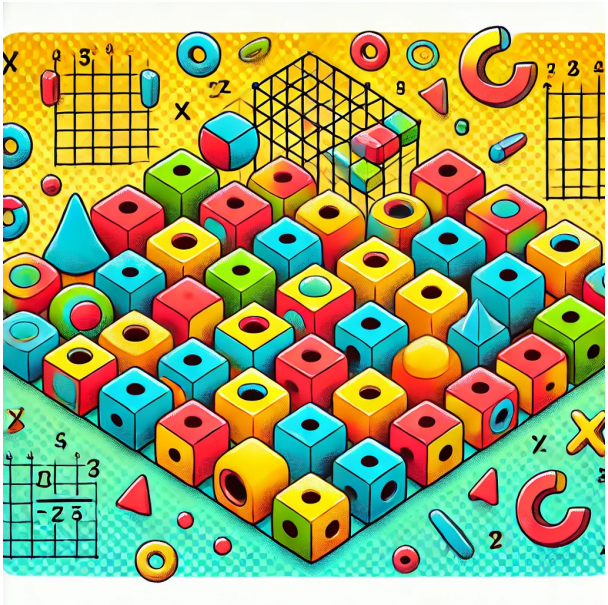
Balancing Random Matrices and the Local Entropy of a Planted Model

Simone Maria Giancola
simonegiancola09@gmail.com

Imrane Zaakour
imrane.zaakour@ensae.fr

February 1, 2025



Abstract

We discuss recent results on the problem of balancing random matrices. The origin is from its deterministic version, one of the most important open questions in discrepancy theory. The first part is a walk through of a very recent paper (Maillard 2024), emphasizing its contribution. Then, drawing up on the final comments of the results, we analyze the geometry of solutions, i.e. how the balancing points are arranged in space. We do this through a companion model that was the main object of the analysis for the analog vector problem. For a specific design, we find that solutions are isolated, meaning that at least $\Omega(n)$ signs have to change to hop from one to the other. This suggests that starting from a solution it is polynomially hopeless to find another.

Contents

I Introduction

In this document we present random properties of matrix discrepancy. The deterministic formulation is as follows.

Matrix discrepancy

Given n square symmetric matrices $\{\boldsymbol{A}^{(i)}\}_{i=1}^n$ in $\mathbb{R}^{d\times d}$ and a tolerance $\Delta \equiv \Delta(n,d) > 0$ find if there is a collection of signs $\{\epsilon_i\}_{i=1}^n \in \{\pm 1\}^n$ such that

$$\left\| \sum_{i=1}^n \epsilon_i \boldsymbol{A}^{(i)} \right\|_{\mathsf{op}} \leq \Delta. \tag{I.1}$$

In words: we want to find a balancing of the matrices such that their magnitude is small. It is important to notice that we only require an answer for *existence*, as finding an *explicit solution* is even harder.

Let us briefly discuss why the question is interesting and non-trivial. For validity: understanding the discrepancy of a collection of objects has applications in many fields of mathematics. We refer to an introductory blog-post (Bandeira 2024) and some works (Spencer 1985; Maillard 2024; Kunisky and Zhang 2024) for more context.
Concerning why it is not immediate, the first observation is that norms are not additive, but there are deeper reasons. If we replace matrices with vectors, the result is classical. Six deviations suffice to get to a "sub-entropic" behavior.

- **Theorem 1** (Spencer (1985)). *There exists $C \in (0,6)$ such that for all $n, d \geq 1$:*

$$\sup_{\{\boldsymbol{u}^{(i)}\}_{i=1}^n : \left\|\boldsymbol{u}^{(i)}\right\|_\infty \leq 1} \min_{\{\epsilon_i\}_{i=1}^n} \left\| \sum_{i=1}^n \epsilon_i \boldsymbol{u}^{(i)} \right\|_\infty \leq C\sqrt{n \max\left\{1, \log \frac{d}{n}\right\}}. \tag{I.2}$$

In words, for any collection of vectors with bounded sup norm taking Δ scaling as $\sqrt{n \log d/n}$ we know there are signs balancing below this threshold.

Coming back to our case, we get stuck on the simplest linear algebra remark: matrices do not commute. Indeed, the result would be extended for free since vectors are just diagonal matrices, but the first annoying difference between diagonal and generic matrices is exactly commutativity. Nevertheless, there is increasing interest in proving a free-lunch theorem, i.e. that the same bound extends. We box it below for future reference.

Conjecture 2 (Matrix Spencer's). *There exists $C > 0$ such that for all $n, d \geq 1$:*

$$\sup_{\{\boldsymbol{A}^{(i)}\}_{i=1}^n : \left\|\boldsymbol{A}^{(i)}\right\|_{\mathsf{op}} \leq 1,\ \boldsymbol{A}^{(i)} = [\boldsymbol{A}^{(i)}]^\top} \min_{\{\epsilon_i\}_{i=1}^n} \left\| \sum_{i=1}^n \epsilon_i \boldsymbol{A}^{(i)} \right\|_{\mathsf{op}} \leq C\sqrt{n \max\left\{1, \log \frac{d}{n}\right\}}. \tag{I.3}$$

We avoid discussing why the current state of solutions is not satisfactory, but remark that we are far from having a bound of the same order for any symmetric collection. The interested

reader can check (Maillard 2024, section 1.1.1).
Motivated by the generalization claim, throughout the text we will discuss the vector case to guess or justify some choices by the principle of analogy.

Notation Most symbols are standard. The only difference is a clear distinction between randomness and size. This is useful in the long run. We denote scalars a, b, c, random scalars $\mathrm{a}, \mathrm{b}, \mathrm{c}$; vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, with entries a_i, b_i, c_i random vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$, with entries $\mathrm{a}_i, \mathrm{b}_i, \mathrm{c}_i$; matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, with entries A_{ij}, B_{ij}, C_{ij}, random matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ with entries $\mathrm{A}_{ij}, \mathrm{B}_{ij}, \mathrm{C}_{ij}$.

I.I Sprinkling randomness

Often studying a random version of a problem is easier. The idea is that laws of probability inject regularity. While apparently useless, it is also common that this improves the understanding of the deterministic case. Let us then rephrase our problem in the language of ensembles. The nicest choice as usual is Gaussian. For convenience, we remind below what it means to be Gaussian (in a particular way) for a matrix.

• **Definition 3** (Gaussian Orthogonal Ensemble (GOE))**.** *The Gaussian Orthogonal Ensemble (GOE) is a distribution over symmetric matrices in $\mathbb{R}^{d\times d}$ that is Gaussian in the sense that each entry is Gaussian and independent. For the purpose of this document, the scaling we choose is such that each entry has mean zero and variance $1/d$ in the off-diagonal and $2/d$ in the diagonal. We denote it as* $\mathrm{GOE}(d)$*, which stresses the dimension and the moments. If we write* $\mathsf{GOE}(d, p)$*, then p is a added in the numerator of the moments.*

Remark. *The normalization choice is without loss of generality but is useful because the law is invariant with respect to rotations, not only asymptotically. Moreover, it has finite variance for any dimension.*

It is very easy to sample from the GOE. For a given d, we throw d iid Gaussians from $\mathcal{N}(0, 2/d)$ and $1/2\, d(d-1)$ iid Gaussians from $\mathcal{N}(0, 1/d)$. Then, we fill the matrix in the diagonal with the first ones, and use the others to complete the table such that $\mathbf{A}_{ij} = \mathbf{A}_{ji}$ for each $j \neq i$. The resulting $\mathbf{A}$ is sampled from the GOE, and we write $\mathbf{A} \sim \mathrm{GOE}(d)$.
The random version of the matrix discrepancy problem just replaces any matrix with our favorite distribution.

GOE-Average-case matrix discrepancy

Given n matrices sampled as $\{\mathbf{A}^{(i)}\}_{i=1}^n \sim \mathrm{GOE}(d)^{\otimes n}$ and a tolerance $\Delta \equiv \Delta(n, d) > 0$ find if there is a collection of signs $\{\epsilon_i\}_{i=1}^n \in \{\pm 1\}^n$ such that

$$\left\| \sum_{i=1}^n \epsilon_i \mathbf{A}^{(i)} \right\|_{\mathsf{op}} \leq \Delta, \tag{I.4}$$

in some "nice" probabilistic sense.

In what follows, we focus on this last problem, and omit the distribution since it will be fixed.

The paper has four other sections and an appendix. In section II we review the results for the random version. In section III we propose a method to build a solution for the planted version of the problem. In section IV we summarize. Appendix A is a walk-through of some ingredients of the main theorems. Throughout, we keep a rather constructive approach and propose non-working solutions commenting over them. The intention is to be as pedagogical as possible. We keep references to a minimum, explicit calculations, state clearly when we take a non-rigorous choice, and criticize at length our approach.

The cover image was made with DALL-E2 (link to website) with the prompt "A cartoon-style illustration of the discrepancy problem in random matrices, featuring a playful and colorful grid of matrix-like blocks".

II Recent developments

The average-case discrepancy problem is recent in literature. The first papers are from the end of 2024 (Kunisky and Zhang 2024; Maillard 2024). Below, we sum up the spirit of some results and their connections with other models.

II.I What is the natural order of tolerance?

Even if we have n matrices, the quickest analysis to make passes by the properties of a single matrix. The intuition behind the simplification is that the Gaussian law is stable with respect to sums with equal weights, and obeys nice limit theorems. Indeed, there is a well-defined limiting law of the spectrum that induces a result for the operator norm.

Since we work with symmetric matrices everything is simplified, and we can associate to each $\mathbf{A} \sim \mathrm{GOE}(d)$ its increasingly ordered real eigenvalues with multiplicities:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d. \tag{II.1}$$

We can study their global behavior of random eigenvectors as d grows large. One direct approach is via the empirical distribution, which we denote as:

$$\mu_{\mathbf{A}} := \frac{1}{d}\sum_{i=1}^{d} \delta_{\lambda_i}. \tag{II.2}$$

Interestingly in the limit, we have an almost sure result in the sense of probability distributions. Related to our problem, it directly implies convergence of the maximum eigenvalue, which in this case is the operator norm. For more details we refer to (Anderson, Alice Guionnet, and Zeitouni 2009; Bordenave and Chafai 2012).

- **Theorem 4** (Wigner's semicircle law and related)**.** *For* $\mathbf{A} \sim \mathrm{GOE}(d)$*, it holds that:*

$$\mu_{\mathbf{A}} \underset{d\to\infty}{\overset{\mathrm{w}}{\to}} \rho_{\mathsf{sc}}, \qquad \textit{almost surely,} \tag{II.3}$$

where the convergence is weak in the sense of distributions and ρ_{sc} *is the semicircular law, with density with respect to Lebesgue given by* $1/2\pi\sqrt{4-x^2}\delta_{[-2,2]}$.

