Large deviations for the largest eigenvalue of sub-Gaussian Wigner matrices

High Dimensional Statistics and Random Matrices

Île de Porquerolles

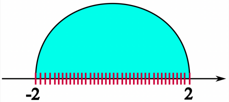
2023/06/14

Nick Cook, Duke University

Based on joint works with Raphaël Ducatez, Alice Guionnet and Amir Dembo

Let H be an $N \times N$ normalized real sub-Gaussian Wigner matrix
i.e. real symmetric with $H_{ij} = \frac{1}{\sqrt{N}} X_{ij}$ for $\{X_{ij}\}_{i \le j}$ iid copies of a standardized variable X with sub-Gaussian law μ.

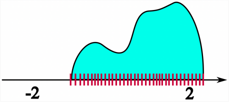
Eigenvalues $\lambda_N \le \cdots \le \lambda_1$.



Semicircle law: The ESD $\hat{\mu}_H = \frac{1}{N}\sum_{i=1}^N \delta_{\lambda_i}$ concentrates around the semicircle measure $d\sigma(x) = \frac{1}{2\pi}(4 - x^2)_+^{1/2} dx$.

* Quantitative: $\mathbb{P}(|\hat{\mu}_H(f) - \sigma(f)| > \varepsilon) \lesssim \exp(-c\varepsilon^2 N^2)$ for f convex, 1-Lipschitz if μ has bounded support [Guionnet–Zeitouni '00].
* Fluctuations: $N\hat{\mu}_H(f)$ converges to a Gaussian for f smooth. (Optimal condition $f \in H^{1/2+\varepsilon}$ [Landon–Sosoe '22].)

* LDP for GOE case [Ben Arous–Guionnet '97]:



$$-\frac{1}{N^2} \log \mathbb{P}(\hat{\mu}_H \sim \nu) \sim \mathcal{I}(\nu)$$

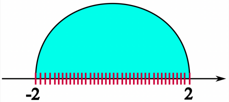
$$= -\frac{1}{2}\int\int \log|x-y| d\nu(x) d\nu(y) + \frac{1}{4}\int x^2 d\nu(x) - c\,.$$

The spectrum of Wigner matrices

Let H be an $N \times N$ normalized real sub-Gaussian Wigner matrix

i.e. real symmetric with $H_{ij} = \frac{1}{\sqrt{N}} X_{ij}$ for $\{X_{ij}\}_{i \leq j}$ iid copies of a standardized variable X with sub-Gaussian law μ.

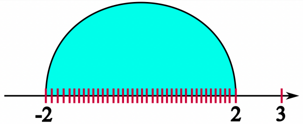
Eigenvalues $\lambda_N \leq \cdots \leq \lambda_1$.



Convergence at the edge: $\lambda_1 \to 2$ w.h.p. [Füredi–Komlós '81].

- ∗ Quantitative: $\mathbb{P}(|\lambda_1 - 2| > \varepsilon) \lesssim \exp(-c\varepsilon^2 N)$ if μ has bounded support.
- ∗ Fluctuations: $N^{2/3}(\lambda_1 - 2) \Rightarrow TW_1$. [Tracy–Widom, Forester '94] (GOE), [Soshnikov '99]
- ∗ LDP for GOE case [Ben Arous–Dembo–Guionnet '99]:

$$-\frac{1}{N} \log \mathbb{P}(\lambda_1 \sim x) \sim \mathcal{I}_\gamma(x) := \begin{cases} \frac{1}{2}\int_2^x \sqrt{y^2 - 4}\,dy & x \geq 2 \\ \infty & x < 2. \end{cases}$$



$\mathbb{P}(\lambda_1 \sim 3) \approx e^{-0.715N}$.

