

# Average-case matrix discrepancy, large deviations, and the HCIZ integral

*Antoine Maillard*

*Inria*

arXiv:2410.17887



Work in progress



w. J-C Mourrat (ENS Lyon)

*Cargèse – August 2025*

## **Part I : Average-case matrix discrepancy**

## Discrepancy: balancing vectors in high dimension

---

# Discrepancy: balancing vectors in high dimension

---

**“Balancing vectors”:** Given  $u_1, \dots, u_n \in \mathbb{R}^d$ , find  $\varepsilon_1, \dots, \varepsilon_n = \pm 1$  to make the “balancing”  $\sum_{i=1}^n \varepsilon_i u_i$  small

$$\text{Discrepancy } \text{disc}(u_1, \dots, u_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_{\infty}$$

# Discrepancy: balancing vectors in high dimension

“Balancing vectors”: Given  $u_1, \dots, u_n \in \mathbb{R}^d$ , find  $\varepsilon_1, \dots, \varepsilon_n = \pm 1$  to make the “balancing”  $\sum_{i=1}^n \varepsilon_i u_i$  small

$$\text{Discrepancy } \text{disc}(u_1, \dots, u_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_{\infty}$$

## Discrepancy theory

For large  $n, d \gg 1$ , and some assumptions on  $\{u_i\}_{i=1}^n$ , can we **compute/upper bound**  $\text{disc}(u_1, \dots, u_n)$



*Applications & motivations:* Combinatorics, perceptron-type problems, computational geometry, experimental design, randomized clinical trials, theory of approximation algorithms, ...

Matousek '09 ; Chen&al '14 ;  
talks of Dan Spielman,...

# Discrepancy: balancing vectors in high dimension

“Balancing vectors”: Given  $u_1, \dots, u_n \in \mathbb{R}^d$ , find  $\varepsilon_1, \dots, \varepsilon_n = \pm 1$  to make the “balancing”  $\sum_{i=1}^n \varepsilon_i u_i$  small

$$\text{Discrepancy } \text{disc}(u_1, \dots, u_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_{\infty}$$

## Discrepancy theory

For large  $n, d \gg 1$ , and some assumptions on  $\{u_i\}_{i=1}^n$ , can we **compute/upper bound**  $\text{disc}(u_1, \dots, u_n)$



*Applications & motivations:* Combinatorics, perceptron-type problems, computational geometry, experimental design, randomized clinical trials, theory of approximation algorithms, ...

Matousek '09 ; Chen&al '14 ;  
talks of Dan Spielman,...

## General discrepancy problem

$$\text{disc}(u_1, \dots, u_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_{\infty}$$



$$\text{disc}(x_1, \dots, x_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|$$

$x_i \in (E, \|\cdot\|)$



# Matrix discrepancy

---

$$\min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|$$

**Matrix setting** of the general discrepancy problem

# Matrix discrepancy

2

$$\min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|$$

**Matrix setting** of the general discrepancy problem

$$x_i = \mathbf{A}_i \in \mathbb{R}^{d \times d} \quad (\mathbf{A}_i = \mathbf{A}_i^\top)$$

$$\|\cdot\| = \|\cdot\|_{\text{op}} := \max\{|\lambda_i(\mathbf{A})|\}$$

**Matrix discrepancy**

$$\text{disc}(\mathbf{A}_1, \dots, \mathbf{A}_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i \mathbf{A}_i \right\|_{\text{op}}$$

Applications and connections to matrix concentration inequalities, quantum random access codes, graph sparsification, ...

Hopkins&al '22; Bansal&al '23; Batson&al '14; Cai&al '25...



# Matrix discrepancy

$$\min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|$$

**Matrix setting** of the general discrepancy problem

$$x_i = \mathbf{A}_i \in \mathbb{R}^{d \times d} \quad (\mathbf{A}_i = \mathbf{A}_i^\top)$$

$$\|\cdot\| = \|\cdot\|_{\text{op}} := \max\{|\lambda_i(\mathbf{A})|\}$$

**Matrix discrepancy**

$$\text{disc}(\mathbf{A}_1, \dots, \mathbf{A}_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i \mathbf{A}_i \right\|_{\text{op}}$$

Applications and connections to matrix concentration inequalities, quantum random access codes, graph sparsification, ...

Hopkins&al '22; Bansal&al '23; Batson&al '14; Cai&al '25...

- Special case:  $\mathbf{A}_i = \text{Diag}(u_i)$



**Vector discrepancy** with  $\|\cdot\|_\infty$

$$\min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_\infty$$

# Matrix discrepancy

$$\min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|$$

**Matrix setting** of the general discrepancy problem

$$x_i = \mathbf{A}_i \in \mathbb{R}^{d \times d} \quad (\mathbf{A}_i = \mathbf{A}_i^\top)$$

$$\|\cdot\| = \|\cdot\|_{\text{op}} := \max\{|\lambda_i(\mathbf{A})|\}$$

**Matrix discrepancy**

$$\text{disc}(\mathbf{A}_1, \dots, \mathbf{A}_n) := \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i \mathbf{A}_i \right\|_{\text{op}}$$

Applications and connections to matrix concentration inequalities, quantum random access codes, graph sparsification, ...

Hopkins&al '22; Bansal&al '23; Batson&al '14; Cai&al '25...

- Special case:  $\mathbf{A}_i = \text{Diag}(u_i)$



**Vector discrepancy** with  $\|\cdot\|_\infty$

$$\min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_\infty$$

**"Matrix Spencer" conjecture**

- Many open questions, e.g.

$$\max_{\|\mathbf{A}_i\|_{\text{op}} \leq 1} \text{disc}(\mathbf{A}_1, \dots, \mathbf{A}_n) \lesssim \sqrt{n}$$



Proven for vector discrepancy by  
J. Spencer in 1985

Zouzias '12; Meka'14 ; Bandeira&al  
'23 ; Bansal&al '23 ; ...

## Average-case matrix discrepancy

---

What about **random matrices** ?

# Average-case matrix discrepancy

$W_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$  for  $i \leq j$

What about **random matrices** ?

Margin  $\kappa > 0$

Given  $\mathbf{W}_1, \dots, \mathbf{W}_n \stackrel{\text{i.i.d.}}{\sim} \text{GOE}(d)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that

$$\left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n}$$



[Kunisky & Zhang '23 ; Wengiel '24]

# Average-case matrix discrepancy

$W_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$  for  $i \leq j$

What about **random matrices** ?

Margin  $\kappa > 0$

Given  $\mathbf{W}_1, \dots, \mathbf{W}_n \stackrel{\text{i.i.d.}}{\sim} \text{GOE}(d)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that

$$\left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n}$$



[Kunisky & Zhang '23 ; Wengiel '24]

Matrix analog of...

$$\mathbf{W}_i = \text{Diag}(w_i)$$

$$(g_i)_j := (w_j)_i$$

## Symmetric Binary Perceptron

Given  $g_1, \dots, g_d \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that  $\max_{i \in [d]} |\langle g_i, \varepsilon \rangle| \leq \kappa \sqrt{n}$

Aubin&al '19 ; Abbe&al '22 ; Gamarnik&al '22 ; ...

# Average-case matrix discrepancy

$W_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$  for  $i \leq j$

What about **random matrices** ?

Margin  $\kappa > 0$

Given  $\mathbf{W}_1, \dots, \mathbf{W}_n \stackrel{\text{i.i.d.}}{\sim} \text{GOE}(d)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that

$$\left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n}$$



[Kunisky & Zhang '23 ; Wengiel '24]

Matrix analog of...

$$\mathbf{W}_i = \text{Diag}(w_i)$$

$$(g_i)_j := (w_j)_i$$

## Symmetric Binary Perceptron

Given  $g_1, \dots, g_d \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that  $\max_{i \in [d]} |\langle g_i, \varepsilon \rangle| \leq \kappa \sqrt{n}$

Aubin&al '19 ; Abbe&al '22 ; Gamarnik&al '22 ; ...

## Goals

- Sharp **satisfiability transitions** (in the regime  $n = \Theta(d^2)$ ) ?
- Structure of solution space ?
- Polynomial-time algorithms ?
- More complex models of  $\mathbf{W}_i$  ?

[Kunisky & Zhang '23]



# Average-case matrix discrepancy

$W_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$  for  $i \leq j$

What about **random matrices** ?

Margin  $\kappa > 0$

Given  $\mathbf{W}_1, \dots, \mathbf{W}_n \stackrel{\text{i.i.d.}}{\sim} \text{GOE}(d)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that

$$\left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n}$$



[Kunisky & Zhang '23 ; Wengiel '24]

Matrix analog of...

$$\mathbf{W}_i = \text{Diag}(w_i)$$

$$(g_i)_j := (w_j)_i$$

## Symmetric Binary Perceptron

Given  $g_1, \dots, g_d \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ , can we find  $\varepsilon \in \{\pm 1\}^n$  such that  $\max_{i \in [d]} |\langle g_i, \varepsilon \rangle| \leq \kappa \sqrt{n}$

Aubin&al '19 ; Abbe&al '22 ; Gamarnik&al '22 ; ...

**This talk** ➤ Sharp **satisfiability transitions** (in the regime  $n = \Theta(d^2)$ ) ?

**Goals**

- Structure of solution space ?
- Polynomial-time algorithms ?
- More complex models of  $\mathbf{W}_i$  ?