Consequently, the operator norm converges with probability one to the edge of the semicircle law, i.e.:

$$\|\mathbf{A}\|_{\mathsf{op}} \underset{d\to\infty}{\to} 2, \qquad \textit{almost surely.} \tag{II.4}$$

We can use this theorem to find a "trivial regime" where the balancing is easy to find. Consider a n-dimensional sample $\{\mathbf{A}^{(i)}\}_{i=1}^n \sim \mathrm{GOE}(d)^{\otimes n}$. If we just sum all of them, we have that $\sum_{i=1}^n \mathbf{A}^{(i)} \sim \mathrm{GOE}(d,n)$, observing that a sum of GOE matrices is just a rescaled GOE, which follows by the classic property of Gaussians being stable. But then, rescaling $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{A}^{(i)} \sim \mathrm{GOE}(d)$, and theorem 4 holds. This implies that:

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{A}^{(i)} \right\|_{\mathsf{op}} \overset{d\to\infty}{\to} 2, \qquad \text{almost surely.} \tag{II.5}$$

Reordering terms, we can conclude that for a candidate tolerance $\Delta = \kappa\sqrt{n}$ the configuration $\epsilon_i \equiv 1$ for all i is such that:

$$\mathbb{P}\left[\left\| \sum_{i=1}^n \mathbf{A}^{(i)} \right\|_{\mathsf{op}} \leq \kappa\sqrt{n} \right] = \begin{cases} 1 - o_d(1) & \text{if } \kappa > 2 \\ \text{non-trivial} & \text{otherwise.} \end{cases} \tag{II.6}$$

In words, since the operator norm concentrates at 2, the $2\sqrt{n}$ threshold is fundamental; above, at $\kappa > 2$, an "all spin up" configuration has a good discrepancy, below, this is not guaranteed.

Proceeding step by step, the second-best configuration to try is uniformly random signs $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(\{\pm 1\})$. By symmetry of Gaussian distributions, it is analogous to conclude that $\sum_{i=1}^n \epsilon_i \mathbf{A}^{(i)} \sim \mathrm{GOE}(d,n)$, and the same reasoning shows that completely random signs do not improve the conclusion. To go beyond the semicircular behavior, we need to let $\Delta \equiv \Delta(n,d) = \kappa\sqrt{n}$ where $\kappa < 2$ at the very least. In particular, we will inspect the existence of non-trivial choices of signs.

II.II Vectors are again useful

Mathematical works suggested the extension of Spencer's theorem to matrices. In a related way, a line of research in statistical physics inspected the average-case vector version of discrepancy. Let us avoid the classical formulation and directly rewrite it in our setting. We sample n iid vectors $\mathbf{u}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and want to find a balancing with tolerance $\tilde{\Delta} \coloneqq K\sqrt{n}$. Since our question is about *existance*, it makes sense to introduce a random set that collects satisfying configurations:

$$\mathbf{V}_K \coloneqq \left\{ \epsilon \in \{\pm 1\}^n : \left\| \sum_{i=1}^n \epsilon_i \mathbf{u}^{(i)} \right\|_\infty \leq K\sqrt{n} \right\}. \tag{II.7}$$

Then, tuning (K,d,n), we seek an all-or-nothing phenomenon on $\#\mathbf{V}_K$. In other words, we study in the $\mathbb{R}^3_+$ plane choices of the three parameters such that the probability that the size is non-empty changes, possibly abruptly. It turns out that the roughest analysis one can make requires a lot of simplifications. Firstly, we do not let the three vary freely, but rather make n,d scale at the same rate, i.e. $n/d \to \beta > 0$. Then, we take a limit $d \to \infty$. This forces the conclusion to be non-asymptotic, but naturally easier to derive because the complexity is not on $\mathbb{R}^3_+$ but rather on $\mathbb{R}^2_+$ for (K,β) only. We then find:

• **Theorem 5** (Counting solutions of average-case vector discrepancy (Aubin, Perkins, and Zdeborová 2019; Abbe, Li, and Sly 2022)). *Let* $n/d \to \beta > 0$ *as* $d \to \infty$, *and* $K > 0$. *Define the critical value:*

$$\beta_{\mathsf{c}} \equiv \beta_{\mathsf{c}}(K) := {}^{-1}/_{\log 2} \log \mathbb{P}\left[|\mathrm{z}| \le K,\ \mathrm{z} \sim \mathcal{N}(0,1)\right]. \tag{II.8}$$

Then, the average case vector discrepancy problem has the following all-or-nothing behavior:

$$\lim_{n\sim\beta d,\ d\to\infty} \mathbb{P}[\#\mathbf{V}_K \ge 1] = \begin{cases} 0 & \text{if } \beta < \beta_{\mathsf{c}} \\ 1 & \text{if } \beta > \beta_{\mathsf{c}}. \end{cases} \tag{II.9}$$

In words, the number of configurations that attain $K\sqrt{n}$ *discrepancy in the proportional regime has a sharp phase transition in the* (β, K) *plane. If* β *is too small, there are no solutions. If* β *is large enough, there are solutions. Crucially, the notion of "too small" and "large enough" are quantitative at the critical value.*

Remark. *The critical value coincides with the transition of* $\mathbb{E}[\#\mathbf{V}_K]$ *from being exponentially small to exponentially large in size wrt the* (n, d) *pair. In some sense, we will see later that this closes the phase diagram in the proportional regime.*

From the theorem and the remark we borrow the conclusions and the technique, which is the first-second moment method. In what follows, we will present the main results of Maillard (2024).

II.III Drawing a partial phase diagram

The first step is noticing that the asymptotically proportional regime changes. The intuitive dimension of vectors is d, while for matrices it is d^2. We then anticipate that we will let $n/d^2 \to \tau > 0$ and expect a critical value $\tau_{\mathsf{c}} \equiv \tau_{\mathsf{c}}(\kappa)$ that depends on the κ tuning parameter of the discrepancy. The analogy is with $(\beta, \beta_{\mathsf{c}})$ of the previous section. However, the situation complicates. We will actually have *two* critical values, $\tau_{\mathsf{c1}}, \tau_{\mathsf{c2}}$ that encode the transition for the two probabilities. This is mainly because the proof technique does not cover the whole phase diagram and there is a dark region. Let us be more specific with statements. Throughout, $\kappa \in (0, 2]$, as otherwise the question is trivial. Also, we define the set of satisfying solutions just like we did for vectors:

$$\mathbf{M}_\kappa := \left\{ \boldsymbol{\epsilon} \in \{\pm 1\}^n : \left\| \sum_{i=1}^n \epsilon_i \mathbf{A}^{(i)} \right\|_{\mathrm{op}} \le \kappa\sqrt{n} \right\}. \tag{II.10}$$

We are now ready to restate the results of Maillard (2024, thms. 1.5, 1.7). Further details about the proof are in appendix A.

• **Theorem 6** (No solutions regime). *Let* $\tau_{\mathsf{c1}} \equiv \tau_{\mathsf{c1}}(\kappa)$ *be an explicit threshold found in (Maillard 2024, eqn. 4). Then sampling* $\{\mathbf{A}^{(i)}\}_{i=1}^n \sim \mathrm{GOE}(d)^{\otimes n}$ *it holds that:*

$$\lim_{n\sim\tau d^2,\ d\to\infty} \frac{1}{d^2} \log \mathbb{E}[\#\mathbf{M}_\kappa] = (\tau - \tau_{\mathsf{c1}}) \log 2. \tag{II.11}$$

In particular, if $\tau < \tau_{\mathsf{c1}}$*:*

$$\lim_{n\sim\tau d^2,\ d\to\infty} \mathbb{P}[\#\mathbf{M}_\kappa = 0] = 1. \tag{II.12}$$

In words, for fixed κ in the regime where $n/d^2 \sim \tau < \tau_{\mathsf{c1}}$, there are asymptotically almost surely no solutions to the discrepancy problem.

We stress that the expression for the no-solutions-threshold τ_{c1} is explicit. From previous discussions, we can also derive its value at the limits. For $\kappa \geq 2$, we expect easy solutions, and indeed it holds that $\tau_{\mathsf{c1}}(2) = 0$ so that any positive proportional rate of observations over complexity is sufficient. Similarly, as the $\kappa \to 0$, we will have $\tau_{\mathsf{c1}} \to \infty$, so that no complexity will satisfy the discrepancy required. The intuition here is also direct: as the discrepancy gets small, we expect to get out of the interesting regime $\sqrt{n}$ and the problem becomes essentially too hard to be interesting.

As previously mentioned, we do not have a clear all-or-nothing phenomenon like for vectors. The careful reader will have already noticed that the object we used is $\mathbb{P}[\#\mathbf{M}_\kappa = 0]$ rather than $\mathbb{P}[\#\mathbf{M}_\kappa \geq 1]$. The result for the latter is slightly more delicate. All depends on an implicit critical value τ_{c2} that solves an optimization problem found in (Maillard 2024, prop. 1.6). For the sake of simplicity, we avoid restating it here and just mention that it is *not* explicit. The result is analogous to one side of the vector version.

• **Theorem 7** (Solutions regime)**.** *Let $\tau_{\mathsf{c2}} \equiv \tau_{\mathsf{c2}}(\kappa)$ be the implicit threshold found in (Maillard 2024, eqn. 8). Then sampling $\{\mathbf{A}^{(i)}\}_{i=1}^n \sim \mathsf{GOE}(d)^{\otimes n}$ it holds that:*

$$\text{if } \tau > \tau_{\mathsf{c2}} \text{ then } \lim_{n\sim\tau d^2, d\to\infty} \mathbb{P}[\#\mathbf{M}_\kappa \geq 1] = 1. \tag{II.13}$$

In words, for fixed κ satisfying the condition on the complexity ratio n/d^2 there exists a solution asymptotically almost surely.

II.IV Known weaknesses in the results

Where is the catch? Why did we define two critical values and change the style of the statement? Comparing theorem 5 with theorems 6-7 we see that they look similar but not quite. With vectors, we were able to report a sharp transition on the same quantity depending on the natural complexity parameter β. For matrices instead, we found something more intricate. It turns out that the proof technique does not extend, and that with some improvements one can basically fill the negative answer up until a value τ_{c1} that does not coincide with τ_{c2} (to trust this, see the simulation in (Maillard 2024, fig. 1)). In other words, there is a region where the two theorems cannot predict the behavior. There are two further conclusions that we draw upon. Both mix formal results and non-rigorous predictions.

The first observation is that building up on different works we can say something more about the unknown region. This is the content of Maillard (2024, cor. 1.9).

• **Corollary 8** (Sharp threshold sequence existence)**.** *In the proportional regime where $n/d^2 \to \tau > 0$, where τ is large enough, there exists a sequence of tolerance parameters $\kappa_{\mathsf{c}}(d, \tau)$ such we have a sequential version of theorem 5. Namely, for any $\epsilon > 0$:*

$$\lim_{n\sim\tau d^2,\, d\to\infty} \mathbb{P}[\#\mathbf{M}_{\kappa_{\mathsf{c}}-\epsilon} \geq 1] = 0, \qquad \lim_{n\sim\tau d^2,\, d\to\infty} \mathbb{P}[\#\mathbf{M}_{\kappa_{\mathsf{c}}+\epsilon} \geq 1] = 1. \tag{II.14}$$

In words, there exists a sequence of tolerances such that slightly above or slightly below it we can replicate theorem 5, without a well-defined limit. By theorems 6-7, the tuple (τ, κ) must satisfy

the inclusion

$$\tau_{\mathsf{c1}}(\kappa_c) \leq \tau \leq \tau_{\mathsf{c2}}(\kappa), \tag{II.15}$$

therefore placing it in closure of the unknown region.