A universal (!) LDP (Guionnet–Husson '18)

For centered μ and $\Lambda_\mu(t) := \log \int_{\mathbb{R}} e^{tx} d\mu(x)$, μ is ...

$$\begin{array}{rll} \textit{sub-Gaussian (SG)} \text{ if} & \Lambda_\mu(t) \le Kt^2 & \forall t \in \mathbb{R}, \text{ constant } K < \infty \\ \textit{sharp sub-Gaussian (SSG)} \text{ if} & \Lambda_\mu(t) \le \frac{1}{2}t^2 = \Lambda_\gamma(t) & \forall t \in \mathbb{R} \end{array}$$

where $\gamma(dx) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. Ex: Rademacher, Unif$[-\sqrt{3}, \sqrt{3}]$ are SSG.

Theorem (Guionnet–Husson '18)

Assume μ is sharp sub-Gaussian. Then λ_1 satisfies an LDP with speed N and good rate function $\mathcal{I}_\gamma$. In particular, for each fixed $x \in \mathbb{R}$,

$$\frac{1}{N} \log \mathbb{P}(|\lambda_1 - x| \le \delta) = -\mathcal{I}_\gamma(x) + o(1).$$

(In this talk we write "$o(1)$" for errors going to zero after $N \to \infty$ then $\delta \downarrow 0$.)

Strategy of tilting by spherical integrals since applied in many contexts: $A + UBU^*$, matrices with a variance profile, perturbed Wigner matrices, generalized sample covariance matrices. [Guionnet–Maïda, Husson, Belinschi–Guionnet–Huang, McKenna, Biroli–Guionnet, Maillard, Husson–McKenna]

Question: How does the matrix "typically" achieve $\lambda_1 \sim x$?

Non-universal LDPs for heavier-tailed/sparse matrices

For some matrices without uniformly sub-Gaussian entries, deviations of the spectrum are due to localization phenomena – events involving $o(N^2)$ entries.

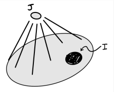
- *Stretched-exponential tails:* if $-\log \mathbb{P}(|X| > t) \asymp t^\alpha$ for $\alpha \in (0, 2)$, large deviations of $\hat{\mu}_H$ and of λ_1 (at scale 1) are due to the appearance of large entries [Bordenave–Caputo '14], [Augeri '16].

- *Sparse Bernoulli(p) matrices:* For the adjacency matrix A of Erdős–Rényi graph $G(N, p)$ with $p = o(1)$, deviations of eigenvalues to scale Np, or spectral moments $\mathrm{Tr}\, A^\ell$ at scale $(Np)^\ell$, are due to appearance of "planted" low-rank structures corresponding to dense subgraphs with $\Theta(N^2 p^2)$ edges; or due to high-degree vertices.
  [Bhattacharya–Ganguly–Lubetzky–Zhao '16], [Augeri '18], [C.–Dembo '18], [Bhattacharya–Ganguly '18], [Bhattacharya–Bhattacharya–Ganguly '20], [Basak '21], [C.–Dembo '22]

- *Randomly weighted diluted networks:* $A \odot Y$ with $\{Y_{ij}\}_{i<j}$ iid: [Ganguly–Nam '21], [Ganguly–Hiesmayr–Nam '22], [Augeri–Basak '23].

Conditional structure of sparse Bernoulli matrices on tail events

For $a, b \geq 0$, $\delta \in (0,1)$ let $\mathcal{E}_{a,b}(\delta)$ be the event that



$$\sum_{i,j \in I} A_{i,j} \geq (1-\delta)|I|^2 , \qquad \sum_{i \in J, j \in J^c} A_{i,j} \geq (1-\delta)|J|(N-|J|)$$

for some $I, J \subset [N]$ with $|I| \sim \sqrt{a}pN$, $|J| \sim bp^2 N$.

