n-1}(X_1, \ldots, X_{n-1}$ with the same properties as in the theorem's assumptions on F the above calculations can be performed iteratively n times to get the upper bound

$$\mathbb{E}e^{tF(X)} \leq \exp(\sum_{i=1}^n O(t^2c_i^2) + t\mathbb{E}F(X)).$$

Dividing by $\exp(t\mathbb{E}F(X)))$ and applying Markov's inequality yields

$$\mathbb{P}(F(X) - \mathbb{E}F(X) \geq \lambda\sigma) \leq \exp(O(t^2\sigma^2) - t\lambda\sigma),$$

which, after minimising the right hand side in $0 \leq t \leq 1$, proves the claim. □

[4]Constants that maintain the same value wherever they occur. In particular applying McDiarmid's inequality in different settings we do not need to consider C_n, c_n, but can still write C, c.

A.1.2 Stieltjes Transform

Definition A.3 (Stieltjes transform)**.** *For a probability measure μ we write s_μ for its Stieltjes transform*

$$\int_{\mathbb{R}} \frac{1}{x-z} \, \mathrm{d}\mu(x).$$

As mentioned above, s_n will be a shorthand for $s_{\mu_{M_n/\sqrt{n}}}$.

Lemma A.1.7 (Properties of the Stieltjes transform)**.** *In the following let μ be some probability measure.*

1. *For $z = a + ib$ we have $Im\frac{1}{x-z} = \frac{b}{(x-a)^2+b^2} > 0$.*
2. *s_μ is analytic in $\mathbb{C} \setminus supp(\mu) \supset \mathbb{C}_+$.*
3. *We can bound the absolute value as well as the derivatives by $|\frac{d^j}{dz^j} s_\mu(z)| \leq O(|Im(z)|^{-(j+1)})$ for all $j \in \{0, 1, \dots\}$.*

Proof. The first property is trivial, the second one can be seen by integrating s_μ over any contour not containing the support of μ, interchanging the order of integration and noting that integrating $\frac{1}{x-z}$ gives 0 by Cauchy's integral formula ($\frac{1}{x-z}$ being holomorphic outside the support of μ). The Stieltjes transform being holomorphic (and thus analytic) follows by Morera's theorem.

The third property can be obtained by using $\frac{1}{x-z} \leq \frac{1}{Im(z)}$ and using Cauchy's integral formula integrating this inequality. □

Corollary A.1.7.1. *From A.1.7.1 it follows that s_μ is a Herglotz function and thus (e.g. Teschl (2009)) $Im(s_\mu(. + ib)) \rightharpoonup \pi\mu$ as $b \to 0^+$ in the vague topology or equivalently (by $\overline{s_\mu(z)} = s_\mu(\overline{z})$)*

$$\frac{s_\mu(. + ib) - s_\mu(. - ib)}{2\pi i} \rightharpoonup \mu. \tag{A.1}$$

Note that this can also be seen by writing $Im(s_\mu)$ as the convolution $\pi\mu * P_b(a)$ with the Poisson kernels $P_b(x) := \frac{1}{\pi}\frac{b}{x^2+b^2} = \frac{1}{b}P_1(\frac{x}{b})$ which form a family of approximations to the identity.

Theorem A.1.8 (Stieltjes continuity theorem)**.** *For μ_n (realisations of) random measures and μ a deterministic measure the following statement holds:*

$\mu_n \rightharpoonup \mu$ almost surely in the vague topology if and only if $s_{\mu_n}(z) \to s_\mu(z)$ almost surely for every $z \in \mathbb{C}_+$.

Proof. "⇒": If $\mu_n \rightharpoonup \mu$ in the vague topology almost surely against a deterministic limit μ, then $\forall \phi \in C_c(\mathbb{R}) : \lim_{n\to\infty} \int_{\mathbb{R}} \phi \, \mathrm{d}\mu_n = \int_{\mathbb{R}} \phi \, \mathrm{d}\mu$ by definition and, by taking the completion, for all bounded, continuous functions vanishing at infinity. The function $x \mapsto \frac{1}{x-z}$ for some $z \in \mathbb{C}$ with $Im(z) > 0$ is bounded and continuous on $\mathbb{R}$ and hence $s_{\mu_n}(z) \to s_\mu(z)$ almost surely.