[Kunisky & Zhang '23]



## Results I: first moment asymptotics

---



$$n/d^2 \rightarrow \tau > 0$$



# Results I: first moment asymptotics

---



$$n/d^2 \rightarrow \tau > 0$$

Number of solutions / Partition function

$$Z_{\kappa} := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{w}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

# Results I: first moment asymptotics



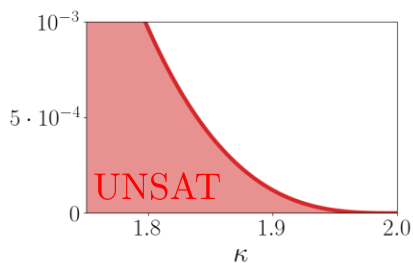
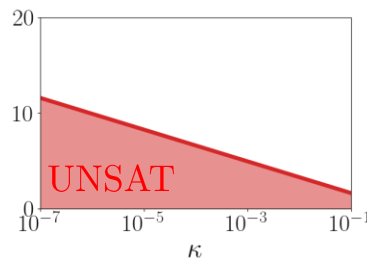
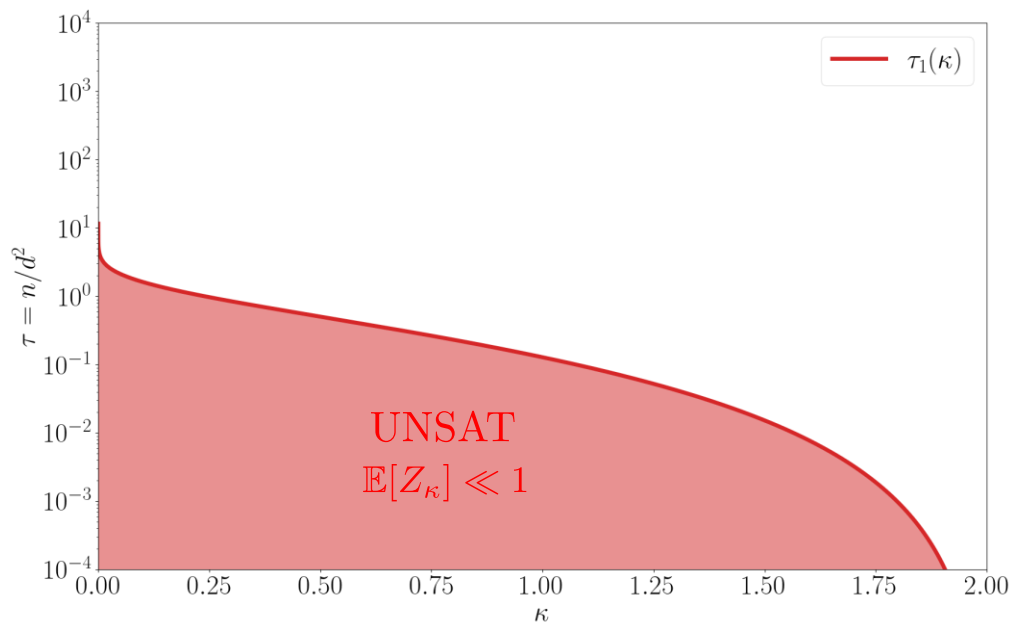
$$n/d^2 \rightarrow \tau > 0$$

Number of solutions / Partition function

$$Z_\kappa := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{w}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

**Theorem:**  $\lim_{d \rightarrow \infty} \frac{1}{d^2} \log \mathbb{E} Z_\kappa = (\tau - \tau_1(\kappa)) \log 2$

$$\tau_1(\kappa) := \frac{1}{\log 2} \left[ -\frac{\kappa^4}{128} + \frac{\kappa^2}{8} - \frac{1}{2} \log \frac{\kappa}{2} - \frac{3}{8} \right]$$

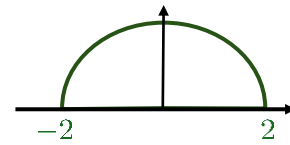


$$\tau < \tau_1(\kappa)$$



$$\mathbb{P}[Z_\kappa = 0] = 1 - o(1)$$

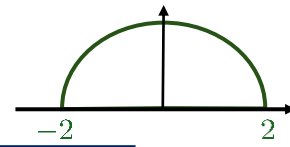
## First moment computation: large deviations



$\mathbf{W} \sim \text{GOE}(d)$

$$\mathbb{E}Z_\kappa = \sum_{\varepsilon \in \{\pm 1\}^n} \mathbb{P} \left[ \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right] = 2^n \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa]$$

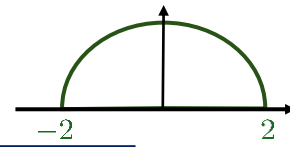
# First moment computation: large deviations



$\mathbf{W} \sim \text{GOE}(d)$

$$\mathbb{E} Z_{\kappa} = \sum_{\varepsilon \in \{\pm 1\}^n} \mathbb{P} \left[ \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right] = 2^n \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \rightarrow \text{Left } (\kappa < 2) \text{ large deviations of } \|\mathbf{W}\|_{\text{op}}$$

# First moment computation: large deviations



$\mathbf{W} \sim \text{GOE}(d)$

$$\mathbb{E} Z_\kappa = \sum_{\varepsilon \in \{\pm 1\}^n} \mathbb{P} \left[ \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right] = 2^n \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \rightarrow \text{Left } (\kappa < 2) \text{ large deviations of } \|\mathbf{W}\|_{\text{op}}$$

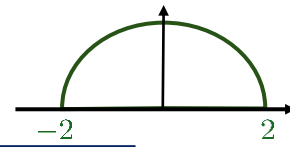
➤ **Idea:** The event  $\|\mathbf{W}\|_{\text{op}} \leq \kappa$  is driven by

**Large deviations of the spectral density**

$$\mathbb{P}[\mu_{\mathbf{W}} \simeq \mu] \simeq \exp\{-d^2 I(\mu)\} \quad \begin{array}{l} \text{Ben Arous \& Guionnet '97;} \\ \text{Dean \& Majumdar '06 '08;} \end{array}$$

$$I(\mu) := -\frac{1}{2} \int \mu(dx) \mu(dy) \log |x - y| + \frac{1}{4} \int \mu(dx) x^2 - \frac{3}{8}$$

# First moment computation: large deviations



$\mathbf{W} \sim \text{GOE}(d)$

$$\mathbb{E} Z_\kappa = \sum_{\varepsilon \in \{\pm 1\}^n} \mathbb{P} \left[ \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right] = 2^n \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \rightarrow \text{Left } (\kappa < 2) \text{ large deviations of } \|\mathbf{W}\|_{\text{op}}$$

➤ **Idea:** The event  $\|\mathbf{W}\|_{\text{op}} \leq \kappa$  is driven by

Large deviations of the spectral density

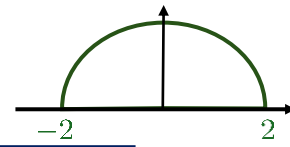
$$\mathbb{P}[\mu_{\mathbf{W}} \simeq \mu] \simeq \exp\{-d^2 I(\mu)\} \quad \begin{array}{l} \text{Ben Arous \& Guionnet '97;} \\ \text{Dean \& Majumdar '06 '08;} \end{array}$$

$$I(\mu) := -\frac{1}{2} \int \mu(\mathrm{d}x) \mu(\mathrm{d}y) \log |x - y| + \frac{1}{4} \int \mu(\mathrm{d}x) x^2 - \frac{3}{8}$$



$$\lim \frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] = - \inf_{\mu \in \mathcal{M}([- \kappa, \kappa])} I(\mu)$$

# First moment computation: large deviations



$\mathbf{W} \sim \text{GOE}(d)$

$$\mathbb{E} Z_\kappa = \sum_{\varepsilon \in \{\pm 1\}^n} \mathbb{P} \left[ \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right] = 2^n \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \rightarrow \text{Left } (\kappa < 2) \text{ large deviations of } \|\mathbf{W}\|_{\text{op}}$$

➤ **Idea:** The event  $\|\mathbf{W}\|_{\text{op}} \leq \kappa$  is driven by

Large deviations of the spectral density

$$\mathbb{P}[\mu_{\mathbf{W}} \simeq \mu] \simeq \exp\{-d^2 I(\mu)\} \quad \text{Ben Arous \& Guionnet '97; Dean\&Majumdar '06 '08;}$$

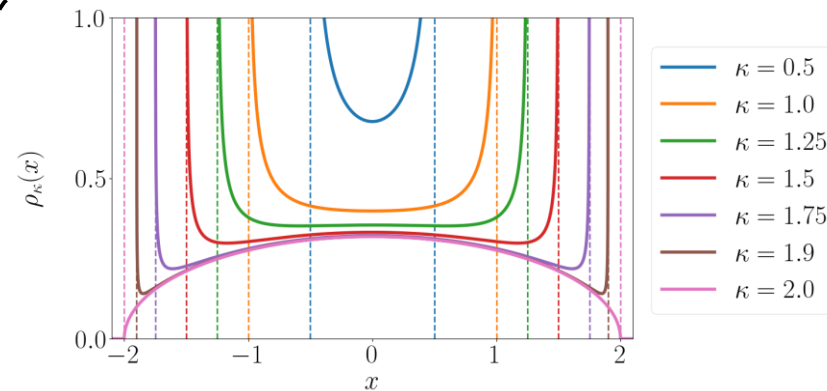


$$\lim \frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] = - \inf_{\mu \in \mathcal{M}([- \kappa, \kappa])} I(\mu)$$

➤ Compute  $\inf_{\mu \in \mathcal{M}([- \kappa, \kappa])} I(\mu)$  from **Tricomi's theorem** ■

Tricomi' 85; Dean\&Majumdar '06 '08; Vivo\&al '07, ...

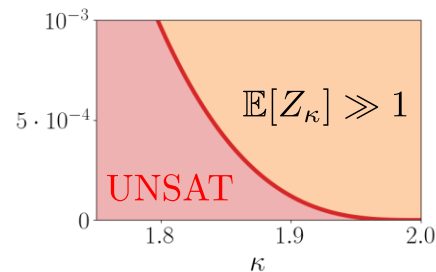
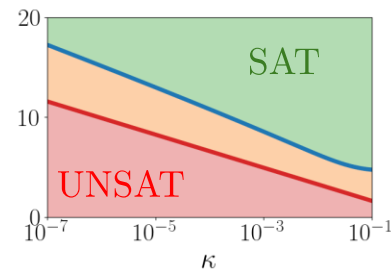
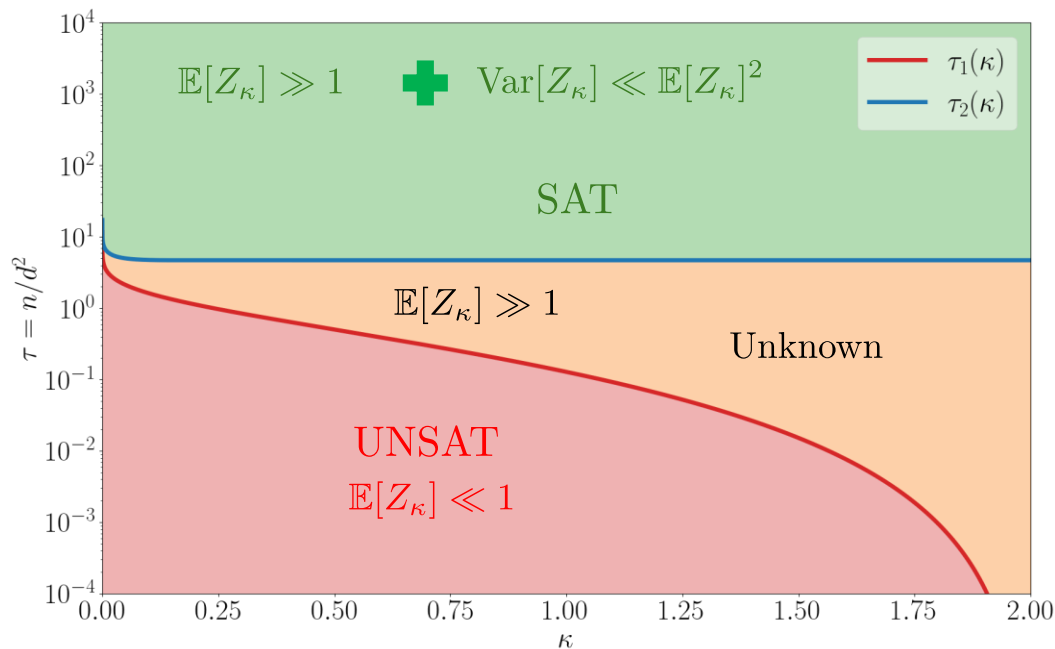
$$I(\mu) := -\frac{1}{2} \int \mu(dx) \mu(dy) \log |x - y| + \frac{1}{4} \int \mu(dx) x^2 - \frac{3}{8}$$



$$\rho_\kappa(x) = \frac{4 + \kappa^2 - 2x^2}{4\pi\sqrt{\kappa^2 - x^2}} = \arg \min_{\mu \in \mathcal{M}([- \kappa, \kappa])} I(\mu)$$