Theorem (C.–Dembo '22)

For $N^{-1/3} \ll p \ll 1$ and fixed $\ell_1, \ldots, \ell_m \geq 3$, $s_1, \ldots, s_m > 0$,

$$\mathbb{P}\Bigg(\bigcup_{(a_*, b_*) \in \mathcal{O}(\underline{\ell}, \underline{s})} \mathcal{E}_{a,b}(\delta) \,\Bigg|\, \mathrm{Tr}(A^{\ell_k}) \geq (1+s_k)(Np)^{\ell_k} \,,\ k = 1, \ldots, m \Bigg) \geq 1 - p^{cN^2 p^2}$$

for some $c(\underline{\ell}, \underline{s}, \delta) > 0$, where $\mathcal{O}(\underline{\ell}, \underline{s})$ is the set of minimizers for a non-convex linear optimization problem determined by $\underline{\ell}, \underline{s}$.

Special case of a result for any fixed collection of graphs. $\Rightarrow$ Typical structure of Exponential Random Graphs, extending [Chatterjee–Diaconis '12].

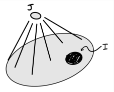
Harel–Mousset–Samotij '18: Case $m = 1$, $\ell_1 = 3$ (and general clique counts).

Conditional structure of sparse Bernoulli matrices on tail events

For $a, b \geq 0$, $\delta \in (0,1)$ let $\mathcal{E}_{a,b}(\delta)$ be the event that

$$\sum_{i,j \in I} A_{i,j} \geq (1-\delta)|I|^2, \qquad \sum_{i \in J, j \in J^c} A_{i,j} \geq (1-\delta)|J|(N - |J|)$$

for some $I, J \subset [N]$ with $|I| \sim \sqrt{a} pN$, $|J| \sim bp^2 N$.



Theorem (C.–Dembo '22)

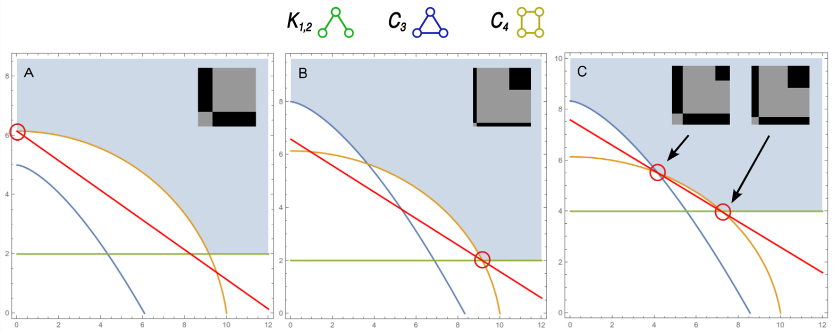
For $N^{-1/3} \ll p \ll 1$ and fixed $\ell_1, \ldots, \ell_m \geq 3$, $s_1, \ldots, s_m > 0$,

$$\mathbb{P}\Bigg(\bigcup_{(a_*, b_*) \in \mathcal{O}(\underline{\ell}, \underline{s})} \mathcal{E}_{a,b}(\delta) \,\Bigg|\, \operatorname{Tr}(A^{\ell_k}) \geq (1 + s_k)(Np)^{\ell_k}\,,\ k = 1, \ldots, m \Bigg) \geq 1 - p^{cN^2 p^2}$$

for some $c(\underline{\ell}, \underline{s}, \delta) > 0$, where $\mathcal{O}(\underline{\ell}, \underline{s})$ is the set of minimizers for a non-convex linear optimization problem determined by $\underline{\ell}, \underline{s}$.

Proof combines quantitative LDPs for Erdős–Rényi (hyper)graphs [C.–Dembo–Pham '20] with stability analysis of the NMF approximation for the upper tail studied in [Bhattacharya–Ganguly–Lubetzky–Zhao '16].

Conditional structure of sparse Bernoulli matrices on tail events



On the 2D manifold of "clique-hub" matrices (up to relabeling rows/columns), level sets of subgraph-counting functions (green/blue/yellow) and relative entropy (red) are ≈smooth curves. Upper tail event is light-blue region, set $\mathcal{O}(F_1, F_2, F_3, s_1, s_2, s_3)$ of optimizers of entropy are circled in red.