Remark. *Maillard (2024) also mentions that the limit is expected to exist from non-rigorous computations.*

Staying on the just mentioned non-rigorous computations, in a forthcoming work the author of (Maillard 2024) hints that the second-moment method used to identify τ_{c2} is not expected to be tight. On one side, the upper bounds for the second-moment part (i.e. theorem 7) are rough. On the other, there is evidence at the rigor of physics that the τ_{c1} threshold and any improved version of τ_{c2} will never coincide. For more details, we refer to (Maillard 2024, conj. 1.10) and the discussion around it.

III Analogy with planting vectors

One captivating question raised in (Maillard 2024, section 1) is about the properties of the solutions to the average case discrepancy problem. In practice, we may be interested in understanding what are the properties of $\mathbf{M}_\kappa$ as a whole set. What is the motivation? The intriguing reason is that theoretical results for the vector version predict that with measure one solutions are isolated and distant, but with a catch. Indeed, when solutions are isolated and distant it is natural to claim that algorithms will fail at finding them, by a simple "needle in a haystack" heuristic. Yet, empirical evidence showed that it is not hard to find solutions. This aspect is under investigation by the statistical physics community, but the current best explanation of the departure theory practice is that there are nearly measure-zero sets of well-connected solutions to which algorithms point in their dynamics. This is yet another example of inductive bias of classical algorithms. For more context on our brief explanation, the shortest argument in favor of the technique we found is in (Barbier et al. 2024), but the context is also largely discussed in (Zdeborová and Krzakala 2016, chap. II).

The first step for studying what happens with matrices is certainly not numerical, as the space to explore is combinatorial in nature. There are 2^n choices for signs, and one cannot hope to even find a solution and explore around it, unless there are many. What is more natural to do is to *inject* a solution of choice and proceed from there. This operation changes the problem, but the intuition is rather to simplify and see what happens to start understanding. In what follows, we try to replicate some results developed for vectors in (Barbier 2024; Barbier et al. 2024).

III.I What is the exact meaning of planting a solution?

We just said we want to "force" a solution. The starting point is to choose one. Let us begin with the reasoning for vectors. We first need to make the identification with the classical formulation

of the problem in statistical physics. From basic operations:

$$\left\|\sum_{i=1}^{n}\epsilon_i\mathbf{u}^{(i)}\right\|_\infty = \max_{1\le k\le d}\left|\sum_{i=1}^{n}\epsilon_i\mathbf{u}_k^{(i)}\right| = \max_{1\le k\le d}\left|\sum_{i=1}^{n}\epsilon_i\mathbf{g}_i^{(k)}\right| = \max_{1\le k\le d}\left|\left\langle \boldsymbol{s},\mathbf{g}^{(k)}\right\rangle\right|, \tag{III.1}$$

where we just "flipped" the vectors and got an expression in terms of an inner product with $\boldsymbol{s}=\{\epsilon_1,\dots,\epsilon_n\}$ a vector of signs. Here by "flipping" we mean that we force an inner product formulation by letting $\mathbf{g}$ be the transpose of $\mathbf{u}$, as done in (Maillard 2024). This is important to make matters easy. We will show that once the optimization is not jointly on $\{\mathbf{u}^{(i)}\}_{i=1}^n$ as on the left-hand side but separately on the $\{\mathbf{g}^{(k)}\}_{k=1}^d$ as on the right-hand side we can construct a solution. We now fix signs $\boldsymbol{s}^{(0)}$. Then, we sample a *biased* dataset as in (Barbier 2024, sec. 2):

$$\mathbf{g}^{(k)} = \frac{\mathrm{v}^{(k)}}{\sqrt{n}}\boldsymbol{s}^{(0)} + \mathbf{g}^{(k\perp)}, \qquad \mathbf{g}^{(k\perp)} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n - {}^1\!/_n\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top), \tag{III.2}$$

for all $1\le k\le d$ where $\mathrm{v}^{(k)}$ is a random variable supported on $[-K,K]$. The idea is that $\left\|\boldsymbol{s}^{(0)}\right\|_2^2 = n$ and that the random variable $\left\langle \mathbf{g}^{(k\perp)},\boldsymbol{s}^{(0)}\right\rangle$ is a Gaussian with mean zero and variance zero. Then,

$$\max_{1\le k\le n}\left|\left\langle \boldsymbol{s}^{(0)},\mathbf{g}^{(k)}\right\rangle\right| \le K\sqrt{n}, \tag{III.3}$$

and we satisfy the discrepancy scaling presented earlier.

Remark. *Notice that the variance of the perpendicular vector above is a projection, so it is idempotent and any power is the same. This will be used throughout the document.*

This approach is neat, and we wish to replicate it below but for matrices. The key ingredient was to pass from an *entangled* problem in the infinity norm to a maximization over *separate* inner products. The main hurdle is that the operator norm is not the infinity norm, and is somehow more implicit. Two methods are possible. We will showcase both, and eventually find out that only one barely works.

Method one Let us denote such solution $\boldsymbol{s}^{(0)}\in\{\pm1\}^n$. We may use the inequality for Hermitian matrices:

$$\|\boldsymbol{A}\|_{\mathsf{op}} \le \max_{1\le r\le d}\|\boldsymbol{A}_{r\cdot}\|_1\,, \tag{III.4}$$

where r stands for row. We see this as an application of a very nice theorem in linear algebra from (Potters and Bouchaud 2020, chap. 1).

• **Theorem 9** (Gershogin's circle theorem)**.** *Let $\boldsymbol{A}\in\mathbb{C}^{d\times d}$. We say a Gershogin disk is a ball centered at A_{ii} with radius $\sum_{j\neq i}|A_{ij}|$, or similarly for the row. Every eigenvalue of $\boldsymbol{A}$ lies within at least one of the Gershogin disks.*

For symmetric real matrices, the theorem holds with the specialty that row-column disks are the same, eigenvalues are real, and the operator norm is the maximum eigenvalue. We can then bound the maximum eigenvalue by the maximum circle:

$$\|\boldsymbol{A}\|_{\mathsf{op}} \le \max_{1\le r\le d}A_{rr} + \sum_{c\neq r}|A_{rc}| \le \max_{1\le r\le d}\|\boldsymbol{A}_{r\cdot}\|_1\,. \tag{III.5}$$

Reconnecting with the work on vectors:

$$\left\|\sum_{i=1}^{n} s_i^{(0)} \mathbf{A}^{(i)}\right\|_{\mathsf{op}} \leq \max_{1\leq r\leq d} \left\|\sum_{i=1}^{n} s_i^{(0)} \mathbf{A}_{r\cdot}^{(i)}\right\|_1 = \max_{1\leq r\leq d} \sum_{c=1}^{d} \left|\sum_{i=1}^{n} s_i^{(0)} \mathrm{A}_{rc}^{(i)}\right| = \max_{1\leq r\leq d} \sum_{c=1}^{d} \left|\left\langle \mathbf{g}^{(rc)}, \boldsymbol{s}^{(0)} \right\rangle\right|, \tag{III.6}$$

where we identified again a flipped version. Now, we want the *sum* of d absolute values to be less than our scaling. This could potentially go as follows. Let us sample iid variables:

$$\mathbf{g}^{(rc)} = \frac{\mathrm{w}^{(rc)}}{d\sqrt{n}} \boldsymbol{s}^{(0)} + \mathbf{g}^{(rc\perp)}, \qquad \mathbf{g}^{(rc\perp)} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n - {}^1\!/_n \boldsymbol{s}^{(0)} [\boldsymbol{s}^{(0)}]^\top), \tag{III.7}$$

for all $r, c \in \{1, \ldots, d\}$ where w is random and independent, from a distribution supported on $[-\kappa, \kappa]$.[1] Now, if we redo the computation, the sum readily cancels the d factor and for each row $1 \leq r \leq d$ we have a $\kappa\sqrt{n}$ bound ready.

Remark. *It is crucial here to realize that we are losing something with the inequality. As our representation is not tight, and $\leq$ might actually be a $<$ strict sign, we are dealing with a different problem. At the same time, we found no direct way to disentangle the matrices while keeping the operator norm. A good idea could be to just notice that since the inequality holds, then equality holds for some unknown $\lambda \in (0, 1]$. This unknown factor can go inside the norms and will just rescale variances by a constant, which makes no difference.*

The approach might be suspicious: as a matter of fact we are changing the problem. From looking for a solution given a distribution we take a solution and choose a very peculiar distribution that accommodates it. This is true, but the intuition is that the geometry of simpler problems ought to bring information for harder ones. There is also a body of work on this concept for a specific class of models. The principle goes by the name of "contiguity" and roughly states that if in the interpretation of thermodynamics the statistics are annealed then the planted model inherits the high probability properties of the non-planted model. For more information, we suggest checking the review (Zdeborová and Krzakala 2016), the lecture notes (Krzakala and Zdeborová 2021), and some recent works for the vector case (Aubin, Perkins, and Zdeborová 2019).

Method two Let us remind once again that for a given matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ it holds:

$$\|\boldsymbol{A}\|_{\mathsf{op}} \coloneqq \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_2, \qquad \|\boldsymbol{A}\|_\infty \coloneqq \max_{i,j} |A_{ij}|, \qquad \|\boldsymbol{A}\|_{\mathsf{op}} \leq \sqrt{d}\, \|\boldsymbol{A}\|_\infty. \tag{III.8}$$

The last inequality should already ring a bell: we can hope to get a $\sqrt{n}$ upper bound on the operator norm if we reduce the problem *directly* to the infinity norm. In our scaling, we have $n \sim d^2$ so $\sqrt{d} \sim n^{1/4}$. By the same principle, we will find that:

$$\left\|\sum_{i=1}^{n} s_i^{(0)} \mathbf{A}^{(i)}\right\|_{\mathsf{op}} \leq n^{1/4} \max_{1\leq r,c\leq d} \sum_{i=1}^{n} |s_i^{(0)} \mathrm{A}_{rc}^{(i)}| = n^{1/4} \max_{1\leq r,c\leq d} \left|\left\langle \boldsymbol{s}^{(0)}, \mathbf{g}^{(rc)} \right\rangle\right|. \tag{III.9}$$

[1] Momentarily, we ignore that on the off-diagonal there should be a two factor playing. This can probably be fixed but notice that it is not immediate how to make it nice. In some sense, we want $\mathbf{g}^{(cc)}$ to have twice the "variance" of $\mathbf{g}^{(rc)}$. Also, the global properties of the GOE ensemble do not care about this aspect. Even the right limit of the bulk of the spectrum is preserved at two (Anderson, Alice Guionnet, and Zeitouni 2009, thm. 2.1.22).