# Results II: Upper bounds via the second moment method

$$n/d^2 \rightarrow \tau > 0$$



$$Z_\kappa := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

## Theorem I

$$\tau < \tau_1(\kappa)$$



$$\mathbb{P}[Z_\kappa = 0] = 1 - o(1)$$



## Theorem II

$$\tau > \tau_2(\kappa)$$



$$\mathbb{P}[Z_\kappa \geq 1] = 1 - o(1)$$




## Second moment computation: sketch

---

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

## Second moment computation: sketch


$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

- 1<sup>st</sup> moment:**  $\mathbb{E}[Z_\kappa] \iff F(\kappa) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \iff I(\mu) := -\frac{1}{d^2} \log \mathbb{P}[\mu(\{\mathbf{W}\}) \simeq \mu]$  

↗  $\mathbf{W} \sim \text{GOE}(d)$

## Second moment computation: sketch

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$


- 1<sup>st</sup> moment:**  $\mathbb{E}[Z_\kappa] \iff F(\kappa) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \iff I(\mu) := -\frac{1}{d^2} \log \mathbb{P}[\mu(\{\mathbf{W}\}) \simeq \mu]$  

$\nearrow \mathbf{W} \sim \text{GOE}(d)$
- 2<sup>nd</sup> moment:**  $\mathbb{E}[Z_\kappa^2] \iff G(\kappa, q) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa \text{ and } \|q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\|_{\text{op}} \leq \kappa]$  ( $\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$ )

$q \in [0, 1]$  “overlap”

## Second moment computation: sketch

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

- 1<sup>st</sup> moment:**  $\mathbb{E}[Z_\kappa] \iff F(\kappa) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \iff I(\mu) := -\frac{1}{d^2} \log \mathbb{P}[\mu(\{\mathbf{W}\}) \simeq \mu]$  

$\mathbf{W} \sim \text{GOE}(d)$
- 2<sup>nd</sup> moment:**  $\mathbb{E}[Z_\kappa^2] \iff G(\kappa, q) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa \text{ and } \|q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\|_{\text{op}} \leq \kappa]$  ( $\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$ )


$q \in [0, 1]$  “overlap”

$\iff \Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices.

## Second moment computation: sketch

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

- **1<sup>st</sup> moment:**  $\mathbb{E}[Z_\kappa] \iff F(\kappa) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \iff I(\mu) := -\frac{1}{d^2} \log \mathbb{P}[\mu(\{\mathbf{W}\}) \simeq \mu]$    $\mathbf{W} \sim \text{GOE}(d)$
- **2<sup>nd</sup> moment:**  $\mathbb{E}[Z_\kappa^2] \iff G(\kappa, q) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa \text{ and } \|q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\|_{\text{op}} \leq \kappa]$   $(\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d))$   
 $q \in [0, 1]$  “overlap”

$$\iff \Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices.

- Our proof gives an **upper bound** on  $G(\kappa, q)$ , using tailored concentration inequalities.



$$\frac{\mathbb{E}[Z_\kappa^2]}{\mathbb{E}[Z_\kappa]^2} \lesssim \left[ 1 - \frac{\tau_2(\kappa)}{\tau} \right]^{-1/2}$$

Not tight in general




**Sharpness of the transition**

$$\implies \mathbb{P}[Z_\kappa \geq 1] = 1 - o(1) \text{ if } \tau > \tau_2(\kappa) \quad \blacksquare$$

Refinements of Gaussian Poincaré inequality [Altschuler '23]

## Second moment computation: sketch

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

- **1<sup>st</sup> moment:**  $\mathbb{E}[Z_\kappa] \iff F(\kappa) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa] \iff I(\mu) := -\frac{1}{d^2} \log \mathbb{P}[\mu(\{\mathbf{W}\}) \simeq \mu]$    $\mathbf{W} \sim \text{GOE}(d)$
- **2<sup>nd</sup> moment:**  $\mathbb{E}[Z_\kappa^2] \iff G(\kappa, q) := -\frac{1}{d^2} \log \mathbb{P}[\|\mathbf{W}\|_{\text{op}} \leq \kappa \text{ and } \|q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\|_{\text{op}} \leq \kappa]$   $(\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d))$   
 $q \in [0, 1]$  “overlap”

$$\iff \Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P}[\mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2]$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices.

- Our proof gives an **upper bound** on  $G(\kappa, q)$ , using tailored concentration inequalities.



$$\frac{\mathbb{E}[Z_\kappa^2]}{\mathbb{E}[Z_\kappa]^2} \lesssim \left[1 - \frac{\tau_2(\kappa)}{\tau}\right]^{-1/2}$$

Not tight in general



**Sharpness of the transition**

$$\implies \mathbb{P}[Z_\kappa \geq 1] = 1 - o(1) \text{ if } \tau > \tau_2(\kappa) \quad \blacksquare$$

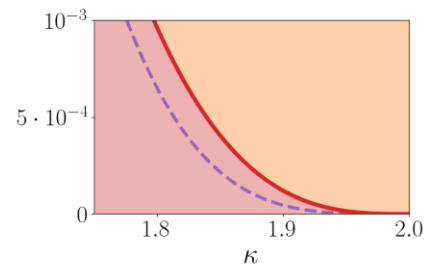
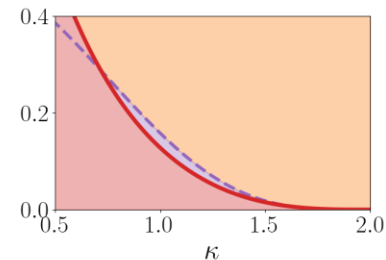
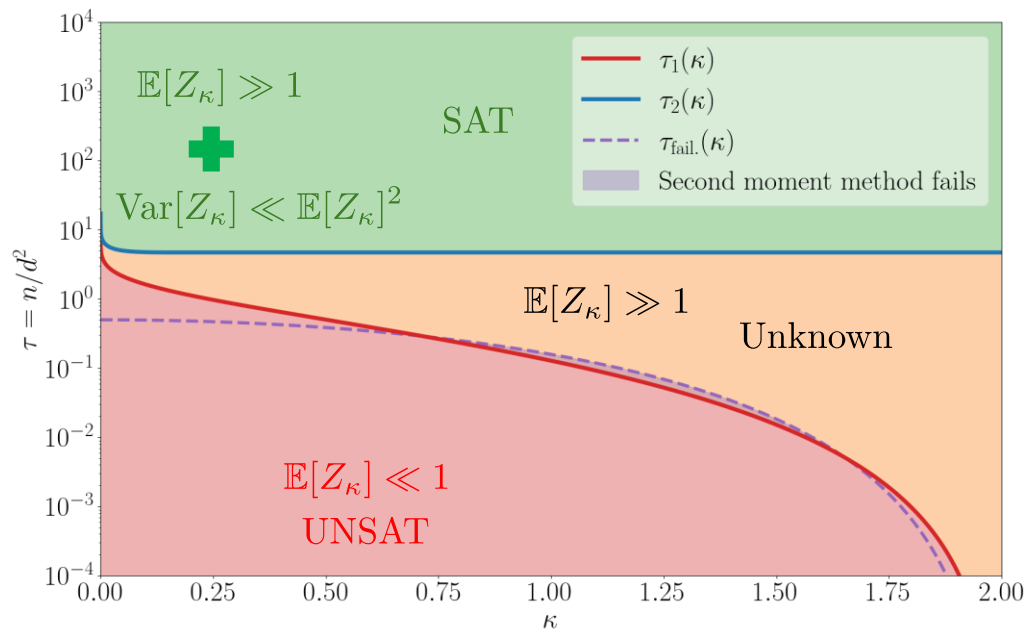
Refinements of Gaussian Poincaré inequality [Altschuler '23]

- An **exact characterization** of  $\mathbb{E}[Z_\kappa^2]$  would require to evaluate  $\Phi(q, \mu_1, \mu_2)$



**Challenging problem**  
more on that later

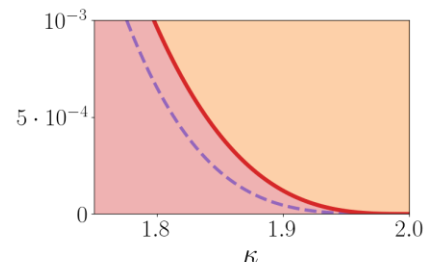
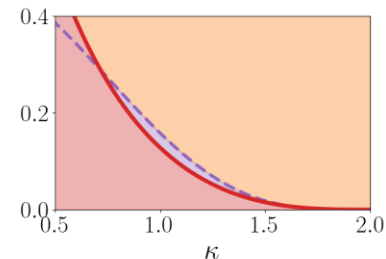
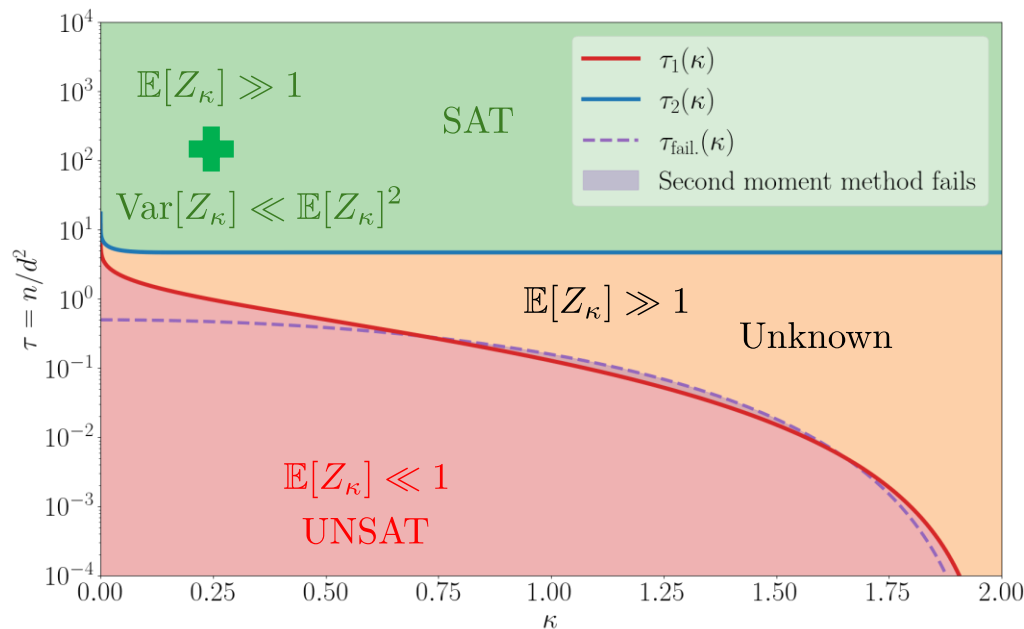
# Failure of the second moment method