Plotted for $s_3 = 100$ and 3 choices of (s_1, s_2): A. $(3, 24)$, B. $(4, 25)$, C. $(4, 31.5)$.

General sub-Gaussian matrices: non-universal LDPs

Back to Wigner matrices, assume law μ of entries is general sub-Gaussian.

Building on strategy of Guionnet–Husson, [Augeri–Guionnet–Husson '19] and [C.–Ducatez–Guionnet '23] show in many cases the existence of $\mathcal{I}_\mu(x)$ such that

$$\frac{1}{N}\log \mathbb{P}(|\lambda_1 - x| \le \delta) = -\mathcal{I}_\mu(x) + o(1). \qquad (*)$$

In particular:

- We always have $\mathcal{I}_\mu(x) \le \mathcal{I}_\gamma(x)$. (Deviations are *at least as likely* as in the sharp sub-Gaussian case.)
- There exists $x_\mu \in (2, \infty]$ such that $\mathcal{I}_\mu(x) = \mathcal{I}_\gamma(x)$ for $x < x_\mu$.
- If $\psi_\mu(t) := \Lambda_\mu(t)/t^2$ is increasing or achieves its maximum at some finite t_* then $(*)$ holds for all $x \in \mathbb{R}$.
- Main technical result of CDG23 gives a non-asymptotic approximation

  $$\frac{1}{N}\log \mathbb{P}(|\lambda_1 - x| \le \delta) \;\approx\; -\mathcal{I}_\mu^{(N)}(x)$$

  in terms of an N-dependent optimization problem over *restricted annealed free energies* for a spiked spherical SK model.

Classical tilting: Cramér LDP

For the sample mean $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ for iid $X_i \sim \mu$, we have

$$\frac{1}{N}\log \mathbb{P}(|\overline{X} - x| \le \delta) = -\Lambda_\mu^*(x) + o(1)$$

where $\Lambda_\mu^*(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda_\mu(\theta)\}$.

Proof: Defining a one-parameter family of tilted measures:

$$\mathbb{P}^{(\theta)}(\cdot) := e^{-N\Lambda_\mu(\theta)}\mathbb{E}e^{\theta N\overline{X}}\,\mathbb{1}(\cdot)\,, \qquad \theta \in \mathbb{R}$$

we can reexpress

$$\begin{aligned}\mathbb{P}(|\overline{X} - x| \le \delta) &= e^{-N(\theta x + o(1))}\,\mathbb{E}e^{\theta N\overline{X}}\,\mathbb{1}(|\overline{X} - x| \le \delta)\\ &= e^{N(\Lambda_\mu(\theta) - \theta x + o(1))}\,\mathbb{P}^{(\theta)}(|\overline{X} - x| \le \delta)\,.\end{aligned}$$

Upper bound: trivially bound $\mathbb{P}^{(\theta)}(|\overline{X} - x| \le \delta) \le 1$ and optimize θ.

Lower bound: show that for the optimizer θ_x, $\mathbb{P}^{(\theta_x)}(|\overline{X} - x| \le \delta) \ge e^{-o(N)}$. $\square$

Tilting by spherical integrals (Guionnet–Husson '18)

For $N \times N$ symmetric M, $\theta \geq 0$ and P the uniform surface measure on $\mathbb{S}^{N-1}$,

$$I(M,\theta) := \int_{\mathbb{S}^{N-1}} e^{\theta N \langle u, Mu \rangle} dP(u)$$

Quenched free energy. Guionnet–Maida '05: on $\{\lambda_1 \sim x\} \cap \{\hat{\mu}_H \sim \sigma\}$,

$$\frac{1}{N} \log I(H,\theta) \sim J(x,\theta) := \begin{cases} \theta^2 & \theta \leq \frac{1}{2} G_\sigma(x) \\ \theta x - \frac{1}{2} \int \log(x-\lambda) d\sigma(\lambda) - \frac{1}{2} \log(2e\theta) & \theta \geq \frac{1}{2} G_\sigma(x) \end{cases}$$

where $G_\sigma(x) = \frac{1}{2}(x - \sqrt{x^2-4})$ is the Stieltjes transform of σ at $x \geq 2$.