"⇐": One can, up to an arbitrary small error $\varepsilon > 0$, approximate $\int_{\mathbb{R}} \phi \, \mathrm{d}\mu$ by $\int_{\mathbb{R}} \phi * P_b \, \mathrm{d}\mu = \frac{1}{\pi} \int_{\mathbb{R}} \phi(a) s_\mu(a+ib) \, \mathrm{d}a$ (and analogously for μ_n). Thus we have $\frac{1}{\pi} \int_{\mathbb{R}} \phi(a)(s_\mu(a+ib) - s_{\mu_n}(a+ib)) \, \mathrm{d}a$ being equal to the difference (we are interested in) $\int_{\mathbb{R}} \phi \, \mathrm{d}\mu - \int_{\mathbb{R}} \phi \, \mathrm{d}\mu_n$ up to an error ε and by dominated convergence (the Stieltjes transform of a measure is bounded for every $z \in \mathbb{C}_+$ and vanishes outside some compact set) we have convergence in the vague topology. □

A.1.3 Stableness and Concentration of Measure

In the following we keep using the notation as defined in A.1.1. To show that $s_n(z) = s_{n-1}(z) + O_z(1/n)$ we first need to prove the following theorem:

Theorem A.1.9 (Cauchy's interlacing theorem)**.** *For any $n \times n$ Hermitian matrix A_n with top left minor A_{n-1} and eigenvalues of descending order ($\lambda_i \geq \lambda_{i+1}$) we have:*

$$\lambda_{i+1}(A_n) \leq \lambda_i(A_{n-1}) \leq \lambda_i(A_n),$$

for all $1 \leq i < n$.

Proof. Using the min-max/max-min theorems ($\lambda_i(A) = \inf_{dim(V)=n-i+1} \sup_{v\in V:\|v\|=1} \langle Av, v\rangle$ and $\lambda_i(A) = \sup_{dim(V)=i} \inf_{v\in V:\|v\|=1} \langle Av, v\rangle$ respectively, c.f. Teschl (2009) p.141) and writing S_{n-i+1} for $\{v \in span\{a_i, \ldots, a_n\} : \|v\| = 1\}$, where $A_{n-1}a_j = \lambda_j a_j$ and P an orthogonal projection such that $P^* A_n P = A_{n-1}$ we have

$$\begin{aligned}\lambda_i(A_{n-1}) = &\sup_{v\in S_i, \|v\|=1} v^* A_{n-1} v = \\ &\sup_{v\in S_i, \|v\|=1} v^* P^* A_n P v \geq \\ &\inf_{dim(V)=n-i} \sup_{v\in V, \|v\|=1} v^* A_n v = \lambda_{i+1}(A_n),\end{aligned}$$

and

$$\lambda_i(A_n) = \inf_{dim(V)=n-i+1} \sup_{v\in V,\|v\|=1} v^* A_n v \geq$$
$$\sup_{v\in S_i,\|v\|=1} v^* P^* A_n P v =$$
$$\sup_{v\in S_i,\|v\|=1} v^* A_{n-1} v = \lambda_i(A_{n-1}).$$

□

Remembering the identity $Im(s_{\mu_n(a+ib)}) = \pi\mu * P_b(a)$ and that $supp\mu_n$ consists of finitely many points, we have $Im(s_{\mu_n}) = \pi\frac{1}{n}\sum_{\lambda_i} \frac{b}{(\lambda_i-a)^2+b^2}$ which suggests that it is important to take a closer look at the function $x \mapsto \frac{b}{(x-a)^2+b^2}$ to compare s_{μ_n} with $s_{\mu_{n-1}}$.