**Theorem III:** Below the purple dashed line  $\sqrt{\text{Var}(Z_\kappa)} \gg \mathbb{E}[Z_\kappa]$

(And possibly in a much larger region)

# Failure of the second moment method



**Theorem III:** Below the purple dashed line  $\sqrt{\text{Var}(Z_\kappa)} \gg \mathbb{E}[Z_\kappa]$

(And possibly in a much larger region)

- The **second moment method fails at least** in the purple region
- Suggests quenched  $\neq$  annealed **at least** in the purple region



**More complex** than in the Symmetric Binary Perceptron !



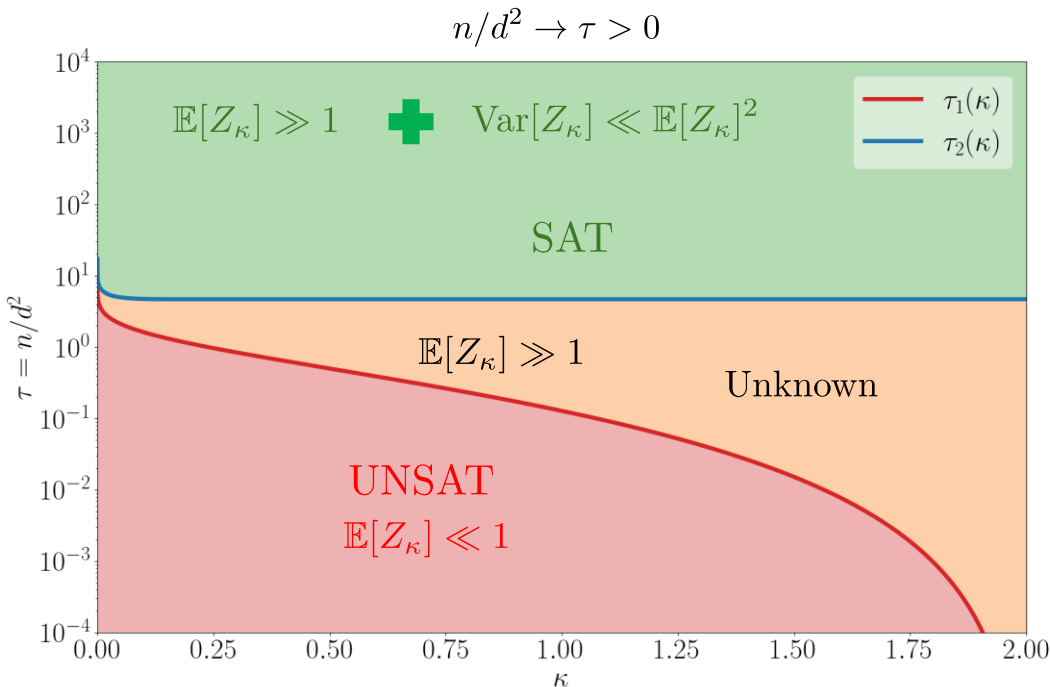
# Summary: average-case matrix discrepancy

$$Z_\kappa := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

## Matrix analog of the SBP



**Failure of the second moment method** in part of the diagram (  $\neq$  Symmetric Binary Perceptron )



# Summary: average-case matrix discrepancy

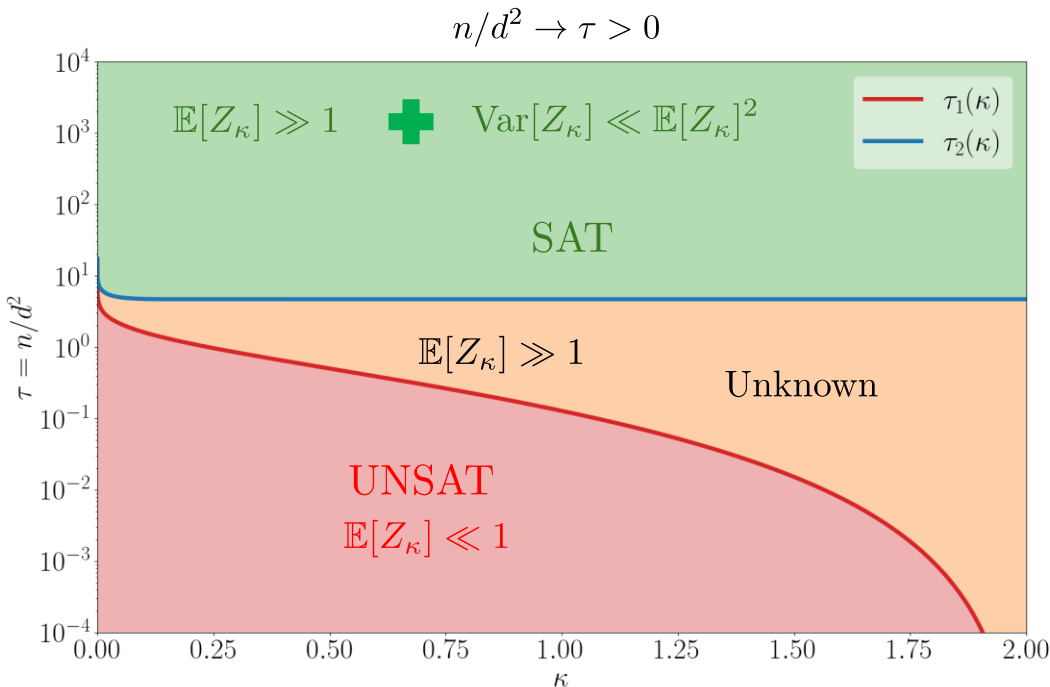
$$Z_\kappa := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

9

## Matrix analog of the SBP



Failure of the second moment method in part of the diagram ( $\neq$  Symmetric Binary Perceptron)



What's next ?

# Summary: average-case matrix discrepancy

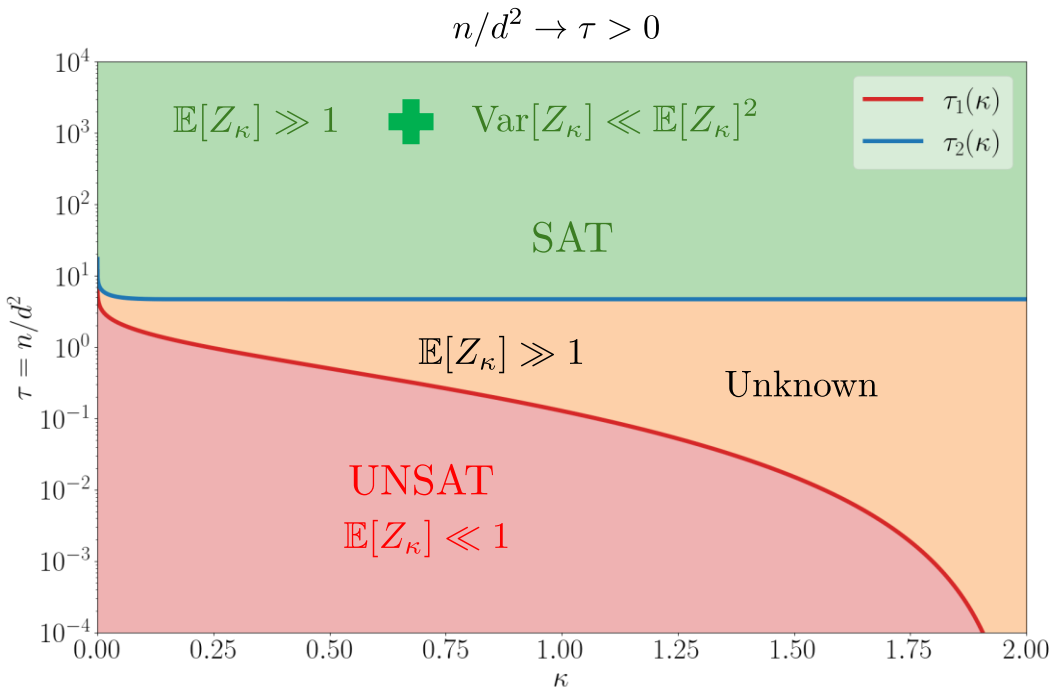
$$Z_\kappa := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

9

## Matrix analog of the SBP



Failure of the second moment method in part of the diagram ( $\neq$  Symmetric Binary Perceptron)



## What's next ?



- ☐ Sharp second moment
- ☐ Replica free energy (at least RS level)



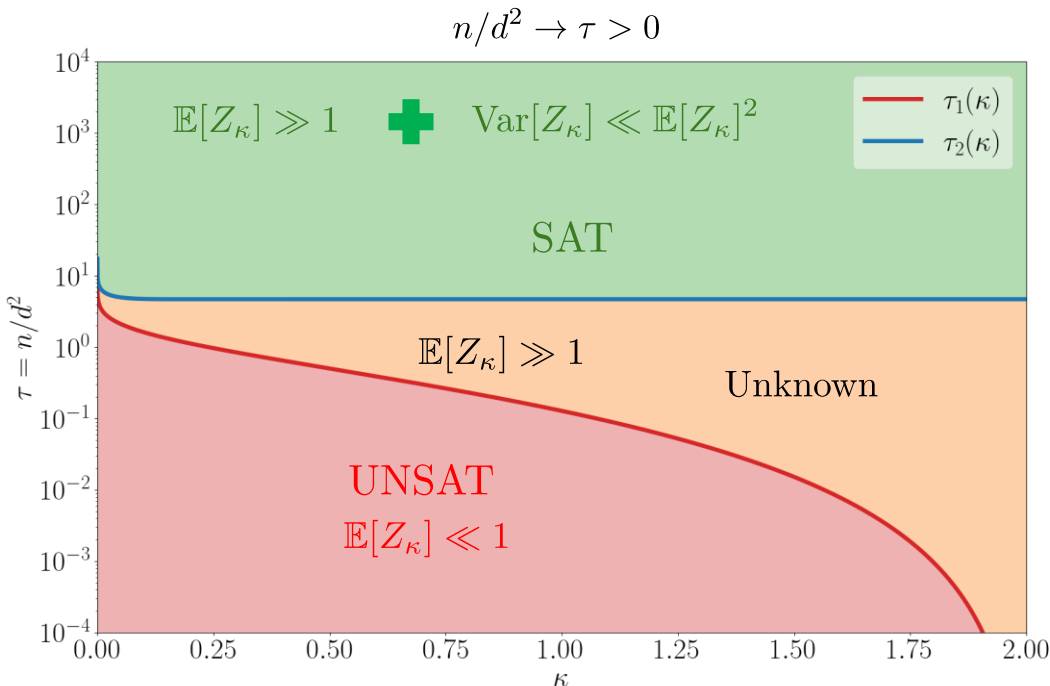
# Summary: average-case matrix discrepancy

$$Z_\kappa := \# \left\{ \varepsilon \in \{\pm 1\}^n \text{ s.t. } \left\| \sum_{i=1}^n \varepsilon_i \mathbf{W}_i \right\|_{\text{op}} \leq \kappa \sqrt{n} \right\}$$