Annealed free energy. $F_N(\theta) := \frac{1}{N} \log \mathbb{E} I(H,\theta)$. By Fubini,

$$\mathbb{E} I(H,\theta) = \int_{\mathbb{S}^{N-1}} \mathbb{E} e^{\theta N \langle u, Hu \rangle} dP(u) = \int_{\mathbb{S}^{N-1}} \prod_{i \leq j} \mathbb{E} e^{2\theta \sqrt{N} X_{ij} u_i u_j} dP(u)$$

$$\implies \quad F_N(\theta) = \frac{1}{N} \log \int_{\mathbb{S}^{N-1}} \exp\Big(\sum_{i \leq j} \Lambda_\mu (2\theta \sqrt{N} u_i u_j) \Big) dP(u) \,.$$

For μ SSG, show: **(A)** $\frac{1}{N} \log \mathbb{P}(\lambda_1 \sim x) \sim \inf_{\theta \geq 0} \{F_N(\theta) - J(x,\theta)\}$,
and **(B)** $F_N(\theta) \to \theta^2$. (Independent of μ!)

A computation gives $\inf_{\theta \geq 0} \{\theta^2 - J(x,\theta)\} = -\mathcal{I}_\gamma(x)$. $\square$

Tilting by spherical integrals: Annealed – Quenched

Argument for (**A**) $\frac{1}{N}\log \mathbb{P}(\lambda_1 \sim x) \sim \inf_{\theta\geq 0}\{F_N(\theta) - J(x,\theta)\}$.

Use *two levels of tilting:* For $\theta \geq 0, u \in \mathbb{S}^{N-1}$, define measures

$$\mathbb{P}^{(\theta,u)}(A) := \frac{\mathbb{E}e^{\theta N\langle u,Hu\rangle}\,\mathbb{1}(A)}{\mathbb{E}e^{\theta N\langle u,Hu\rangle}}, \qquad \text{on } (\Omega,\mathcal{F}),$$

$$Q^{(\theta)}(B) := \frac{\int_B \mathbb{E}e^{\theta N\langle u,Hu\rangle}dP(u)}{\int_{\mathbb{S}^{N-1}}\mathbb{E}e^{\theta N\langle u,Hu\rangle}dP(u)} = e^{-NF_N(\theta)}\int_B \mathbb{E}e^{\theta N\langle u,Hu\rangle}dP(u) \qquad \text{on } \mathbb{S}^{N-1}.$$

With $\mathcal{E}_x := \{\lambda_1 \sim x, \hat{\mu}_H \sim \sigma\}$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_x) &= e^{-N(J(x,\theta)+o(1))}\mathbb{E}I(H,\theta)\,\mathbb{1}(\mathcal{E}_x) \\
&= e^{-N(J(x,\theta)+o(1))}\int_{\mathbb{S}^{N-1}}\mathbb{E}e^{\theta N\langle u,Hu\rangle}\,\mathbb{1}(\mathcal{E}_x)dP(u) \\
&= e^{-N(J(x,\theta)+o(1))}\int_{\mathbb{S}^{N-1}}\mathbb{P}^{(\theta,u)}(\mathcal{E}_x)\,\mathbb{E}e^{\theta N\langle u,Hu\rangle}dP(u) \\
&= e^{N(F_N(\theta)-J(x,\theta)+o(1))}\int_{\mathbb{S}^{N-1}}\mathbb{P}^{(\theta,u)}(\mathcal{E}_x)\,dQ^{(\theta)}(u).
\end{aligned}$$

<u>Upper bound:</u> trivially bound $\mathbb{P}^{(\theta,u)}(\mathcal{E}_x) \leq 1$ and optimize θ.