Therefore, we have a feasible configuration if:

$$\max_{1\leq r,c\leq d}\left|\left\langle \boldsymbol{s}^{(0)},\mathbf{g}^{(rc)}\right\rangle\right|\leq \kappa n^{1/4}. \tag{III.10}$$

Replicating the approach above, we generate d^2 vectors $\mathbf{g}^{(rc)}\in\mathbb{R}^n$ that satisfy a slightly different relation:

$$\mathbf{g}^{(rc)}=\frac{\mathrm{w}^{(rc)}}{n^{3/4}}\boldsymbol{s}^{(0)}+\mathbf{g}^{(rc\perp)},\qquad \mathbf{g}^{(rc\perp)}\sim\mathcal{N}(\mathbf{0}_n,\boldsymbol{I}_n,-1/n\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top). \tag{III.11}$$

A moment of thought reveals that the scaling are just correct to get the right $\kappa\sqrt{n}$ upper bound. Additionally, by $n\sim d^2$ the problem we study in both plantings will be of the same nature. The only apparent difference in the sampling is the scaling, but $n^{3/4}\sim d\sqrt{n}$

A non-method We can also check how other authors massaged the equations for matrices. In particular, Kunisky and Zhang (2024) deal with the same problem, in generic scaling (not proportional), and even derive an algorithm! While most of their results are not pertinent to our choice as they are sub-optimal (see (Maillard 2024, sec. 1)), some clever ideas remain. Adapting the notation and the scaling of (n,d), they found two bounds. From (Kunisky and Zhang 2024, cor. 2.5, eqn. 2.11) we have:

$$\min_{\epsilon_i\in\{\pm1\}\,\forall i}\left\|\sum_{i=1}^n\epsilon_i\mathbf{A}^{(i)}\right\|_{\mathsf{op}}\leq C\sqrt{d}\max_i\left\|\mathbf{A}^{(i)}\right\|_{\mathsf{F}}\leq\sqrt{d}\lambda_{\max}(\mathbf{M}), \tag{III.12}$$

where $\mathbf{M}$ is the gram matrix (w.r.t. the Frobenius inner product) of $\{\mathbf{A}^{(i)}\}_{i=1}^n$. However, we can see that we lost the handle of the problem as on both bounds the signs disappeared, and we cannot plant signs.

III.II Reasons why the planted solution is interesting

The following section is based largely on speculations from the vector problem. As a matter of fact, getting to a positive result would require extensive computations. For the sake of brevity, we summarize here an idea to draw conclusions about the behavior of the average-case matrix discrepancy problem from the point of view of its planted formulation. We draw inspiration from two works (Barbier et al. 2024; Barbier 2024).

In search for a discrepancy of $\kappa\sqrt{n}$ order, we consider a margin $\kappa_0\leq\kappa$. The natural distance in a hypercube is commonly termed Hamming distance, and just counts the number of different components between two vectors. From now onwards, we denote it as:

$$\mathrm{d}_{\mathsf{H}}(\boldsymbol{s},\boldsymbol{s}')\coloneqq\sum_{i=1}^n\delta_{s_i\neq s_i'}\in\{0,\dots,n\}. \tag{III.13}$$

We are interested in the geometry of solutions. Then, a natural object to inspect is the number of solutions that are close in distance to a given one. For notation purposes, we refer to the formulation in terms of vectors $\{\mathbf{g}^{(rc)}\}_{r,c=1}^d$. They lie in $\mathbb{R}^n$ and are stored in a matrix $\mathbf{G}\in\mathbb{R}^{d^2\times n}$, which upon flipping forms the matrices $\{\mathbf{A}^{(i)}\}_{i=1}^n$. Given a margin κ_0, and a matrix $\mathbf{G}$, the set of

solutions is again $\mathsf{M}_{\kappa_0} \equiv \mathsf{M}_{\kappa_0,\mathbf{G}}$ as in the previous sections. For a reference solution $\boldsymbol{s}^{(0)} \in \{\pm 1\}^n$, we can search others that are close via the (random) quantity:

$$\mathrm{z}(\boldsymbol{s}^{(0)}, \kappa_0, r) \coloneqq \#\left\{\boldsymbol{s} \in \mathsf{M}_{\kappa_0,\mathbf{G}} \mid \mathrm{d}_{\mathsf{H}}(\boldsymbol{s}^{(0)}, \boldsymbol{s}) = nr\right\}. \tag{III.14}$$

Remark. *The n scaling in the distance is forced to derive a result that will be in the line of "we need to flip $\Omega(n)$ signs to get to the closest solution, hence solutions are isolated". This justifies the choice above.*

We then define the local entropy of our problem at rate (n, δ) and distance r as the average number of solutions that are nr close to any other. Here, the $\delta > 0$ parameter rules out cases in which the logarithm is sub-linear. Mathematically we adapt (Barbier et al. 2024, eqn. 6):

$$\phi_{n,\delta}(r) \coloneqq \frac{1}{n}\mathbb{E}_{\mathbf{G}}\left[\frac{1}{\#\mathsf{M}_{\kappa_0,\mathbf{G}}}\sum_{\mathbf{s}^{(0)}\in\mathsf{M}_{\kappa_0,\mathbf{G}}} \log_{n\delta}\mathrm{z}(\mathbf{s}^{(0)}, \kappa_0, r) \mid \mathsf{M}_{\kappa_0,\mathbf{G}} \neq 0\right], \tag{III.15}$$

where we define the truncated log as:

$$\log_{n\delta}(x) \coloneqq \max\{\log(x), n\delta\}. \tag{III.16}$$

For vectors, Barbier et al. (2024) show that under certain assumptions the behavior of this local entropy is characterized by the annealed entropy of the planted model. From this and other considerations already mentioned about the concept of contiguity, we will study the planted version. Some words of caution are needed here to unravel all the terminologies. The result is in the limit of a precise scaling, hence being non-asymptotic. Secondly, it will get rid of δ, by letting it go to zero. Moreover, the behavior will also depend on how we change the margin κ, which plays the role of a temperature. The full statement with context is (Barbier et al. 2024, thm. 1).

The term "annealed" originates in physics, and is life-saving for us. When dealing with a random problem, if we want to access its statistics, it is natural to inspect the partition function or log-partition function, also termed cumulant generating function. Early in the process, one encounters the issue that its integral is a black box: the form is essentially $\mathbb{E}[\log\sum]$, where the sum is over a large space. In statistical physics literature, it is termed quenched average. What comes to the rescue is the annealed version, which in plain mathematical terms is applying Jensen's inequality and studying $\log\mathbb{E}\sum$. While only an inequality holds, it gives non-trivial information on the problem, as it is basically a first-moment bound. Moreover, in some low temperature regimes, it is expected to be the right behavior of the problem. We reroute the reader to pedagogical works on the intersection of machine learning and thermodynamics for more precise statements (Mezard and Montanari 2009; Krzakala and Zdeborová 2021).

III.III Computation of the planted annealed local entropy

While we could proceed with further justification of the intention to study the annealed entropy of the planted problem, we make a leap of faith; it is the simplest thing we can try, so we just go on with it. For further technical aspects, the reader can check (Barbier et al. 2024), and the references therein. From now onwards, we take this for granted and also presume that a certain first (conditional) moment we will use is the right object to consider following the reasoning of

Barbier et al. (2024, thm. 3), where they also prove that its variance vanishes. Then, we will postulate that such annealed local entropy is faithful in certain regimes of interest, and study its sign perturbatively, as done in (Barbier 2024, sec. 2).[2] If said sign is negative, we will guess that the solutions are typically isolated in the planted model, and hence too in the original model. Let us unravel once again the heuristic route. We:

1. had a random problem over matrices;
2. rewrote it in terms of a certain collection of Gaussian vectors stored in a matrix $\mathbf{G}$;
3. formulated a planted version, where we choose one vector that is a solution and can look around it;
4. came back to the original problem and proposed a quantity that counts neighboring solutions;
5. believed in the possibility of contiguity between planted and original problem, for which roughly high probability events are shared;
6. believed in the idea that the local entropy of the original problem is well approximated in certain regimes by the annealed local entropy of the planted model;
7. having an explicit formula for said last object, we can simulate it. If its sign will be negative at small distances then we will conclude that solutions are isolated.

It is clear that many passages are rough, but hopefully, the thread of the idea is valuable. In step 2 we lose something by the inequality, in steps 5 & 6 we presume that some results could be proved, but stress that nevertheless what we do is often the first attempt. In step 7, if the formula will be nice enough, it will be numerically stable.

Remark. *As a last point, why would a negative log-expectation tell us that solutions are isolated? Roughly, this is because if we count in log-scale some large object, a negative number indicates that there are no solutions at order $o(n)$. In other words, we need to flip about $\Omega(n)$ signs from a planted solution to find another. In the larger sense, this is the principle of overlap-gap property (see (Gamarnik et al. 2022) and the line of work of the first author).*

Let us proceed by the reasoning of (Barbier et al. 2024, thm. 3). We sample as in equation III.7, and specify that the random margins $\mathrm{w}^{(rc)}$ are Gaussians conditioned on being less than κ_0. Their p.d.f. is:

$$f(\omega) \coloneqq \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\omega^2/2} \delta_{|\omega| \leq \kappa_0}}{\mathbb{P}[|\mathrm{z}| \leq \kappa_0]}, \qquad \mathrm{z} \sim \mathcal{N}(0,1). \tag{III.17}$$

Remark. *It is important to notice that Barbier (2024) claims that the local entropy is non-zero when the margin distribution satisfies the condition $f(\kappa) + f(-\kappa) \neq 0$ which holds here if we let $\kappa_0 = \kappa$.*

The planted local entropy reads (Barbier et al. 2024, eqn. 17):

$$\phi^{\mathsf{pl}}_{n,\delta}(r) \coloneqq \frac{1}{n} \mathbb{E}_{\mathsf{pl}} \left[\log_{n\delta} \mathrm{z}(\mathbf{s}^{(0)}, \kappa, r) \right]. \tag{III.18}$$

[2] In statistical physics this regime is the annealed replica: it is an ansatz on the optimal solutions of a large optimization problem that appears when doing replica computations. The intention of this document is not to be an exposé of this magical computation. The original works of Parisi (Parisi 1979a; Parisi 1979b; Parisi 1983) lead to a Nobel prize in Physics. For a more computation oriented perspective one can consult (Krzakala and Zdeborová 2021), for a review of the potential (Zdeborová and Krzakala 2016). It was in particular applied in the vector case (Aubin, Perkins, and Zdeborová 2019; Barbier et al. 2024; Barbier 2024) and it should be one of the ingredients of the working paper mentioned by Maillard (2024).

In particular, the expectation is with respect to the random quantities:

- $\mathbf{s}^{(0)}$, uniform on the hypercube $\{\pm 1\}^n$;
- $\mathbf{G}$, which conditional on $\mathbf{s}^{(0)}$ is sampled as in equation III.7.