Lemma A.1.10. *For fixed $z \in \mathbb{C}_+$ the Stieltjes transform is "stable" in n, i.e.*

$$s_n(z) = s_{n-1}(z) + O\left(\frac{1}{n}\right)$$

Proof. The idea is to use the Cauchy interlacing law and apply it to the previously mentioned identity by seeing that

$$\sum_{j=1}^{n-1} \frac{b}{\lambda_j(M_{n-1})/\sqrt{n}-a} - \sum_{j=1}^{n} \frac{b}{\lambda_j(M_n)/\sqrt{n}-a}$$

Up to the dimensional factors these two sums correspond to s_n and s_{n-1} and because of Cauchy's interlacing law this is an alternating sum, giving

$$\sqrt{n(n-1)}s_{n-1}(\sqrt{n/(n-1)}(a+ib)) - ns_n(a+ib) = O(1).$$

Now using the fact that the Stieltjes transform s_n is analytic away from the support of μ_n (A.1.7.2) and using the bound for its derivatives (A.1.7.3) we can approximate $s_{n-1}(\cdot)$ by $s_{n-1}(\sqrt{n/(n-1)}\cdot)$ and hence the statement holds. □

Using McDiarmid's inequality one gets

$$\mathbb{P}(|s_n(z) - \mathbb{E}s_n(z)| \geq \lambda/\sqrt{n}) \leq C\exp^{-c\lambda^2}, \tag{A.2}$$

for all $\lambda > 0$ and some constants $c, C > 0$.

From the Borel-Cantelli lemma we see that for every z away from the real line $s_n(z) - \mathbb{E}s_n(z)$ converges almost surely to zero since, for every fixed $\varepsilon > 0$, the sum $\sum_n \mathbb{P}(d(s_n - \mathbb{E}s_n, 0) \geq \varepsilon) \leq C\sum_n \exp^{-cn\varepsilon^2} < \infty$ which is obtained by setting $\lambda = \varepsilon\sqrt{n}$.

A.1.4 Finding the Semicircle Law

We start off by using the following identity

$$s_n(z) := \int_{\mathbb{R}} \frac{1}{x-z} d\mu_n(x) = \frac{1}{n} tr\left(\frac{1}{\sqrt{n}} M_n - zI_n\right)^{-1},$$

which holds for every $z \in \mathbb{C} \setminus supp(\mu_n)$. Because of the linearity of the trace we also have

$$\mathbb{E}s_n(z) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} M_n - zI_n\right)^{-1}\right]_{jj} = \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} M_n - zI_n\right)^{-1}\right]_{nn}, \tag{A.3}$$

where the last equality holds because all of the random variables $\left[(M_n/\sqrt{n} - zI_n)^{-1}\right]_{jj}$ have the same distribution.

To calculate one entry of an inverse of a matrix we use Schur's complement, which tells us that (under the assumptions that all the occurring inverse matrices exist)

$$[(M_n/\sqrt{n} - zI_n)^{-1}]_{nn} = -\left(z + \frac{1}{n} X^*(\frac{1}{\sqrt{n}} M_{n-1} - zI_{n-1})^{-1} X\right)^{-1},$$

where $X \in \mathbb{C}^{n-1}$ is the top right column of M_n with the bottom entry removed and the diagonal elements have been set to zero as justified in A.1.1.

This inverse exists because, for $z \in \mathbb{C}_+$, the imaginary part $Q := Im\left((\frac{1}{\sqrt{n}} M_{n-1} - zI_{n-1})^{-1}\right)$ is positive definite according to the spectral theorem. To see this notice that this holds for arbitrary Hermitian matrices M (instead of $\frac{1}{\sqrt{n}} M_{n-1}$) since their spectrum is on the real line. Thus, by the spectral theorem, we can write $Q = Im \int \frac{1}{\mu_M - z} dM(\mu_M)$ for some projection valued measure dM and since $x \mapsto \frac{1}{x-z}$ is a Herglotz function its imaginary part will be greater than zero for $z \in \mathbb{C}_+$. As a result the imaginary part of the integrand (which is the imaginary part of the eigenvalues) will be greater than zero. We conclude by noticing that $Im(z) > 0$ plus something of the form $\langle Qx, x\rangle$ for $Q \geq 0$ will have imaginary part strictly greater than zero and hence the inverse exists.

The next step is to get a better understanding of the resolvent $R := (\frac{1}{\sqrt{n}} M_{n-1} - zI_{n-1})^{-1}$ and its product $\langle RX, X\rangle$. Clearly R and X are independent, so we may treat R almost like a deterministic matrix. Furthermore,

again due to the spectral theorem, $\|R\|_{op}$ is at most $O(1)$. By the strong law of large numbers $\|X\| = O(\sqrt{n})$ almost surely and by Chernoff's inequality this holds with overwhelming probability.