Tilting by spherical integrals: Annealed – Quenched

Lower bound:

Showed $\quad \mathbb{P}(\mathcal{E}_x) = e^{N(F_N(\theta)-J(x,\theta))+o(N)} \int_{\mathbb{S}^{N-1}} \mathbb{P}^{(\theta,u)}(\mathcal{E}_x) dQ^{(\theta)}(u)$

where

$$\mathcal{E}_x = \{\lambda_1 \sim x, \hat{\mu}_H \sim \sigma\}, \qquad \frac{d\mathbb{P}^{(\theta,u)}}{d\mathbb{P}} \propto e^{\theta N \langle u, Hu \rangle}, \qquad \frac{dQ^{(\theta)}}{dP}(u) \propto \mathbb{E} e^{\theta N \langle u, Hu \rangle}.$$

With θ_x the optimizing choice of θ from the upper bound, only remains to show $\{\lambda_1 \sim x\}$ is likely under $\mathbb{P}^{(\theta_x,u)}$, at least for all u in some $\mathcal{D} \subset \mathbb{S}^{N-1}$ such that $Q^{(\theta_x)}(\mathcal{D}) \geq e^{-o(N)}$.

Take $\mathcal{D} = \{u \in \mathbb{S}^{N-1} : \|u\|_\infty \leq N^{-\frac{1}{4}-\varepsilon}\}$ set of delocalized unit vectors. Then

(1) $Q^{(\theta)}(\mathcal{D}) \geq e^{-o(N)}$ (easy).
(2) For any $u \in \mathcal{D}$, $H \overset{d}{\approx} 2\theta u u^\mathsf{T} + \widetilde{H}$ under $\mathbb{P}^{(\theta,u)}$ for a Wigner matrix $\widetilde{H}$.

(Note that $\mathbb{E}^{(\theta,u)} H_{ij} = \frac{1}{\sqrt{N}} \Lambda'_\mu(2\theta\sqrt{N} u_i u_j) \sim 2\theta u_i u_j$.)

By the BBP transition we get $\lambda_1 \sim x$ w.h.p. under $\mathbb{P}^{(\theta_x,u)}$ for any $u \in \mathcal{D}$. $\square$

Tilting by spherical integrals: Universal annealed free energy

Now to show (**B**) $F_N(\theta) := \frac{1}{N} \log \mathbb{E} I(H, \theta) \to \theta^2$,

$$F_N(\theta) = \frac{1}{N} \log \int_{\mathbb{S}^{N-1}} \exp\Big(\sum_{i \le j} \Lambda_\mu(2\theta\sqrt{N} u_i u_j) \Big) dP(u)$$
$$\sim \frac{1}{N} \log \int_{\mathcal{D}} \exp\Big(\sum_{i \le j} \Lambda_\mu(2\theta\sqrt{N} u_i u_j) \Big) dP(u).$$

For u delocalized we can use the Taylor expansion $\Lambda_\mu(t) \sim \frac{1}{2} t^2$ for $t = o(1)$:

$$\Lambda_\mu(2\theta\sqrt{N} u_i u_j) \sim \sum_{i \le j} 2\theta^2 N u_i^2 u_j^2 \sim \theta^2 N.$$

Universality comes from expansion of Λ_μ near 0 (as for the CLT!).

New ideas to capture localization

For μ SSG, showed: **(A)** $\frac{1}{N}\log\mathbb{P}(\lambda_1\sim x)\sim\sup_{\theta\geq 0}\{F_N(\theta)-J(x,\theta)\}$,
(B) $F_N(\theta)\to\theta^2$.

Proof suggests that on $\{\lambda_1\sim x\}$, $H\overset{d}{\approx}2\theta uu^{\mathsf{T}}+\widetilde{H}$ for a random delocalized u.