As said previously, this is rather hard to compute in general; it is the expectation of a log of a sum. We believe that in appropriate cases the annealed entropy is interesting to compute. It has the easier expression:

$$\phi_n^{\mathsf{pl;ann}}(r) \coloneqq \frac{1}{n} \log \mathbb{E}[\mathrm{z}(\mathbf{s}^{(0)}, \kappa, r)]. \tag{III.19}$$

Notice that now we can sum over $\mathbf{s}^{(0)}$ and just compute the randomness w.r.t. the number of solutions that are close to it, and that we blindly took $\delta \to 0$. In this perspective, the annealed assumption is like saying that "any configuration behaves like the other". This is one of the reasons why it is expected to be not always true: in certain regimes, the properties of the problem will break symmetries. Again, for more context, full statements for the vector case are mostly in (Barbier et al. 2024). For more standard theory, we refer to (Mezard and Montanari 2009; Krzakala and Zdeborová 2021).
Now, we argue that conditional on $\mathbf{w} \in \mathbb{R}^{d^2}$, the argument inside the logarithm converges almost surely to a deterministic limit. We have by linearity of expectation:

$$\mathbb{E}[\mathrm{z}(\boldsymbol{s}^{(0)}, \kappa, r) \mid \boldsymbol{w}] = \sum_{\boldsymbol{s} \in \{\pm 1\}^n} \mathbb{E}\left[\mathbb{1}\left\{\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}}, |\left\langle \boldsymbol{s}, \mathbf{s}^{(0)} \right\rangle| = nr \mid \boldsymbol{w}\right\}\right], \tag{III.20}$$

now it is natural to let the inner product express its sign: we parameterize $r = {}^{(1-m)}/_2$ for $m \in (-1, 1)$, and write:

$$\mathbb{E}[\mathrm{z}(\boldsymbol{s}^{(0)}, \kappa, r) \mid \boldsymbol{w}] = \sum_{\boldsymbol{s} \in \{\pm 1\}^n} \mathbb{E}\left[\mathbb{1}\left\{\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}}, \left\langle \boldsymbol{s}, \mathbf{s}^{(0)} \right\rangle = nm \mid \boldsymbol{w}\right\}\right]. \tag{III.21}$$

The expectation is partly w.r.t $\mathbf{s}^{(0)}$, which is uniform over the hypercube with equal weights ${}^1/_{2^n}$ so we can add a second sum over the hypercube and recognize that the indicator splits into two indicators. In equations now we have fixed $\boldsymbol{s}^{(0)}$:

$$\mathbb{E}[\mathrm{z}(\boldsymbol{s}^{(0)}, \kappa, r) \mid \boldsymbol{w}] = \frac{1}{2^n} \sum_{\boldsymbol{s} \in \{\pm 1\}^n} \sum_{\boldsymbol{s}^{(0)} \in \{\pm 1\}^n} \mathbb{E}\left[\mathbb{1}\{\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}}\} \mathbb{1}\left\{\left\langle \boldsymbol{s}, \boldsymbol{s}^{(0)} \right\rangle = nm\right\} \mid \boldsymbol{w}, \boldsymbol{s}^{(0)}\right]. \tag{III.22}$$

The second indicator is now deterministic so we can take it out, and it remains to compute the probability that $\boldsymbol{s}$ is in $\mathsf{M}_{\kappa,\mathbf{G}}$ given a reference planted solution $\boldsymbol{s}^{(0)}$ and reference margins $\boldsymbol{w}$. This might be complicated, but we also remind that the indicator outside forces us to consider configurations that are exactly at nm distance from $\boldsymbol{s}^{(0)}$ in each term. We will see that it boils down to some computations that use the sampling model of equation III.7. We have:

$$\mathbb{E}[\mathrm{z}(\boldsymbol{s}^{(0)}, \kappa, r) \mid \boldsymbol{w}] = \frac{1}{2^n} \sum_{\boldsymbol{s} \in \{\pm 1\}^n} \sum_{\boldsymbol{s}^{(0)} \in \{\pm 1\}^n} \mathbb{1}\left\{\left\langle \boldsymbol{s}, \boldsymbol{s}^{(0)} \right\rangle = nm\right\} \mathbb{P}\left[\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}} \mid \boldsymbol{w}, \boldsymbol{s}^{(0)}\right], \tag{III.23}$$

with:

$$\mathbb{P}\left[\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}} \mid \boldsymbol{w}, \boldsymbol{s}^{(0)}\right] = \prod_{1 \le r \le d} \mathbb{P}\left[\sum_{c=1}^{d} \left|\left\langle \mathbf{g}^{(rc)}, \boldsymbol{s} \right\rangle\right| \le \kappa\sqrt{n} \mid (w^{(r1)}, \dots, w^{(rd)}), \boldsymbol{s}^{(0)}\right]. \tag{III.24}$$

In particular, we rewrite equation III.7 as:

$$\mathbf{g}^{(rc)} = \frac{\mathrm{w}^{(rc)}}{d\sqrt{n}}\boldsymbol{s}^{(0)} + \left[\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top\right]\widetilde{\mathbf{g}}^{(rc)}, \qquad \widetilde{\mathbf{g}}^{(rc)} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n), \tag{III.25}$$

So the inner products inside the probability take form (putting the $\sqrt{n}$ factor on the left-hand side of the inequality):

$$\frac{1}{\sqrt{n}}\left\langle \mathbf{g}^{(rc)}, \boldsymbol{s}\right\rangle = \frac{\mathrm{w}^{(rc)}}{dn}\left\langle \boldsymbol{s}^{(0)}, \boldsymbol{s}\right\rangle + \frac{1}{\sqrt{n}}\left\langle \left[\boldsymbol{I}_n - 1/n\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top\right]\widetilde{\mathbf{g}}^{(rc)}, \boldsymbol{s}\right\rangle \tag{III.26}$$

$$= \frac{1}{d}\frac{\mathrm{w}^{(rc)}}{dn}nm + \frac{1}{\sqrt{n}}\left\langle \widetilde{\mathbf{g}}^{(rc)}, \left[\boldsymbol{I}_n - 1/n\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top\right]^\top \boldsymbol{s}\right\rangle \quad \text{by the indicator in the sum;} \tag{III.27}$$

$$= \frac{1}{d}\mathrm{w}^{(rc)}m + \frac{1}{\sqrt{n}}\left\langle \widetilde{\mathbf{g}}^{(rc)}, \left[\boldsymbol{I}_n - 1/n\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top\right]\boldsymbol{s}\right\rangle \quad \text{by symmetry of the matrix;} \tag{III.28}$$

$$= \frac{1}{d}\mathrm{w}^{(rc)}m + \frac{1}{\sqrt{n}}\left\langle \widetilde{\mathbf{g}}^{(rc)}, \boldsymbol{s} - m\boldsymbol{s}^{(0)}\right\rangle \quad \text{again by the indicator in the sum.} \tag{III.29}$$

The inner product has a Gaussian distribution, since $\widetilde{\mathbf{g}}^{(rc)}$ is a standard Gaussian vector. The mean is zero. The variance is then the second moment, which is a standard computation since the vector has a diagonal covariance matrix. We make it explicit below:

$$\mathbb{E}\left[\left\langle \widetilde{\mathbf{g}}^{(rc)}, \boldsymbol{s} - m\boldsymbol{s}^{(0)}\right\rangle^2\right] = \sum_{i,j=1}^{n}\mathbb{E}\left[\mathrm{g}_i^{(rc)}\mathrm{g}_j^{(rc)}(s_i - ms_i^{(0)})(s_j - ms_j^{(0)})\right] \tag{III.30}$$

$$= \sum_{i=1}^{n}(s_i - ms_i^{(0)})^2 \tag{III.31}$$

$$= \left\|\boldsymbol{s} - m\boldsymbol{s}^{(0)}\right\|^2 \tag{III.32}$$

$$= \|\boldsymbol{s}\|^2 + m^2\left\|\boldsymbol{s}^{(0)}\right\|^2 - 2m\left\langle \boldsymbol{s}, \boldsymbol{s}^{(0)}\right\rangle \tag{III.33}$$

$$= n + m^2n - 2m^2n \tag{III.34}$$

$$= n - m^2n, \tag{III.35}$$

where we used that the vectors have inner product mn and that the norm of a vector in the hypercube is n. The normalization by $1/\sqrt{n}$ in the inner product makes it overall a centered Gaussian with variance $1 - m^2$. In other words, we can write:

$$\frac{1}{\sqrt{n}}\left\langle \widetilde{\mathbf{g}}^{(rc)}, \boldsymbol{s} - m\boldsymbol{s}^{(0)}\right\rangle \overset{\mathrm{d}}{=} \sqrt{1 - m^2}\mathrm{z}^{(rc)}, \qquad \mathrm{z}^{(rc)} \sim \mathcal{N}(0, 1), \tag{III.36}$$

where $\mathrm{z}^{(rc)}$ is independent of the other random variables. We know reach a fundamental hurdle that will be solved by a switch. In practice, we found that:

$$\mathbb{P}\left[\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}} \mid \boldsymbol{w}, \boldsymbol{s}^{(0)}\right] = \prod_{1\le r\le d}\mathbb{P}\left[\sum_{c=1}^{d}\left|\left\langle \mathbf{g}^{(rc)}, \boldsymbol{s}\right\rangle\right| \le \kappa\sqrt{n} \mid (w^{(r1)}, \dots, w^{(rd)}), \boldsymbol{s}^{(0)}\right] \tag{III.37}$$

$$= \prod_{1\le r\le d}\mathbb{P}\left[\sum_{c=1}^{d}\left|\frac{1}{d}w^{(rc)}m + \sqrt{1 - m^2}\mathrm{z}^{(rc)}\right| \le \kappa \mid (w^{(r1)}, \dots, w^{(rd)})\right]. \tag{III.38}$$

What we would like to say is the following. First, we plug our expressions into the conditional,

$$\frac{1}{n}\log\mathbb{E}\left[\mathrm{z}(\boldsymbol{s}^{(0)},\kappa,m)\mid \boldsymbol{w}\right] \tag{III.39}$$

$$=\frac{1}{n}\log\left(\frac{1}{2^n}\sum_{\boldsymbol{s},\boldsymbol{s}^{(0)}\in\{\pm1\}^n}\mathbb{1}\left\{\left\langle \boldsymbol{s},\boldsymbol{s}^{(0)}\right\rangle=nm\right\}\prod_{1\leq r\leq d}\mathbb{P}\left[\sum_{c=1}^{d}\left|\frac{1}{d}w^{(rc)}m+\sqrt{1-m^2}\mathrm{z}^{(rc)}\right|\leq\kappa\mid(w^{(r1)},\dots,w^{(rd)})\right]\right) \tag{III.40}$$

While apparently monstrous, this expression is not super bad. First of all, the normalization factor can be placed outside. Then, the double sum inside is independent of the probability, so we eventually have:

$$\frac{1}{n}\log\mathbb{E}\left[\mathrm{z}(\boldsymbol{s}^{(0)},\kappa,m)\mid \boldsymbol{w}\right]=\frac{1}{n}\sum_{1\leq r\leq d}\log\left(\mathbb{P}\left[\sum_{c=1}^{d}\left|\frac{1}{d}w^{(rc)}m+\sqrt{1-m^2}\mathrm{z}^{(rc)}\right|\leq\kappa\mid(w^{(r1)},\dots,w^{(rd)})\right]\right) \tag{III.41}$$

$$+\frac{1}{n}\log\left(\frac{1}{2^n}\sum_{\boldsymbol{s},\boldsymbol{s}^{(0)}\in\{\pm1\}^n}\mathbb{1}\left\{\left\langle \boldsymbol{s},\boldsymbol{s}^{(0)}\right\rangle=nm\right\}\right). \tag{III.42}$$

Barbier et al. (2024) now do the following. They simplify with Stirling's formula the first term (for more details, one might check (Krzakala and Zdeborová 2021, sec. 1.1)), and find that:

$$\frac{1}{n}\log\left(\frac{1}{2^n}\sum_{\boldsymbol{s},\boldsymbol{s}^{(0)}\in\{\pm1\}^n}\mathbb{1}\left\{\left\langle \boldsymbol{s},\boldsymbol{s}^{(0)}\right\rangle=nm\right\}\right)=\mathcal{H}(m)+o_n(1), \tag{III.43}$$

where we wrote Shannon's binary entropy,

$$\mathcal{H}(m)\coloneqq-\frac{(1+m)}{2}\log\left(\frac{1+m}{2}\right)-\frac{(1-m)}{2}\log\left(\frac{1-m}{2}\right). \tag{III.44}$$

Instead, the first term should converge by the law of large numbers as $n\to\infty$ with $n\sim\tau d^2$ almost surely to its expectation, which in this case would be w.r.t. the $(w^{(r1)},\dots,w^{(rd)})$ random variable; some truncated Gaussians. However, there are two issues:

- the normalization is $1/n$, but the sum is over d elements; in our scaling, we have $n\sim\tau d^2$ so this is not really a normalized sum;
- the probability inside is yes a probability, but it is not nice w.r.t. the vector $(w^{(r1)},\dots,w^{(rd)})$, so what is its limit?