In the following we will show some results for some deterministic matrix A which has roughly the same properties as R (i.e. $A \geq 0$ and $\|A\|_{OP} = O(1)$).

Noting that the function $X \mapsto \sqrt{\langle AX, X\rangle}$ is a Lipschitz function with operator norm $O(1)$ and remembering from A.1.1 that we can safely assume the entries to be bounded, we can invoke Talagrand's concentration inequality to get

$$\mathbb{P}(|\sqrt{\langle AX, X\rangle} - \mathbb{M}\sqrt{\langle AX, X\rangle}| \geq \lambda) \leq C \exp^{-c\lambda^2}$$

for any $\lambda > 0$. On the other hand we have $\sqrt{\langle AX, X\rangle} = O(\|X\|) = O(\sqrt{n})$ with overwhelming probability (since the operator norm in the non-deterministic case is only controlled with overwhelming probability). Hence the median $\mathbb{M}\sqrt{\langle AX, X\rangle} = O(\sqrt{n})$ and considering the square $\langle AX, X\rangle$, we conclude that

$$\mathbb{P}(|\langle AX, X\rangle - \mathbb{M}\langle AX, X\rangle| \geq \lambda\sqrt{n}) \leq C \exp^{-c\lambda^2}$$

with some (possibly different) $c, C > 0$, since taking the square of $\sqrt{\langle AX, X\rangle}$ amounts to getting an additional factor $O(\sqrt{n})$ and the median is of the same magnitude as the random variable.

Because of this concentration of measure result we may replace the median with the expected value, yielding

$$\mathbb{P}(|\langle AX, X\rangle - \mathbb{E}\langle AX, X\rangle| \geq \lambda\sqrt{n}) \leq C \exp^{-c\lambda^2} \tag{A.4}$$

for the the case where A is deterministic and positive definite. One can extend this result to arbitrary matrices of operator norm $O(1)$ by noting that it holds for Hermitian matrices $M = M^* = M_+ - M_-$ ($M_+ \geq 0, M_- \geq 0, \|M\|_{OP} = O(1)$) by applying the triangle inequality and for general matrices M (which are diagonalizable almost surely) of operator norm $\|M\|_{op} = O(1)$ by using their diagonalization $M = U^*DU = U^*D_+U - U^*D_-U$, U unitary and $D_+, D_- \geq 0$, to get the bound $|X^*MX| \leq |X^*U^*D_+UX| + |X^*U^*D_-UX|$.

Remark. *By using conditional expectations the above results also hold true for random matrices R with $R \geq 0$ and $\|R\|_{OP} = O(1)$ as long as it is independent of X. The idea is to write all the above statements as $\mathbb{P}(E|\{A = R\}) = \frac{\mathbb{P}(E \cap \{A=R\})}{\mathbb{P}(\{A=R\})} = \mathbb{P}(E)$ for all events E.*

Now we want to know what $\mathbb{E}\langle RX, X\rangle$ actually is. Because of the linearity of the expectation we write it as $\sum_{i,j=1}^{n-1} \mathbb{E}[\overline{X_i} R_{ij} X_j]$. Since the X_i and R_{ij} are independent we can write that as $\sum_{i,j=1}^{n-1} \mathbb{E}[\overline{X_i} X_j]\mathbb{E}[R_{ij}]$, but as the X_i are iid with mean zero and variance one this double sum simplifies to the expectation of the trace of R

$$\sum_{i=1}^{n-1} \mathbb{E} R_{ii}.$$

Noticing that, up to some "almost correct" normalization factors, $tr(R) = tr\left((M_{n-1}/\sqrt{n} - zI_{n-1})^{-1}\right)$ is the Stieltjes transform $s_{n-1}(z)$. To be more precise we have

$$tr(R) = n\sqrt{\frac{n}{n-1}} s_{n-1}\left(\sqrt{\frac{n}{n-1}} z\right),$$

but because of the smoothness of the Stieltjes transform for $z \in \mathbb{C}_+$ these factors do not play a role in the limit $n \to \infty$, i.e. $tr(R) = n(s_{n-1}(z) + o(1))$.