When μ is not SSG, both (A) and (B) are false.

- Upper bound $\frac{1}{N}\log\mathbb{P}(\lambda_1\sim x)\leq F_N(\theta)-J(x,\theta)+o(1)$ still true, but not always sharp [AGH19].
- No longer true that $F_N(\theta)\sim\theta^2$. Contribution of $u\in\mathbb{S}^{N-1}$ with large entries carries too much weight – $\mathcal{D}$ no longer typical under $Q^{(\theta)}$.
- For u not delocalized, no longer true that $H\overset{d}{\approx}2\theta uu^{\mathsf{T}}+\widetilde{H}$ under $\mathbb{P}^{(\theta,u)}$. Can't do a BBP computation!

What is happening? Heavier tails open up non-universal localization strategies that compete with delocalized tilt.

In fact, large deviations of λ_1 result from a combination of the two!

Localization will be reflected by large entries of the associated eigenvector v_1. So we do a spherical integral tilt with a fixed choice of large entries, then optimize them at the end (there are $o(N)$ of them).

New ideas to capture localization

Let $\mathcal{L}_\eta$ be the set of $N^{1-2\eta}$-sparse vectors in the ball $\mathbb{B}^N$. For $z \in \mathcal{L}_\eta$ we let

$$\mathcal{U}_z := \left\{ u \in \mathbb{S}^{N-1} : u^{large} \approx z,\ \| u|_{\mathrm{supp}(z)^c} \|_\infty \leq N^{\eta - 1/2} \right\}$$

and denote the *restricted annealed free energy*

$$F_N(\theta; z) := \frac{1}{N} \log \mathbb{E} \int_{\mathcal{U}_z} e^{\theta N \langle u, Hu \rangle} dP(u).$$

Theorem (C.–Ducatez–Guionnet '23)

(With technical conditions) If η is a sufficiently small constant, for any $x \geq 2$,

$$\frac{1}{N} \log \mathbb{P}(|\lambda_1 - x| \leq \delta) = \sup_{w \in \mathcal{L}_\eta} \inf_{\theta \geq 0} \left\{ F_N\big(\theta; q_x(\theta) w\big) - J(x, \theta) \right\} + o(1)$$

where $q_x(\theta) := (1 - \frac{G_\sigma(x)}{2\theta})_+^{1/2}$.

Note $\mathcal{U}_0 = \mathcal{D}$ so $F_N(\theta; 0) \sim \theta^2$, and RHS is bounded below by $-\mathcal{I}_\gamma(x) + o(1)$. For $x < x_\mu$ the supremum is in fact attained at $w = 0$, giving $\mathcal{I}_\mu(x) = \mathcal{I}_\gamma(x)$.

New ideas to capture localization

Recall

$$F_N(\theta; z) = \frac{1}{N} \log \mathbb{E} \int_{\mathcal{U}_z} e^{\theta N \langle u, Hu \rangle} dP(u) = \frac{1}{N} \log \int_{\mathcal{U}_z} e^{\sum_{i \le j} \Lambda_\mu(2\theta\sqrt{N} u_i u_j)} dP(u).$$

Theorem (C.–Ducatez–Guionnet '23)

For any $z \in \mathcal{L}_\eta$, and $\theta \ge 0$,

$$F_N(\theta; z) = \varphi^{del}(\theta, z) + \varphi_N^{loc}(\theta, z) + \varphi_N^{cross}(\theta, z) + O_\theta(N^{-\eta/2})$$

where

$$\varphi^{del}(\theta, z) := \theta^2 (1 - \|z\|_2^2)^2, \qquad \varphi_N^{loc}(\theta, z) := \frac{1}{N} \sum_{i \le j} \Lambda_\mu(\theta \sqrt{N} z_i z_j)$$

$$\varphi_N^{cross}(\theta, z) := \sup_{\substack{\nu \in \mathcal{P}([-N^\eta, N^\eta]) : \\ \int s^2 d\nu(s) = 1 - \|z\|_2^2}} \left\{ \int \sum_{i=1}^N \Lambda_\mu(2\theta z_i s) d\nu(s) - \mathrm{H}(\nu|\gamma) \right\} - \frac{1}{2} \|z\|_2^2 .$$