We resolve this perspective by changing the way we plant a solution. Indeed, we presented another possibility to build a solution around a given vector, that used another inequality, i.e. **Method two**. It also had a different normalization, and we will need just a minimal (but potentially not nice) adjustment. Many equations are the same, what we need to fix is the way we evaluate the probability of having a neighbor solution. Indeed, sampling as in equation III.11 we may write:

$$\mathbf{g}^{(rc)}=\frac{\mathrm{w}^{(rc)}}{n^{3/4}}\boldsymbol{s}^{(0)}+\left[\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top\right]\widetilde{\mathbf{g}}^{(rc)},\qquad\widetilde{\mathbf{g}}^{(rc)}\sim\mathcal{N}(\mathbf{0}_n,\boldsymbol{I}_n). \tag{III.45}$$

Notice that we have a different normalization in the first term, which is because the constraint we want to satisfy is that of equation III.10 and it loses a $n^{1/2}$ factor in the inequality. If we write the analog of equation III.24 we have:

$$\mathbb{P}\left[\boldsymbol{s} \in \mathsf{M}_{\kappa,\mathbf{G}} \mid \boldsymbol{w}, \boldsymbol{s}^{(0)}\right] = \prod_{1\le r\le d}\prod_{1\le c\le d} \mathbb{P}\left[\left|\left\langle \mathbf{g}^{(rc)}, \boldsymbol{s}\right\rangle\right| \le \kappa n^{1/4} \mid w^{(rc)}, \boldsymbol{s}^{(0)}\right]. \tag{III.46}$$

The expression above is better for two reasons:

- it is rightfully a product of $d^2 \sim \frac{1}{\tau}n$ terms, which once we take the log will allow us to apply the law of large numbers;
- inside there is no sum, so we might hope to get something that converges to constant random variables like before.

However, it is not nice all the way until the end. Replicating the computation for the inner product inside the probability, we find:

$$\frac{1}{n^{1/4}}\left\langle \mathbf{g}^{(rc)}, \boldsymbol{s}\right\rangle = w^{(rc)}m + \frac{1}{n^{1/4}}\left\langle \widetilde{\mathbf{g}}^{(rc)}, \boldsymbol{s} - m\boldsymbol{s}^{(0)}\right\rangle. \tag{III.47}$$

Again, the first term is nice, even better than before, but the second diverges as $n \to \infty$ since the normalization is not enough.[3] We fix this last issue by decreasing the variance. Let us then propose a third way to sample planted solutions:

$$\mathbf{g}^{(rc)} = \frac{w^{(rc)}}{n^{3/4}}\boldsymbol{s}^{(0)} + \frac{1}{n^{1/4}}\left[\boldsymbol{I}_d - \frac{1}{n}\boldsymbol{s}^{(0)}[\boldsymbol{s}^{(0)}]^\top\right]\widetilde{\mathbf{g}}^{(rc)}, \qquad \widetilde{\mathbf{g}}^{(rc)} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n). \tag{III.48}$$

In words, we add just what we need: a normalizing factor on the perpendicular component.

Remark. *Notice that this adjustment is not great in the sense that as $n \to \infty$ we will sample with an orthogonal Gaussian bias that becomes small, albeit slowly. This might be an indication that this planted model is not the right one to observe contiguity and that we might need better ways to not lose factors with an inequality on the operator norm.*

Borrowing the reasoning in the proof of (Barbier et al. 2024, thm. 3), we may conclude that in our proportional scaling seeing the distance $r = {}^{(1-m)}/_2$ to account for symmetries:

$$\phi_n^{\mathsf{pl;ann}}(r) = \frac{1}{n}\log \mathbb{E}\left[\mathrm{z}(\boldsymbol{s}^{(0)}, \kappa, m) \mid \boldsymbol{w}\right] \tag{III.49}$$

$$= \mathcal{H}(m) + \frac{1}{n}\sum_{1\le r\le d}\sum_{1\le c\le d}\log\left(\mathbb{P}\left[\left|w^{(rc)}m + \sqrt{1-m^2}\mathrm{z}^{(rc)}\right| \le \kappa \mid w^{(rc)}\right]\right) + o_n(1) \tag{III.50}$$

$$\sim \mathcal{H}(m) + \frac{1}{\tau}\frac{1}{d^2}\sum_{1\le r\le d}\sum_{1\le c\le d}\log\left(\mathbb{P}\left[\left|w^{(rc)}m + \sqrt{1-m^2}\mathrm{z}^{(rc)}\right| \le \kappa \mid w^{(rc)}\right]\right) + o_n(1) \tag{III.51}$$

$$\underset{n\to\infty, n\sim\tau d^2}{\overset{\text{a.s.}}{\to}} \mathcal{H}(m) + \frac{1}{\tau}\mathbb{E}_{\mathrm{w}}\left[\log\left(\mathbb{P}\left[\left|\mathrm{w}m + \sqrt{1-m^2}\mathrm{z}\right| \le \kappa \mid \mathrm{w}\right]\right)\right] \tag{III.52}$$

$$:= \phi^{\mathsf{pl;ann}}(m). \tag{III.53}$$

[3] Proceeding as before: if we compute the variance it is not of constant order, but rather $\propto \sqrt{n}$.

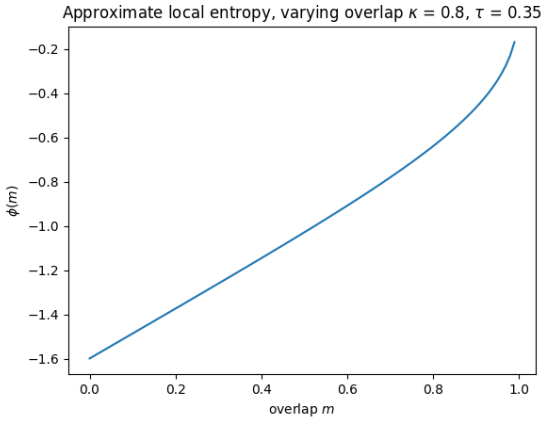


Figure 1: Local entropy approximation plot.
We remark that this function is a good approximation for $m \approx 1$.

by an application of the law of large numbers to the probabilities inside, which are random w.r.t. the collection of $\mathrm{w}^{(rc)}$ that are independent and identically distributed with support on $[-\kappa_0, \kappa_0]$. In particular, the probability is for z being random and i.i.d. conditioned on the value of w, then we take the log, then we integrate w.r.t. w. This expression is the same as the one of Barbier (2024, eqn. A.12) and Barbier et al. (2024, eqn. 24, eqn. 8). Additionally, we can compute its Taylor expansion around $m = 1$ which counts "close in order n" solutions. Reporting (Barbier 2024, eqn. A.13) as is:

$$\phi^{\mathsf{pl;ann}}(m) \overset{m\approx 1}{\approx} -\frac{(1-m)}{2}\log\left(\frac{1-m}{2}\right) + \frac{2}{\tau}\sqrt{1-m}\,[f(\kappa)+f(-\kappa)]\int_0^\infty \log\left(\frac{1+\mathrm{erf(x)}}{2}\right)\mathrm{d}x. \tag{III.54}$$

In particular, if $\kappa_0 < \kappa$ then it is always negative, while if $\kappa_0 = \kappa$, we can check numerically that there is a range of values $(m_0, 1)$ such that it is negative. We propose two quick simulations in figures 1-2.

IV Conclusion and perspectives

We have only scratched the surface, but will try to review our attempt and infer the next steps. Starting from an interesting conjecture that would extend non-asymptotic results of vectors to matrices, we presented the average-case matrix discrepancy problem. It was very recently formulated in two works (Maillard 2024; Kunisky and Zhang 2024). For the sake of space, we ignored the second, and briefly reviewed the interesting results in the first. The current sketch of the phase diagram tells us that in the proportional scaling of $n \sim \tau d^2$ there is a threshold at which solutions do not appear, and another at which solutions appear. However, the two seem to be distinct and not possible to join with previous methods. This last aspect is the main difference with the vector case, where the moment method fills the phase diagram.
Nevertheless, it appears that some techniques (but not all!) can be transferred from the analysis made for vectors. The last comments of Maillard (2024) hint at some potential directions. One

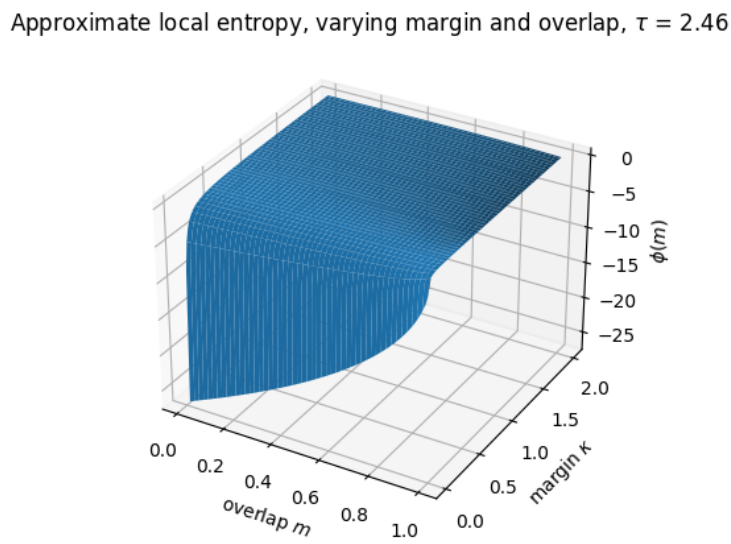


Figure 2: Local entropy approximation in (m, κ) plane.
We remark that this function is a good approximation for $m \approx 1$. Also, in 3 dimensions it is not nice at all, but it is always negative.

of these is understanding better the *geometry* of solutions, meaning how the good signs are arranged in the hypercube. The first question in this case is how close they are.