9

## Matrix analog of the SBP

✚ Failure of the second moment method in part of the diagram ( $\neq$  Symmetric Binary Perceptron)



## What's next ?

- ☐ Sharp second moment
- ☐ Replica free energy (at least RS level)



- Structure of the solution space ?  
RS/RSB, ...
- (Efficient) algorithms ? [Kunisky & Zhang '23]
- Extensions to **non-GOE matrices** ?  
Large deviations of spectra of **structured matrices**

## Part II : The HCIZ integral

*w. J-C Mourrat (ENS Lyon)*



## From matrix discrepancy to large deviations...

---

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

# From matrix discrepancy to large deviations...

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices [Guionnet '04]

Exact asymptotics of  $\mathbb{E}[Z_\kappa^2]$



$$\Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

$\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$

# From matrix discrepancy to large deviations...

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices [Guionnet '04]

Exact asymptotics of  $\mathbb{E}[Z_\kappa^2]$



$$\Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

$\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$

$$\Phi(q, \mu_1, \mu_2) = -\frac{1}{d^2} \log \int d\mathbf{W} d\mathbf{Z} \frac{e^{-\frac{d}{4} \text{Tr}[\mathbf{W}^2 + \mathbf{Z}^2]}}{Z_d} \mathbb{1} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$



# From matrix discrepancy to large deviations...

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices [Guionnet '04]

Exact asymptotics of  $\mathbb{E}[Z_\kappa^2]$   $\Rightarrow$   $\Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$

$\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$

$$\Phi(q, \mu_1, \mu_2) = -\frac{1}{d^2} \log \int d\mathbf{W} d\mathbf{Z} \frac{e^{-\frac{d}{4} \text{Tr}[\mathbf{W}^2 + \mathbf{Z}^2]}}{Z_d} \mathbb{1} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

$$\Downarrow \quad \mathbf{W}_2 = q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}$$

$$= C_d(q) - \frac{1}{d^2} \log \int d\mathbf{W}_1 d\mathbf{W}_2 e^{-\frac{d}{4(1-q^2)} \text{Tr}[\mathbf{W}_1^2 + \mathbf{W}_2^2] + \frac{dq}{2(1-q^2)} \text{Tr}[\mathbf{W}_1 \mathbf{W}_2]} \mathbb{1} [\mu(\{\mathbf{W}_1\}) \simeq \mu_1] \mathbb{1} [\mu(\{\mathbf{W}_2\}) \simeq \mu_2]$$

# From matrix discrepancy to large deviations...

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices [Guionnet '04]

Exact asymptotics of  $\mathbb{E}[Z_\kappa^2]$   $\Rightarrow$   $\Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$

$\downarrow$   
 $\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$

$$\Phi(q, \mu_1, \mu_2) = -\frac{1}{d^2} \log \int d\mathbf{W} d\mathbf{Z} \frac{e^{-\frac{d}{4} \text{Tr}[\mathbf{W}^2 + \mathbf{Z}^2]}}{Z_d} \mathbb{1} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

$$\Downarrow \quad \mathbf{W}_2 = q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}$$

$$= C_d(q) - \frac{1}{d^2} \log \int d\mathbf{W}_1 d\mathbf{W}_2 e^{-\frac{d}{4(1-q^2)} \text{Tr}[\mathbf{W}_1^2 + \mathbf{W}_2^2] + \frac{dq}{2(1-q^2)} \text{Tr}[\mathbf{W}_1 \mathbf{W}_2]} \mathbb{1} [\mu(\{\mathbf{W}_1\}) \simeq \mu_1] \mathbb{1} [\mu(\{\mathbf{W}_2\}) \simeq \mu_2]$$

$$\mathbf{W}_a = \mathbf{O}_a \Lambda_a \mathbf{O}_a^\top \quad \Downarrow \quad \Lambda_a = \text{Diag}(\{\lambda_i^{(a)}\}_{i=1}^d)$$

# From matrix discrepancy to large deviations...

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices [Guionnet '04]

Exact asymptotics of  $\mathbb{E}[Z_\kappa^2]$   $\Rightarrow$   $\Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$

$\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$

$$\Phi(q, \mu_1, \mu_2) = -\frac{1}{d^2} \log \int d\mathbf{W} d\mathbf{Z} \frac{e^{-\frac{d}{4} \text{Tr}[\mathbf{W}^2 + \mathbf{Z}^2]}}{\mathcal{Z}_d} \mathbb{1} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

$$\Downarrow \quad \mathbf{W}_2 = q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}$$

$$= C_d(q) - \frac{1}{d^2} \log \int d\mathbf{W}_1 d\mathbf{W}_2 e^{-\frac{d}{4(1-q^2)} \text{Tr}[\mathbf{W}_1^2 + \mathbf{W}_2^2] + \frac{dq}{2(1-q^2)} \text{Tr}[\mathbf{W}_1 \mathbf{W}_2]} \mathbb{1} [\mu(\{\mathbf{W}_1\}) \simeq \mu_1] \mathbb{1} [\mu(\{\mathbf{W}_2\}) \simeq \mu_2]$$

$$\mathbf{W}_a = \mathbf{O}_a \mathbf{\Lambda}_a \mathbf{O}_a^\top \quad \Downarrow \quad \mathbf{\Lambda}_a = \text{Diag}(\{\lambda_i^{(a)}\}_{i=1}^d)$$

Jacobian of the change of variables

$$= \tilde{C}_d(q) - \frac{1}{d^2} \log \int \prod_{i=1}^d d\lambda_i^{(1)} d\lambda_i^{(2)} e^{-\frac{d}{4(1-q^2)} \text{Tr}[\mathbf{\Lambda}_1^2 + \mathbf{\Lambda}_2^2]} \mathbb{1} [\mu(\{\mathbf{\Lambda}_1\}) \simeq \mu_1] \mathbb{1} [\mu(\{\mathbf{\Lambda}_2\}) \simeq \mu_2] \prod_{i < j} |\lambda_i^{(1)} - \lambda_j^{(1)}| |\lambda_i^{(2)} - \lambda_j^{(2)}|$$

$$\times \mathbb{E}_{\mathbf{O}_1, \mathbf{O}_2} \left[ e^{\frac{dq}{2(1-q^2)} \text{Tr}[\mathbf{\Lambda}_1 \mathbf{O}_1^\top \mathbf{O}_2 \mathbf{\Lambda}_2 \mathbf{O}_2^\top \mathbf{O}_1]} \right]$$

$$\mathbf{O}_1, \mathbf{O}_2 \sim \text{Haar}(\mathcal{O}(d))$$

# From matrix discrepancy to large deviations...

$$\mu(\{\mathbf{W}\}) := \frac{1}{d} \sum_{i=1}^d \delta_{\lambda_i(\mathbf{W})}$$

Large deviations of the spectra of **correlated**  $\text{GOE}(d)$  matrices [Guionnet '04]

Exact asymptotics of  $\mathbb{E}[Z_\kappa^2] \implies \Phi(q, \mu_1, \mu_2) := -\frac{1}{d^2} \log \mathbb{P} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$

$\mathbf{W}, \mathbf{Z} \sim \text{GOE}(d)$

$$\Phi(q, \mu_1, \mu_2) = -\frac{1}{d^2} \log \int d\mathbf{W} d\mathbf{Z} \frac{e^{-\frac{d}{4} \text{Tr}[\mathbf{W}^2 + \mathbf{Z}^2]}}{Z_d} \mathbb{1} \left[ \mu(\{\mathbf{W}\}) \simeq \mu_1 \text{ and } \mu(\{q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}\}) \simeq \mu_2 \right]$$

$$\Downarrow \quad \mathbf{W}_2 = q\mathbf{W} + \sqrt{1-q^2}\mathbf{Z}$$

$$= C_d(q) - \frac{1}{d^2} \log \int d\mathbf{W}_1 d\mathbf{W}_2 e^{-\frac{d}{4(1-q^2)} \text{Tr}[\mathbf{W}_1^2 + \mathbf{W}_2^2] + \frac{dq}{2(1-q^2)} \text{Tr}[\mathbf{W}_1 \mathbf{W}_2]} \mathbb{1} [\mu(\{\mathbf{W}_1\}) \simeq \mu_1] \mathbb{1} [\mu(\{\mathbf{W}_2\}) \simeq \mu_2]$$

$$\mathbf{W}_a = \mathbf{O}_a \mathbf{\Lambda}_a \mathbf{O}_a^\top \quad \Downarrow \quad \mathbf{\Lambda}_a = \text{Diag}(\{\lambda_i^{(a)}\}_{i=1}^d)$$

Jacobian of the change of variables

$$= \tilde{C}_d(q) - \frac{1}{d^2} \log \int \prod_{i=1}^d d\lambda_i^{(1)} d\lambda_i^{(2)} e^{-\frac{d}{4(1-q^2)} \text{Tr}[\mathbf{\Lambda}_1^2 + \mathbf{\Lambda}_2^2]} \mathbb{1} [\mu(\{\mathbf{\Lambda}_1\}) \simeq \mu_1] \mathbb{1} [\mu(\{\mathbf{\Lambda}_2\}) \simeq \mu_2] \prod_{i < j} |\lambda_i^{(1)} - \lambda_j^{(1)}| |\lambda_i^{(2)} - \lambda_j^{(2)}|$$

$$\times \mathbb{E}_{\mathbf{O}} \left[ e^{\frac{dq}{2(1-q^2)} \text{Tr}[\mathbf{\Lambda}_1 \mathbf{O} \mathbf{\Lambda}_2 \mathbf{O}^\top]} \right]$$

**Key quantity**

$\mathbf{O} \sim \text{Haar}(\mathcal{O}(d))$

## ... to the HCIZ integral

---

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

$$\mu(M_a) \rightarrow \mu_a \text{ as } d \rightarrow \infty \quad (a \in \{1, 2\})$$

Harish-Chandra '57, Itzykson and Zuber '80 ;  
Matytsin '94 ; Guionnet & Zeitouni '02 ; ...