Case of ψ_μ increasing

For e.g. sparse Gaussian entries, $\psi_\mu(t) = \Lambda_\mu(t)/t^2$ is increasing as $|t| \to \infty$. In this case we can show

$$F_N(\theta; z) \leq F_N(\theta; \|z\|_2 e_1) + o(1)$$

for any $z \in \mathbb{B}^N$. Then we obtain

Theorem (C.–Ducatez–Guionnet '23)

Assume $\psi_\mu(t)$ is increasing as $|t| \to \infty$. Then λ_1 satisfies a full large deviation principle with speed N and good rate function $\mathcal{I}_\mu$ which is infinite on $(-\infty, 2)$ and is otherwise given by

$$\mathcal{I}_\mu(x) = \inf_{\alpha \in [0,1]} \sup_{\theta \geq 0} \Big\{ J(x,\theta) + \frac{1}{2}\alpha q_x(\theta)^2 - \psi_\infty \alpha^2 \theta^2 q_x(\theta)^4 - \theta^2 (1 - \alpha q_x(\theta)^2)^2 \\ - \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}) \\ \int x^2 d\nu(x) = 1-\alpha}} \Big\{ \int \Lambda_\mu(2\alpha\theta q_x(\theta)^2 s) d\nu(s) - \mathrm{H}(\nu|\gamma) \Big\} \Big\}.$$

Moreover, on $\{\lambda_1 \sim x\}$, v_1 has a coordinate of size $\Omega(1)$.

New ideas to capture localization

With $\mathcal{E}_{x,w} = \{\lambda_1 \sim x, v_1^{large} \sim w\}$, we can show

$$\mathbb{P}(\mathcal{E}_{x,w}) = e^{N(F_N(\theta;q_x(\theta)w) - J(x,\theta) + o(1))} \int_{\mathcal{U}_{q_x(\theta)w}} \mathbb{P}^{(\theta,u)}(\lambda_1 \sim x)\, dQ^{(\theta,q_x(\theta)w)}(u)$$

where $\frac{d\mathbb{P}^{(\theta,u)}}{d\mathbb{P}} \propto e^{\theta N \langle u, Hu \rangle}$ as before, and we take $\frac{dQ^{(\theta,qw)}}{dP} \propto 1_{\mathcal{U}_{qw}} \mathbb{E} e^{\theta N \langle u, Hu \rangle}$.

Upper bound: trivially bound the integral by 1, optimize θ, worst case w.

Lower bound: task is to show the integral is $\geq e^{-o(N)}$ for some $\theta = \theta_{x,w}$.

Problem: As $u \in \mathcal{U}_{q_x(\theta)w}$ are not delocalized, we can't compute $\mathbb{E}^{(\theta,u)}\lambda_1$ by a BBP computation as before.

Solution: Can show λ_1 concentrates under $\mathbb{P}^{(\theta,u)}$, with mean $\approx$ continuous in θ and u (under the ℓ^2 metric). $\mathbb{E}^{(\theta,u)}\lambda_1 \sim 2$ for small θ, $\mathbb{E}^{(\theta,u)}\lambda_1 \to \infty$ as $\theta \to \infty$.

Moreover, we can show the measures $Q^{(\theta,qw)}$ concentrate on a small ball in the 2-Wasserstein metric, with center $v_{\theta,w} \in \mathbb{S}^{N-1}$ that varies continuously with θ.

Intermediate Value Theorem yields $\theta = \theta_{x,w}$ such that $\mathbb{E}^{(\theta,v_{\theta,w})}\lambda_1 \sim x$. $\square$

Thanks for your attention!