Believing in the potential of a parallel, we took more practical inspiration from earlier works on vectors. The methodology which extends well over this case, is to analyze a companion problem where can choose to inject a solution: this is the principle of planting. We propose some ways to plant and analyze the expectation of a certain object that is supposed to count close solutions. We find that in regimes of closeness (i.e. $m \approx 1$), there are no solutions. This suggests in layman's terms that solutions are isolated to order $\Omega(n)$.

Weaknesses Our approach needs critique for many reasons, with no regard to completeness, we overview some below. Firstly, we plant with an inequality: we could not find an exact representation of the problem over which we could plant, so we found a solution for an upper bound on our objective quantity and then wrote down a model that satisfied the original requirement. As a matter of fact, there might be a way to write a model that satisfies directly the operator norm bound. Secondly, we glossed over many aspects of the steps needed to conclude that a planted model is interesting at all, and just proceeded by the principle of doing the first feasible thing.

What is next? At the same time, there are many ways to improve our understanding. Drawing on the methods of Barbier et al. (2024), we could show that the planted local entropy is really a good approximation of the local entropy of the original problem, and that in the right regime, the annealed local entropy we found is again faithful. If this were true, we would again "rigorously" conclude that there are only isolated solutions in the average-case matrix discrep-

ancy problem, in line with the result for vectors. Here, by "rigorously", we mean that the works we relied on either use physics methods (e.g. (Barbier 2024)) or a mix of physics, proofs, and unproven assumptions (e.g. (Barbier et al. 2024)). As always, this is the first step. The next best thing would be closing the argument formally as done in (Abbe, Li, and Sly 2022) for vectors, but we did not focus much on this second step here.

From the practical point of view, pushing on the development of (Barbier et al. 2024) looks like the natural choice to make, while keeping an eye on the incoming publication mentioned by Maillard (2024) that might include some new aspects of the problem. Once we will have understood the model from the point of view of physics, it will be the time for non-asymptotic refinements, and possibly more information about the original deterministic problem.

References

Maillard, Antoine (Oct. 2024). *Average-Case Matrix Discrepancy: Satisfiability Bounds*. DOI: 10.48550/arXiv.2410.17887. arXiv: 2410.17887 [math]. (Visited on 12/26/2024).

Bandeira, Afonso S (2024). *Welcome, and Matrix Discrepancy (Problems 1-2) | Randomstrasse 101*. (Visited on 01/28/2025).

Spencer, Joel (1985). "Six Standard Deviations Suffice". In: *Transactions of the American Mathematical Society* 289.2, pp. 679–706. ISSN: 0002-9947. DOI: 10.2307/2000258. JSTOR: 2000258. (Visited on 01/25/2025).

Kunisky, Dmitriy and Peiyuan Zhang (July 2024). *Average-Case Matrix Discrepancy: Asymptotics and Online Algorithms*. DOI: 10.48550/arXiv.2307.10055. arXiv: 2307.10055 [cs]. (Visited on 01/19/2025).

Anderson, Greg W., Alice Guionnet, and Ofer Zeitouni (2009). *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press. ISBN: 978-0-521-19452-5. DOI: 10.1017/CBO9780511801334. (Visited on 01/25/2025).

Bordenave, Charles and Djalil Chafai (Jan. 2012). "Around the Circular Law". In: *Probability Surveys* 9.none. ISSN: 1549-5787. DOI: 10.1214/11-PS183. arXiv: 1109.3343 [math]. (Visited on 01/28/2025).

Aubin, Benjamin, Will Perkins, and Lenka Zdeborová (July 2019). "Storage Capacity in Symmetric Binary Perceptrons". In: *Journal of Physics A: Mathematical and Theoretical* 52.29, p. 294003. ISSN: 1751-8113, 1751-8121. DOI: 10.1088/1751-8121/ab227a. arXiv: 1901.00314 [cond-mat]. (Visited on 01/25/2025).

Abbe, Emmanuel, Shuangping Li, and Allan Sly (Feb. 2022). "Proof of the Contiguity Conjecture and Lognormal Limit for the Symmetric Perceptron". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 327–338. DOI: 10.1109/FOCS52979.2021.00041. (Visited on 01/19/2025).

Barbier, Damien et al. (Apr. 2024). "On the Atypical Solutions of the Symmetric Binary Perceptron". In: *Journal of Physics A: Mathematical and Theoretical* 57.19, p. 195202. ISSN: 1751-8121. DOI: 10.1088/1751-8121/ad3a4a. (Visited on 01/19/2025).

Zdeborová, Lenka and Florent Krzakala (Sept. 2016). "Statistical Physics of Inference: Thresholds and Algorithms". In: *Advances in Physics* 65.5, pp. 453–552. ISSN: 0001-8732, 1460-6976. DOI: 10.1080/00018732.2016.1211393. arXiv: 1511.02476 [cond-mat, stat]. (Visited on 09/27/2022).

Barbier, Damien (Aug. 2024). *How to Escape Atypical Regions in the Symmetric Binary Perceptron: A Journey through Connected-Solutions States*. DOI: 10.48550/arXiv.2408.04479. arXiv: 2408.04479 [cond-mat]. (Visited on 01/19/2025).

Potters, Marc and Jean-Philippe Bouchaud (2020). *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge: Cambridge University Press. ISBN: 978-1-108-48808-2. DOI: 10.1017/9781108768900. (Visited on 01/25/2025).

Krzakala, Florent and Lenka Zdeborová (2021). "Statistical Physics Methods in Optimization and Machine Learning". In: p. 225.

Mezard, Marc and Andrea Montanari (2009). *Information, Physics, and Computation*. Oxford Graduate Texts. Oxford ; New York: Oxford University Press. ISBN: 978-0-19-857083-7.

Parisi, Giorgio (Dec. 1979a). "Infinite Number of Order Parameters for Spin-Glasses". In: *Physical Review Letters* 43.23, pp. 1754–1756. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.43.1754. (Visited on 08/05/2023).

Parisi, Giorgio (Sept. 1979b). "Toward a Mean Field Theory for Spin Glasses". In: *Physics Letters A* 73.3, pp. 203–205. ISSN: 03759601. DOI: 10.1016/0375-9601(79)90708-4. (Visited on 08/15/2023).

— (June 1983). "Order Parameter for Spin-Glasses". In: *Physical Review Letters* 50.24, pp. 1946–1948. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.50.1946. (Visited on 08/05/2023).

Gamarnik, David et al. (Mar. 2022). *Algorithms and Barriers in the Symmetric Binary Perceptron Model*. DOI: 10.48550/arXiv.2203.15667. arXiv: 2203.15667 [cs]. (Visited on 01/19/2025).

Arous, G. Ben and A. Guionnet (Aug. 1997). "Large Deviations for Wigner's Law and Voiculescu's Non-Commutative Entropy". In: *Probability Theory and Related Fields* 108.4, pp. 517–542. ISSN: 1432-2064. DOI: 10.1007/s004400050119. (Visited on 01/30/2025).

Voiculescu, Dan (Dec. 1994). "The Analogues of Entropy and of Fisher's Information Measure in Free Probability Theory, II". In: *Inventiones mathematicae* 118.1, pp. 411–440. ISSN: 1432-1297. DOI: 10.1007/BF01231539. (Visited on 01/30/2025).

Saff, Edward B. and Vilmos Totik (2024). *Logarithmic Potentials with External Fields*. Vol. 316. Grundlehren Der Mathematischen Wissenschaften. Cham: Springer International Publishing. ISBN: 978-3-031-65132-8 978-3-031-65133-5. DOI: 10.1007/978-3-031-65133-5. (Visited on 01/30/2025).

Forrester, Peter J. and S. Ole Warnaar (Oct. 2007). *The Importance of the Selberg Integral*. DOI: 10.48550/arXiv.0710.3981. arXiv: 0710.3981 [math]. (Visited on 01/30/2025).

Mehta, M. L. (2004). *Random Matrices*. 3rd ed. Pure and Applied Mathematics v. 142. Amsterdam San Diego, CA: Elsevier/Academic Press. ISBN: 978-0-12-088409-4.

Altschuler, Dylan J. (July 2024). *Zero-One Laws for Random Feasibility Problems*. DOI: 10.48550/arXiv.2309.13133. arXiv: 2309.13133 [math]. (Visited on 01/30/2025).

A More words on the techniques

The two main results of Maillard (2024) build on previous knowledge extensively. In this appendix, we will not focus on giving the details but rather the intuition. In simple words, we partially dissect the work for less experienced audiences.

A.I Proof outline of theorem 6

We want to count solutions. By Markov's inequality, to show that solutions are asymptotically almost surely zero, we may use the bound:

$$\mathbb{P}[\#\mathsf{M}_\kappa > 0] \leq \mathbb{E}\left[\#\mathsf{M}_\kappa\right], \tag{A.1}$$

and if we show that the expectation is null in the limit, we are done. Thus, we we focus on the expectation of $\#\mathsf{M}_\kappa$. Practically, it just counts the number of satisfying solutions, so that:

$$\mathbb{E}\left[\#\mathsf{M}_\kappa\right] = \sum_{\boldsymbol{\epsilon}\in\{\pm 1\}^n} \mathbb{E}\left[\mathbb{1}\left\{\left\|\sum_{i=1}\epsilon_i \mathbf{A}^{(i)}\right\|_{\mathsf{op}} \leq \kappa\sqrt{n}\right\}\right] = \sum_{\boldsymbol{\epsilon}\in\{\pm 1\}^n} \mathbb{P}\left[\left\|\sum_{i=1}\epsilon_i \mathbf{A}^{(i)}\right\|_{\mathsf{op}} \leq \kappa\sqrt{n}\right]. \tag{A.2}$$

By invariance and centering, as we saw earlier, the large sum of GOE matrices is just a large GOE matrix, and the probability inside the sum is invariant w.r.t. the specific sign. Then, we can rewrite:

$$\mathbb{E}\left[\#\mathsf{M}_\kappa\right] = 2^n \mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \leq \kappa\right]. \tag{A.3}$$

Obviously, it will *not always* be small; we need top find the right conditions on (κ, τ). Let us cheat a little beat, or put more simply, use previous results. In particular, we want some upper bound on a probability, that beats the 2^n factor. This suggests looking at large deviations theory, which describes the limiting behavior of the log-probability of events. Since we have an operator norm, the natural object to look at is the empirical measure. If we had scalar random variables we would roughly look for a rate function $\mathscr{I} : \mathbb{R} \to \mathbb{R}$ such that:

$$\mathbb{P}\left[\mathrm{x}_n > x\right] \approx \mathrm{e}^{-n\mathscr{I}(x)}, \qquad \text{for } n \gg 1. \tag{A.4}$$

Moreover, such rate function is expected to be the result of a balancing between energetic and entropic contributions, in the sense of a regularized companion problem (i.e. a Legendre-Fenchel transform).
However, given an empirical measure, we cannot compute probabilities at points, and will need a generalization, but the principle is the same. Also, we remark that:

- the natural dimension of matrices is $n^{1}/_{\tau} \sim d^2$;
- we cannot take the limit in equation A.4 as the n is on both sides.

This quickly suggests looking at log probabilities in their natural dimensions, i.e. $^{1}/_{d^2} \log \mathbb{P}[[\mu_{\mathbf{A}} \in S]$, where $S \subset \mathbb{R}$. The generalization to empirical measures of GOE matrices was studied in the 90s by Arous and A. Guionnet (1997). The authors find that the large deviation principle is described by two optimization problems over the space of measures. Let us briefly define some

objects and unroll the content of (Maillard 2024, prop. 2.3). We write $\mathcal{M}_1(\mathbb{R})$ for probability measures in the real line.