## ... to the HCIZ integral

---

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

Harish-Chandra '57, Itzykson and Zuber '80 ;  
Matytsin '94 ; Guionnet & Zeitouni '02 ; ...

$$\mu(M_a) \rightarrow \mu_a \text{ as } d \rightarrow \infty \quad (a \in \{1, 2\})$$

**A random matrix integral with several motivations and applications...**

### ❖ Large deviations theory

- Multi-matrix models (our example)
- Matrix models with an external field

Guionnet's ICM talk (2022)

## ... to the HCIZ integral

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

Harish-Chandra '57, Itzykson and Zuber '80 ;  
Matytsin '94 ; Guionnet & Zeitouni '02 ; ...

$$\mu(M_a) \rightarrow \mu_a \text{ as } d \rightarrow \infty \quad (a \in \{1, 2\})$$

### A random matrix integral with several motivations and applications...

#### ❖ Large deviations theory

- Multi-matrix models (our example)
- Matrix models with an external field

Guionnet's ICM talk (2022)

#### ❖ Free energy in Bayesian inference/learning when parameters are **large matrices, out of Bayes-optimality**

- Matrix denoising / Matrix sensing with rotationally-invariant priors
- Overparametrized and wide two-layers neural networks with quadratic activations

Bun&al '16; M.&al '22; M.&al '24;  
Semerjian '24; Barbier & al '25;  
Erba&al'25 ; ...

## ... to the HCIZ integral

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

Harish-Chandra '57, Itzykson and Zuber '80 ;  
Matytsin '94 ; Guionnet & Zeitouni '02 ; ...

$$\mu(M_a) \rightarrow \mu_a \text{ as } d \rightarrow \infty \text{ } (a \in \{1, 2\})$$

### A random matrix integral with several motivations and applications...

#### ❖ Large deviations theory

- Multi-matrix models (our example)
- Matrix models with an external field

Guionnet's ICM talk (2022)

#### ❖ Free energy in Bayesian inference/learning when parameters are **large matrices, out of Bayes-optimality**

- Matrix denoising / Matrix sensing with rotationally-invariant priors
- Overparametrized and wide two-layers neural networks with quadratic activations

Bun&al '16; M.&al '22; M.&al '24;  
Semerjian '24; Barbier & al '25;  
Erba&al'25 ; ...



Other problems in mathematical physics (quantum gravity, ...)

Terry Tao's blog post on the HCIZ integral (2013) ; McSwiggen '18



## ... to the HCIZ integral

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

Harish-Chandra '57, Itzykson and Zuber '80 ;  
Matytsin '94 ; Guionnet & Zeitouni '02 ; ...

$$\mu(M_a) \rightarrow \mu_a \text{ as } d \rightarrow \infty \text{ } (a \in \{1, 2\})$$

### A random matrix integral with several motivations and applications...

#### ❖ Large deviations theory

- Multi-matrix models (our example)
- Matrix models with an external field

Guionnet's ICM talk (2022)

#### ❖ Free energy in Bayesian inference/learning when parameters are **large matrices, out of Bayes-optimality**

- Matrix denoising / Matrix sensing with rotationally-invariant priors
- Overparametrized and wide two-layers neural networks with quadratic activations

Bun&al '16; M.&al '22; M.&al '24;  
Semerjian '24; Barbier & al '25;  
Erba&al'25 ; ...



Other problems in mathematical physics (quantum gravity, ...) Terry Tao's blog post on the HCIZ integral (2013) ; McSwiggen '18

**This talk**

Given arbitrary  $(\mu_1, \mu_2)$ , how to numerically evaluate  $I_{\text{HCIZ}}(\mu_1, \mu_2)$



## Matytsin's solution

---

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(\mathrm{d}O) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

# Matytsin's solution

---

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{O(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

[Matytsin '94 ; Guionnet & Zeitouni '02]

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]$$

Constraints

$$\partial_t \rho + \partial_x(\rho v) = 0$$

$$\rho(0, \cdot) = \mu$$

$$\rho(1, \cdot) = \nu$$

# Matytsin's solution

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{\mathcal{O}(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

[Matytsin '94 ; Guionnet & Zeitouni '02]

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]$$

Constraints

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0 \\ \rho(0, \cdot) &= \mu \\ \rho(1, \cdot) &= \nu \end{aligned}$$

➤ A unique minimizer  $(\rho^*, v^*)$ , which satisfies

$$\partial_t v + v \partial_x v = \pi^2 \rho \partial_x \rho$$

Euler's equations of hydrodynamics, with a **negative** pressure field  $P = -\frac{\pi^2}{3} \rho^3$

# Matytsin's solution

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{O(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

[Matytsin '94 ; Guionnet & Zeitouni '02]

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]$$

Constraints

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0 \\ \rho(0, \cdot) &= \mu \\ \rho(1, \cdot) &= \nu \end{aligned}$$

➤ A unique minimizer  $(\rho^*, v^*)$ , which satisfies

$$\partial_t v + v \partial_x v = \pi^2 \rho \partial_x \rho$$

Euler's equations of hydrodynamics, with a **negative** pressure field  $P = -\frac{\pi^2}{3} \rho^3$



A hard problem for (relatively simple) hydrodynamical PDE solvers

# Matytsin's solution

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{O(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

## Constraints

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]$$

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0 \\ \rho(0, \cdot) &= \mu \\ \rho(1, \cdot) &= \nu \end{aligned}$$

➤ A unique minimizer  $(\rho^*, v^*)$ , which satisfies

$$\partial_t v + v \partial_x v = \pi^2 \rho \partial_x \rho$$

Euler's equations of hydrodynamics, with a **negative** pressure field  $P = -\frac{\pi^2}{3} \rho^3$



A hard problem for (relatively simple) hydrodynamical PDE solvers

Connections to **complex analysis** and **integrable systems**



Particular solutions, but **no general numerical approach**

[Matytsin '94; Menon '17; Schmidt '18; ...]

[Matytsin '94 ; Guionnet & Zeitouni '02]

# Matytsin's solution

$$I_{\text{HCIZ}}(\mu_1, \mu_2) := \lim_{d \rightarrow \infty} \frac{2}{d^2} \log \int_{O(d)} \mu_{\text{Haar}}(dO) e^{\frac{d}{2} \text{Tr}[OM_1 O^\top M_2]}$$

[Matytsin '94 ; Guionnet & Zeitouni '02]

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]$$

Constraints

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0 \\ \rho(0, \cdot) &= \mu \\ \rho(1, \cdot) &= \nu \end{aligned}$$

➤ A unique minimizer  $(\rho^*, v^*)$ , which satisfies

$$\partial_t v + v \partial_x v = \pi^2 \rho \partial_x \rho$$

Euler's equations of hydrodynamics, with a **negative** pressure field  $P = -\frac{\pi^2}{3} \rho^3$



A hard problem for (relatively simple) hydrodynamical PDE solvers

Connections to **complex analysis** and **integrable systems**



Particular solutions, but **no general numerical approach**

[Matytsin '94; Menon '17; Schmidt '18; ...]

**Our approach**

Discretize the infimum over trajectories  $(\rho, v)$

Similar to ideas used in optimal transport  
[Benamou & Brenier '00 '01, ...]

# Discretization, and large-scale minimization



Convex problem in  $(\rho, \rho v)$

[Matytsin '94 ; Guionnet & Zeitouni '02]

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \underbrace{\frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]}_{=: J(\mu, \nu)}.$$

Constraints

$$\rho(0, \cdot) = \mu$$

$$\rho(1, \cdot) = \nu$$

$$\partial_t \rho + \partial_x(\rho v) = 0$$



# Discretization, and large-scale minimization



Convex problem in  $(\rho, \rho v)$

Constraints

$$\rho(0, \cdot) = \mu$$

$$\rho(1, \cdot) = \nu$$

$$\partial_t \rho + \partial_x(\rho v) = 0$$

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \underbrace{\frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]}_{=: J(\mu, \nu)}.$$

**Theorem [M. & Mourrat '25] (informal)**

$(\mu, \nu)$  compactly supported, and with continuous densities  $\Rightarrow J(\mu, \nu) = \lim_{N, T \rightarrow \infty} J_{N, T}(\mu, \nu)$

# Discretization, and large-scale minimization



Convex problem in  $(\rho, \rho v)$

Constraints

$$\rho(0, \cdot) = \mu$$

$$\rho(1, \cdot) = \nu$$

$$\partial_t \rho + \partial_x(\rho v) = 0$$

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \underbrace{\frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]}_{=: J(\mu, \nu)}.$$

**Theorem [M. & Mourrat '25] (informal)**

$(\mu, \nu)$  compactly supported, and with continuous densities  $\implies J(\mu, \nu) = \lim_{N, T \rightarrow \infty} J_{N, T}(\mu, \nu)$

$$J_{N, T}(\mu, \nu) = \inf_{\{x_i(t_k)\}} \left[ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^T \Delta t_k \left( v_i(t_k)^2 + \frac{\pi^2}{3} \underbrace{\frac{1}{N^2 (x_{i+1}(t_k) - x_i(t_k))^2}}_{\simeq \rho(x_i(t_k))^2} \right) dt \right]$$

$v_i(t_k) := \frac{x_i(t_{k+1}) - x_i(t_k)}{\Delta t_k}$

$\{x_i(0)\}$  quantiles of  $\mu$

$\{x_i(1)\}$  quantiles of  $\nu$

$$x_i(t_k) < x_{i+1}(t_k)$$

Linear constraints

# Discretization, and large-scale minimization



Convex problem in  $(\rho, \rho v)$

Constraints

$$\rho(0, \cdot) = \mu$$

$$\rho(1, \cdot) = \nu$$

$$\partial_t \rho + \partial_x(\rho v) = 0$$

$$I_{\text{HCIZ}}(\mu, \nu) = F(\mu) + F(\nu) - \underbrace{\frac{1}{2} \inf_{\rho, v} \int_0^1 dt \int dx \rho(t, x) \left[ v(t, x)^2 + \frac{\pi^2}{3} \rho(t, x)^2 \right]}_{=: J(\mu, \nu)}.$$

**Theorem [M. & Mourrat '25] (informal)**

$(\mu, \nu)$  compactly supported, and with continuous densities  $\implies J(\mu, \nu) = \lim_{N, T \rightarrow \infty} J_{N, T}(\mu, \nu)$

$$J_{N, T}(\mu, \nu) = \inf_{\{x_i(t_k)\}} \left[ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^T \Delta t_k \left( v_i(t_k)^2 + \frac{\pi^2}{3} \underbrace{\frac{1}{N^2 (x_{i+1}(t_k) - x_i(t_k))^2}}_{\simeq \rho(x_i(t_k))^2} \right) dt \right]$$

$v_i(t_k) := \frac{x_i(t_{k+1}) - x_i(t_k)}{\Delta t_k}$

$\{x_i(0)\}$  quantiles of  $\mu$

$\{x_i(1)\}$  quantiles of  $\nu$

$$x_i(t_k) < x_{i+1}(t_k)$$

Linear constraints



- Proof requires careful handling of potential singularities in  $v(t, x)$
- The discretization **preserves the convexity** of the minimization problem