So using the concentration of measure results for the Stieltjes transform (A.2) and for $\langle AX, X\rangle$(A.4), remembering that latter also holds for random matrices as long as they are independent (A.1.4), we see that

$$\langle RX, X\rangle = n(s_{n-1}(z) + o(1))$$

with overwhelming probability. Substituting back in Schur's complement (A.3) we get[5]

$$\mathbb{E} s_n(z) = -(z + \mathbb{E} s_n(z))^{-1} + o(1).$$

To say something about the limit we first need to ensure $\lim_{n\to\infty} \mathbb{E} s_n$ exists. This is indeed the case since $\mathbb{E} s_n$ is locally equicontinuous and locally uniformly bounded away from the real line. Applying the Arzelá-Ascoli theorem we get the existence of a subsequence that converges locally uniformly to a limit s, which is again a Herglotz function. Note that, by the concentration of measure for Stieltjes transforms, there is only one possible limit (so $\lim_{n\to\infty} \mathbb{E} s_n$ is well defined) and $s_n(z)$ even converges almost surely to $s(z)$. As a further result we get

$$s(z) = -(z + s(z))^{-1},$$

[5]Note that we need the concentration of measure results from above to justify "interchanging" expectation and taking the resolvent!

where the quadratic formula gives

$$s(z) = -\frac{z \pm \sqrt{z^2 - 4}}{2}$$

From $\lim_{a\to\infty} s_\mu(a + ib) = 0$ for every Stieltjes transform of a fixed measure μ we see that we need to take $s(z) = \frac{-z+\sqrt{z^2-4}}{2}$.

We conclude the proof with the Stieltjes inversion formula, yielding the famous result

$$\frac{s(. + ib) - s(. - ib)}{2\pi i} \rightharpoonup \frac{1}{2\pi}\sqrt{4 - x^2}_+ \, \mathrm{d}x = \mu_{sc}$$

as $b \to 0^+$, which can be verified by an application of the Cauchy integral formula.

Bibliography

Adler, R. J. and J. E. Taylor (2007), *Random Fields and Geometry*. Springer-Verlag New York.

Auffinger, Antonio, Gérard Ben Arous, and Jiří Černỳ (2013), "Random matrices and complexity of spin glasses." *Communications on Pure and Applied Mathematics*, 66, 165–201.

Ben Arous, G., Amir Dembo, and Alice Guionnet (2001), "Aging of spherical spin glasses." *Probability theory and related fields*, 120, 1–67.

Ben Arous, G. and A. Guionnet (1997), "Large deviations for wigner's law and voiculescu's non-commutative entropy." *Probability Theory and Related Fields*, 108, 517–542, URL `http://dx.doi.org/10.1007/s004400050119`.

Boutet de Monvel, A. and A. Khorunzhy (2015), "Some elementary results around the wigner semicircle law." URL `https://www.physik.uni-bielefeld.de/bibos/old-bibos-site/01-03-035.pdf`.

Choromanska, Anna, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun (2014), "The loss surface of multilayer networks." *arXiv preprint arXiv:1412.0233*.

Dembo, Amir and Ofer Zeitouni (2009), *Large deviations techniques and applications*, volume 38. Springer Science & Business Media.

Liu, Y (2000), "Statistical behavior of the eigenvalues of random matrices."

Loh, Po-Ling and Martin J Wainwright (2013), "Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima." In *Advances in Neural Information Processing Systems*, 476–484.

Mehta, M.L. (2004), *Random Matrices*. Pure and Applied Mathematics, Elsevier Science.

Sagun, Levent, V Ugur Guney, Gerard Ben Arous, and Yann LeCun (2014), "Explorations on high dimensional landscapes." *arXiv preprint arXiv:1412.6615*.

Tao, T. (2010), "254a, notes 4: The semi-circular law." URL `http://terrytao.wordpress.com/2010/02/02/254a-notes-4-the-semi-circular-law/`.

Teschl, G. (2009), *Mathematical methods in quantum mechanics*, volume 99 of *Graduate Studies in Mathematics*. American Mathematical Society, URL `http://www.mat.univie.ac.at/~gerald/ftp/book-schroe/schroe.pdf`. With applications to Schrödinger operators.