## A word on the numerical scheme

---

$$J_{N,T}(\mu, \nu) = \inf_{\{x_i(t_k)\}} \left[ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^T \Delta t_k \left( v_i(t_k)^2 + \frac{\pi^2}{3} \frac{1}{N^2 (x_{i+1}(t_k) - x_i(t_k))^2} \right) dt \right]$$

$\{x_i(0)\}$  quantiles of  $\mu$

$\{x_i(1)\}$  quantiles of  $\nu$

$x_i(t_k) < x_{i+1}(t_k)$

**Linear constraints**

## A word on the numerical scheme

$$J_{N,T}(\mu, \nu) = \inf_{\{x_i(t_k)\}} \left[ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^T \Delta t_k \left( v_i(t_k)^2 + \frac{\pi^2}{3} \frac{1}{N^2 (x_{i+1}(t_k) - x_i(t_k))^2} \right) dt \right]$$

$\{x_i(0)\}$  quantiles of  $\mu$

$\{x_i(1)\}$  quantiles of  $\nu$

$x_i(t_k) < x_{i+1}(t_k)$

**Linear constraints**

❖ A convex problem but very **ill-conditioned in general**  $\implies$  Naïve 1<sup>st</sup> order methods struggle for large  $(N, T)$

The KKT matrix has condition number  $\kappa \sim \Theta(N^2 T^2)$

## A word on the numerical scheme

$$J_{N,T}(\mu, \nu) = \inf_{\{x_i(t_k)\}} \left[ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^T \Delta t_k \left( v_i(t_k)^2 + \frac{\pi^2}{3} \frac{1}{N^2 (x_{i+1}(t_k) - x_i(t_k))^2} \right) dt \right]$$

$\{x_i(0)\}$  quantiles of  $\mu$

$\{x_i(1)\}$  quantiles of  $\nu$

$x_i(t_k) < x_{i+1}(t_k)$

**Linear constraints**

- ❖ A convex problem but very **ill-conditioned in general**  $\implies$  Naïve 1<sup>st</sup> order methods struggle for large  $(N, T)$

The KKT matrix has condition number  $\kappa \sim \Theta(N^2 T^2)$

- ❖ We use approximate **second-order methods** with strong **preconditioning** techniques.

Using fast approximations of the inverse Hessian of  $J_{N,T}$

## A word on the numerical scheme

$$J_{N,T}(\mu, \nu) = \inf_{\{x_i(t_k)\}} \left[ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^T \Delta t_k \left( v_i(t_k)^2 + \frac{\pi^2}{3} \frac{1}{N^2 (x_{i+1}(t_k) - x_i(t_k))^2} \right) dt \right]$$

$\{x_i(0)\}$  quantiles of  $\mu$   
 $\{x_i(1)\}$  quantiles of  $\nu$   
 $x_i(t_k) < x_{i+1}(t_k)$

Linear constraints

- ❖ A convex problem but very **ill-conditioned in general**  $\implies$  Naïve 1<sup>st</sup> order methods struggle for large  $(N, T)$

The KKT matrix has condition number  $\kappa \sim \Theta(N^2 T^2)$

- ❖ We use approximate **second-order methods** with strong **preconditioning** techniques.

Using fast approximations of the inverse Hessian of  $J_{N,T}$

- ❖ Allows for **large-scale computation**

$N, T \sim (10^3, 10^4)$  is typically solved in  $\Theta(1 \text{ min.})$  on a single GPU

## First applications (1)...

---

### Benchmark I

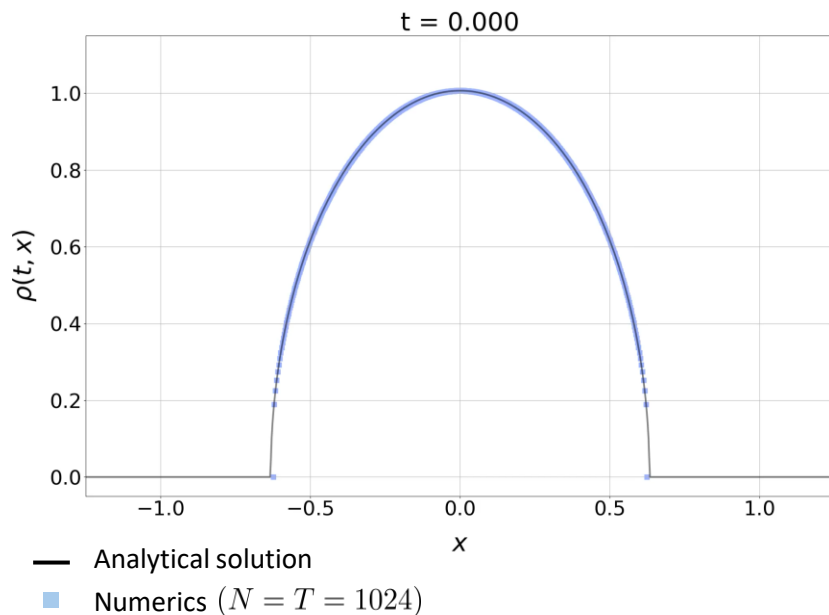
$\mu = \nu = \sigma_{\text{s.c.}}$  (semicircle), with variance  $\sigma^2$  [Bun & al '16]



# First applications (1)...

## Benchmark I

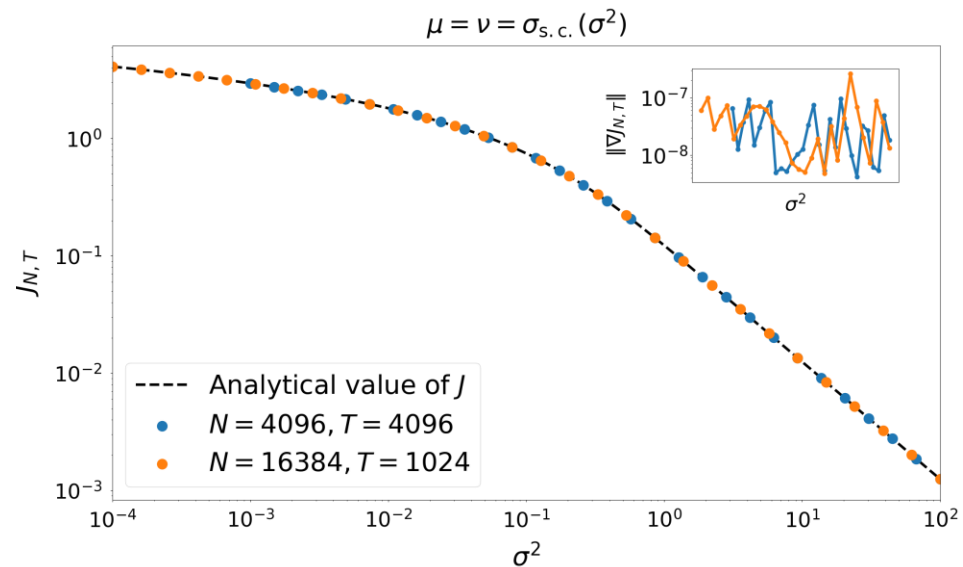
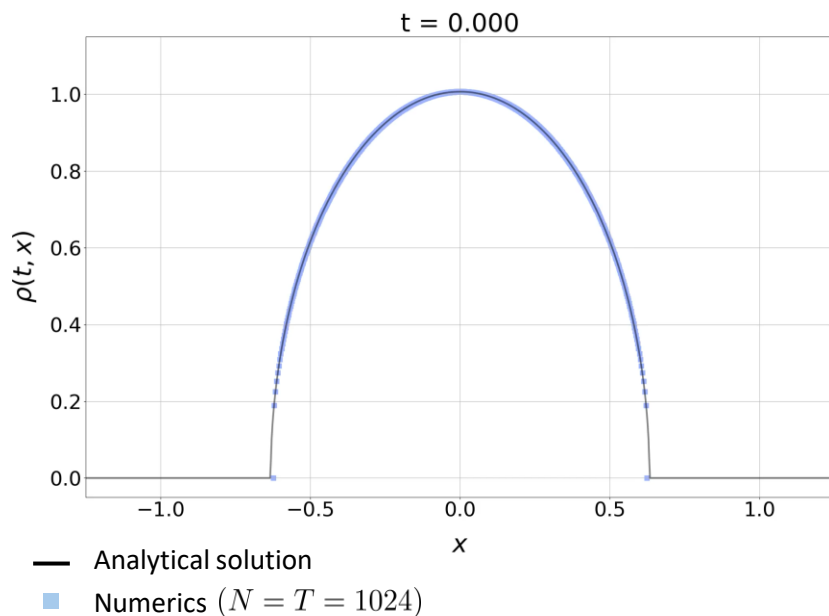
$\mu = \nu = \sigma_{\text{s.c.}}$  (semicircle), with variance  $\sigma^2$  [Bun & al '16]



# First applications (1)...

## Benchmark I

$\mu = \nu = \sigma_{\text{s.c.}}$  (semicircle), with variance  $\sigma^2$  [Bun & al '16]



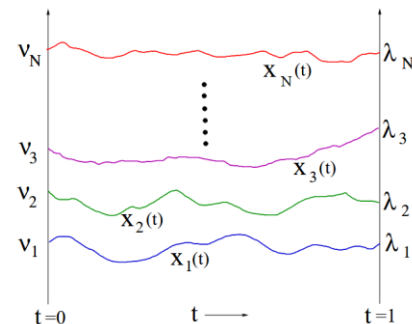
## First applications (2)...

[Matytsin '94; Guionnet-Zeitouni '02; Bun&a'l'16]

$\rho^*(t)$  is the eigenvalue density of the Dyson Brownian motion  $A(t) = M + \sqrt{t}W$ ,  
**constrained** to have eigenvalue density  $\nu$  at time  $t = 1$

$$\mu(M) \sim \mu$$

$$W \sim \text{GOE}(d)$$



## First applications (2)...

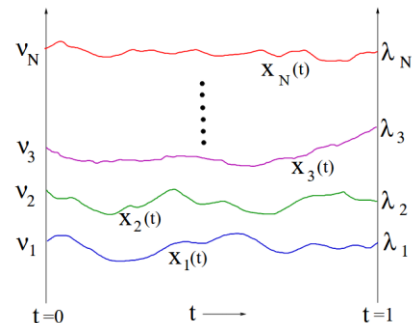
[Matytsin '94; Guionnet-Zeitouni '02; Bun&a'16]

$\rho^*(t)$  is the eigenvalue density of the Dyson Brownian motion  $A(t) = M + \sqrt{t}W$ ,

**constrained** to have eigenvalue density  $\nu$  at time  $t = 1$

$\mu(M) \sim \mu$

$W \sim \text{GOE}(d)$

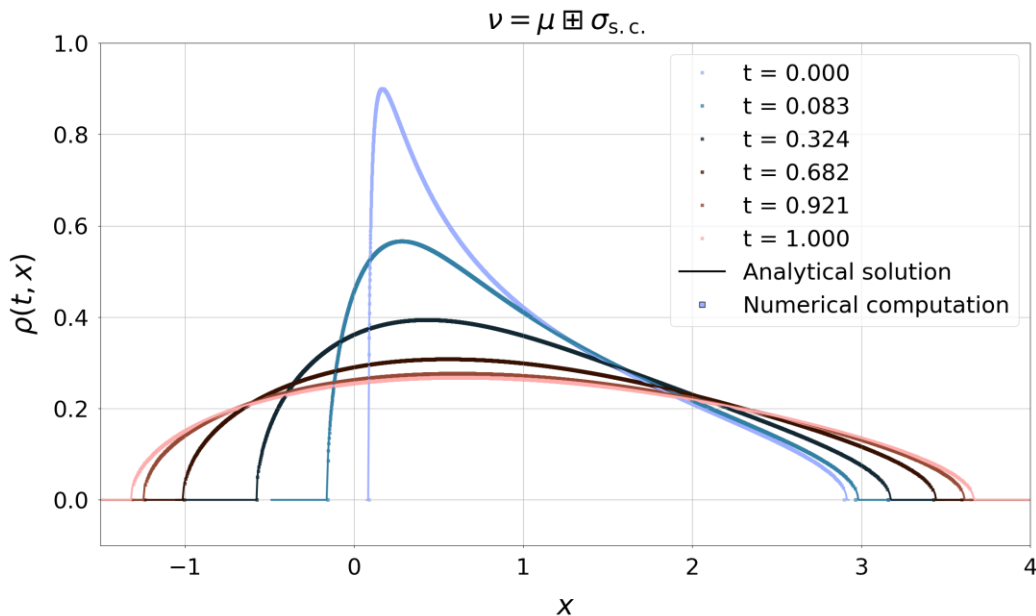


**Remark:** Without any constraints,  $\rho(t, \cdot) = \mu \boxplus \sigma_{s.c., \sqrt{t}}$  (free convolution)



### Benchmark II

Free convolution  $\nu = \mu \boxplus \sigma_{s.c.}$



## First applications (2)...

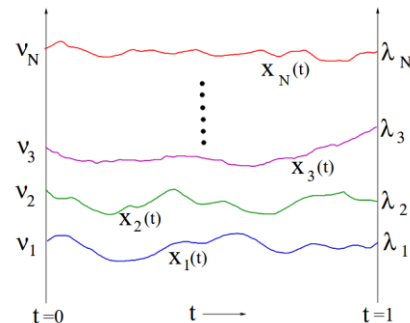
[Matytsin '94; Guionnet-Zeitouni '02; Bun&a'16]

$\rho^*(t)$  is the eigenvalue density of the Dyson Brownian motion  $A(t) = M + \sqrt{t}W$ ,

**constrained** to have eigenvalue density  $\nu$  at time  $t = 1$

$\mu(M) \sim \mu$

$W \sim \text{GOE}(d)$



**Remark:** Without any constraints,  $\rho(t, \cdot) = \mu \boxplus \sigma_{s.c., \sqrt{t}}$  (free convolution)



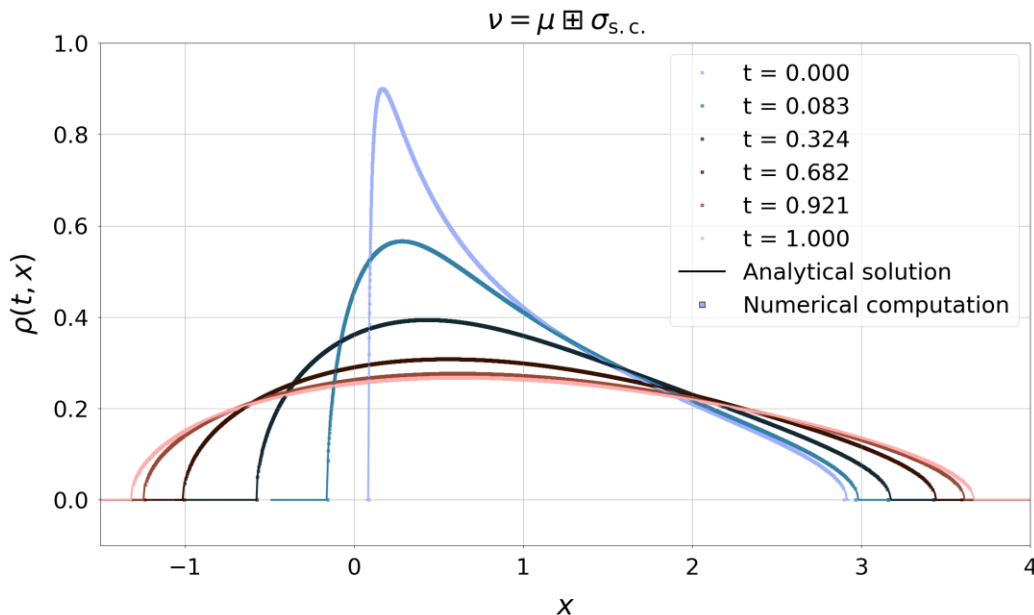
### Benchmark II

Free convolution  $\nu = \mu \boxplus \sigma_{s.c.}$



These 2 benchmarks are the **only known analytical solutions** to Matytsin's equations !

(to the best of my knowledge)



## First applications (3)...

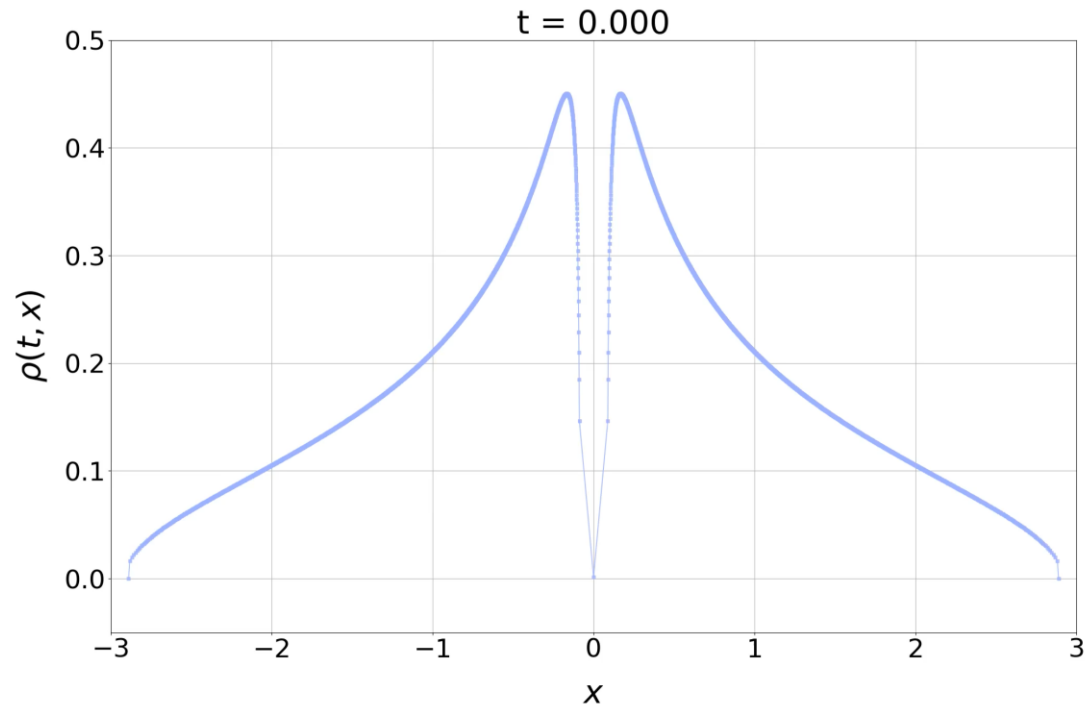
---

Our solver can be applied to arbitrary  $(\mu, \nu)$ , for which **no analytical solution exists**.

## First applications (3)...

Our solver can be applied to arbitrary  $(\mu, \nu)$ , for which **no analytical solution exists**.

### Example



## ... and work in progress

---

### Conclusion

A provably correct general-purpose solver to compute  $I_{\text{HCIZ}}(\mu, \nu)$  for **arbitrary**  $(\mu, \nu)$



## ... and work in progress

---

### Conclusion

A **provably correct general-purpose solver** to compute  $I_{\text{HCIZ}}(\mu, \nu)$  for **arbitrary**  $(\mu, \nu)$

#### ☐ Other numerical solvers ?

PDE solvers, integrable systems, augmented Lagrangian methods, ...

#### ☐ Applications in **disordered systems and statistical learning**

- Sharp phase diagram in random matrix discrepancy
- High-rank and mismatched (non Bayes-optimal) matrix denoising
- Overparametrized two-layers neural networks with quadratic activations

#### ☐ ... Other applications (large deviations theory) ?

What's  
next



## ... and work in progress

### Conclusion

A provably correct general-purpose solver to compute  $I_{\text{HCIZ}}(\mu, \nu)$  for **arbitrary**  $(\mu, \nu)$

### What's next



- ❑ Other numerical solvers ?  
PDE solvers, integrable systems, augmented Lagrangian methods, ...
- ❑ Applications in **disordered systems and statistical learning**
  - Sharp phase diagram in random matrix discrepancy
  - High-rank and mismatched (non Bayes-optimal) matrix denoising
  - Overparametrized two-layers neural networks with quadratic activations
- ❑ ... Other applications (large deviations theory) ?



Needs to solve variational problems:

$$\sup_{\mu} [G(\mu) + I_{\text{HCIZ}}(\mu, \nu)]$$

e.g. all  $\mu$  such that  $\text{supp}(\mu) \subseteq [-\kappa, \kappa]$  in random matrix discrepancy



Generalize our approach to optimize as well over the boundary densities



## ... and work in progress

### Conclusion

A provably correct general-purpose solver to compute  $I_{\text{HCIZ}}(\mu, \nu)$  for **arbitrary**  $(\mu, \nu)$

#### ❑ Other numerical solvers ?

PDE solvers, integrable systems, augmented Lagrangian methods, ...

#### ❑ Applications in **disordered systems and statistical learning**

- Sharp phase diagram in random matrix discrepancy
- High-rank and mismatched (non Bayes-optimal) matrix denoising
- Overparametrized two-layers neural networks with quadratic activations

#### ❑ ... Other applications (large deviations theory) ?



Needs to solve variational problems:

$$\sup_{\mu} [G(\mu) + I_{\text{HCIZ}}(\mu, \nu)]$$

e.g. all  $\mu$  such that  $\text{supp}(\mu) \subseteq [-\kappa, \kappa]$  in random matrix discrepancy



Generalize our approach to optimize as well over the boundary densities



What's  
next



[arXiv:25XX.XXXX ?](#)

# THANK YOU !