Our rate function can be seen as a free energy that gives weights in the limit. We will have an energetic contribution and an entropic contribution. For a given probability measure $\mu \in \mathcal{M}_1(\mathbb{R})$, a good notion of entropy first studied by Voiculescu (1994) is the so-called non-commutative entropy. We denote it as:

$$\mathscr{S}(\mu) \coloneqq \int \log|x-y|\mu(\mathrm{d}x)\mu(\mathrm{d}y). \tag{A.5}$$

It turns out that GOE matrices, and more generally Wigner matrices, obey a large deviation principle with rate function that has Voicolescu's entropy inside:

$$\mathscr{I} : \mathcal{M}_1(\mathbb{R}) \to \mathbb{R} \tag{A.6}$$

$$\mu \mapsto -\frac{1}{2}\mathscr{S}(\mu) + \frac{1}{4}\int x^2\mu(\mathrm{d}x) - \frac{3}{8}. \tag{A.7}$$

The analogy with the scalar case is now fundamental, if a random variable may live in $\mathbb{R}$, measures live in $\mathcal{M}_1(\mathbb{R})$, and we may write the large deviation principle in the style of Portmanteau's theorem, i.e. in terms of open or closed sets. In other words, we will find some upper and lower bounds in terms of optimization problems over the space of measures. Mathematically, in (Arous and A. Guionnet 1997) we find:

$$\liminf_{d\to\infty} {}^1\!/_{d^2} \log \mathbb{P}\left[\mu_{\mathbf{A}} \in O\right] \geq -\inf_{\mu\in O} \mathscr{I}(\mu), \qquad \forall O \subseteq \mathcal{M}_1(\mathbb{R}) \text{ open;} \tag{A.8}$$

$$\limsup_{d\to\infty} {}^1\!/_{d^2} \log \mathbb{P}\left[\mu_{\mathbf{A}} \in F\right] \leq -\inf_{\mu\in F} \mathscr{I}(\mu) \qquad \forall F \subseteq \mathcal{M}_1(\mathbb{R}) \text{ closed.} \tag{A.9}$$

Originally, we were interested in the probability of the event $\{\|\mathbf{A}\|_{\mathsf{op}} \leq \kappa\}$, how can we bound it? It turns out that we can write it as:

$$\{\|\mathbf{A}\|_{\mathsf{op}} \leq \kappa\} = \{\mu_{\mathbf{A}} \in \mathcal{Q}\}, \qquad \mathcal{Q} \coloneqq \{\mu : \mathcal{M}_1(\mathbb{R}) \mid \mu([-\kappa,\kappa]) = 1\}, \tag{A.10}$$

where $\mathcal{Q}$ is closed (Maillard 2024, sec. 2.2). The upper bound of the limit is direct, the lower bound is by deduction.

Upper bound By the large deviation upper bound, we have the asymptotic inequality:

$$\mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \leq \kappa\right] \lesssim \mathrm{e}^{-d^2 \inf_{\mu\in\mathcal{Q}} \mathscr{I}(\mu)} \tag{A.11}$$

It remains to compute the true value of the optimization problem. As previously mentioned, the rate function(al) has an energetic and entropic contribution. Using results from earlier works on potential theory (Saff and Totik 2024) and (Maillard 2024, lem. 2.5), we get that the existing and unique optimizer of the rate function satisfies:

$$\int \log|x-y|\mu(\mathrm{d}y) = \frac{x^2}{4} + C, \qquad \text{for some } C, \text{ for all } x \in (-\kappa,\kappa). \tag{A.12}$$

Remark. *The equation above is an optimality first order condition. If we take the functional derivative of the rate functional in equation A.7 it should be analogous. Indeed, we want to study the infimum of $\mathscr{I}$ over a closed set.*

It turns out that after some computations equation A.12 is solved by the following measure (Maillard 2024, prop. 2.7):

$$\mu_\kappa^\star(\mathrm{d}x) \coloneqq \frac{4+\kappa^2-2x^2}{4\pi\sqrt{\kappa^2-x^2}}\delta_{(-\kappa,\kappa)}(x)\mathrm{d}x. \tag{A.13}$$

In particular, the case $\kappa = 2$ is Wigner's semicircle law, the result of theorem 4. Once the optimal measure is known, a little more patience gives us the optimal value of the rate function:

$$\inf_{\mu\in\mathcal{Q}} \mathscr{I}(\mu) = \mathscr{I}(\mu_\kappa^\star) = -\frac{\kappa^4}{128} + \frac{\kappa^2}{8} - \frac{1}{2}\log\frac{\kappa}{2} - \frac{3}{8}. \tag{A.14}$$

Lower bound For the other side, there is no clever open set to use to lower bound. The intuition will be to seek an asymptotic lower bound that attains the same optimal rate functional. As a matter of fact, Maillard (2024) takes inspiration from he proof technique of the lower bound in (Arous and A. Guionnet 1997) but proves directly the matching result. Since it is a long exercise in inequalities, we just remark the main aspects of it. The first crucial passage is recognizing that the GOE comes with an explicit expression for the joint law of the eigenvalues. We have from (Anderson, Alice Guionnet, and Zeitouni 2009, sec. 2.5.4) that for a function depending only on the eigenvalues like the operator norm (take $\beta = 1$ in their notation):

$$\mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \le \kappa\right] = \mathbb{E}\left[\mathbb{1}_{\{|\lambda_i|\le\kappa \text{ for all } i\in\{1,\dots,d\}\}}\right] = \frac{1}{\mathcal{Z}}\int_{\mathbb{R}^d}\prod_{i=1}^d \mathbb{1}_{\{|\lambda_i|\le\kappa\}}\prod_{i<j}^d|\lambda_i-\lambda_j|\mathrm{e}^{-\sum_{i=1}^d\lambda_i^2}\prod_{i=1}^d\mathrm{d}\lambda_i, \tag{A.15}$$

where $\mathcal{Z}$ is a normalizing factor. Interestingly, the denominator has an explicit value at the limit, which is computed via the so-called Selberg's integrals. The story of this result is interesting and well explained in (Forrester and Warnaar 2007). In particular in (Forrester and Warnaar 2007, sec. 1) the authors explain how in Random Matrix Theory the conjecture popularized by Mehta (2004) is solved. Cutting matters short, we can say that in our case:

$$\mathcal{Z} = \int_{\mathbb{R}}\prod_{i<j}^d|\lambda_i-\lambda_j|\mathrm{e}^{-\sum_{i=1}^d\lambda_i^2}\prod_{i=1}^d\mathrm{d}\lambda_i \sim \mathrm{e}^{-d^2\frac{3}{8}}. \tag{A.16}$$

Remark. *Notice that the* 3/8 *appears also in the rate functional, i.e. eqn. A.7. This is not by chance. The normalizaiton factor is a constant term once we take the rescaled limit, so its limit will naturally appear in the rate functional.*

Now, we have to compute a lower bound for the numerator. This is more technical; in substance, we take a perturbed density ν_δ that removes δ from the κ margin to have stability. With careful bounds, we will see that the non-commutative entropy and the rate function for ν_δ appear (refer to (Maillard 2024, sec. 2.3)). Taking the limit $\delta \to 0$, we eventually recover a statement of the form:

$$\mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \le \kappa\right] \gtrsim \mathrm{e}^{-d^2\mathscr{I}(\mu_\kappa^\star)}, \tag{A.17}$$

which matches the upper bound in equation A.11.

Finalization Joining the bounds with the optimal value of the rate function in equation A.14 we have:

$$\mathbb{E}\left[\#\mathbf{M}_\kappa\right] = 2^n \mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \leq \kappa\right] \sim 2^n \mathrm{e}^{-d^2 \mathscr{I}(\mu_\kappa^\star)}, \qquad \mathscr{I}(\mu_\kappa^\star) = -\frac{\kappa^4}{128} + \frac{\kappa^2}{8} - \frac{1}{2}\log\frac{\kappa}{2} - \frac{3}{8}. \quad \text{(A.18)}$$

The expected size of the set of solutions obeys the asymptotic, so we might as well focus on the argument of the exponential. recalling that $n \sim \tau d^2$ and putting everything in e-scale:

$$\mathbb{E}\left[\#\mathbf{M}_\kappa\right] \sim \exp\left\{-d^2\left[\mathscr{I}(\mu_\kappa^\star) - \tau \log 2\right]\right\}. \quad \text{(A.19)}$$

It is evident that we have an all or nothing behavior: if the argument is positive it is exponentially fast exploding, if it is negative it is exponentially fast going to zero. The critical point in the (τ, κ) plane is exactly that of theorem 6, i.e. $\tau > {}^1\!/_{\log 2}\mathscr{I}(\mu_\kappa^\star)$ gives infinity, while the opposite sign gives zero. In particular, the value is ${}^1\!/_{\log 2}\mathscr{I}(\mu_\kappa^\star) = \tau_{\mathsf{c}1}$, the explicit threshold of (Maillard 2024, eqn. 4) mentioned in theorem 6. If we take the log, bring d^2 to the left hand side and take the limit $d \to \infty$ with $n \sim \tau d^2$ we recover the exact same statement of (Maillard 2024, theorem 1.5) reported in theorem 6.

A.II Proof outline of theorem 7

The combination of the second-moment method and abrupt transition techniques allows us to demonstrate that for $\tau > \tau_2(\kappa)$, the mean matrix imbalance problem has a solution with margin κ with probability approaching 1 as the dimension d tends to infinity:

- **Lemma 10.** *Let $\kappa \in (0, 2]$. Recall the definition of $\mathbf{M}_\kappa$ in Equation (3) and that of $\bar{\tau}(\kappa)$.*

Assume that $\tau > \bar{\tau}(\kappa)$. Then, for $n, d \to \infty$ with $n/d^2 \to \tau$:

$$\limsup_{d\to\infty} \frac{\mathbb{E}[\mathbf{M}_k^2]}{\mathbb{E}[\mathbf{M}_k]^2} \leq L\left[1 - \frac{\bar{\tau}(\kappa)}{\tau}\right]^{-1/2}$$

for an absolute constant $L > 0$.

Where $\bar{\tau}(\kappa)$ is given by $\bar{\tau}(\kappa) := \min_{\eta>0}\tau(\eta, \kappa)$. This quantity is the subject of lemma 3.1 in (Maillard 2024). Basically, we will notice that $\kappa \to \bar{\tau}(\kappa)$ is continuous. Also, we have $\tau > \tau_2(\kappa) > \bar{\tau}(\kappa)$ by definition.

Lemma 10 characterizes the second-order method used to prove the result of theorem 7. An important first observation is that the limiting term is strictly greater than 1, which is not strong enough to prove the result with high probability. However, Altschuler (2024) has developed a general method for this kind of transition. By applying it to our problem, we have:

- **Lemma 11** (Theorem 7 of Altschuler (2024))**.** *Let $d \geq 1$, and $\{\mathbf{A}^{(i)}\}_{i=1}^n \sim \mathrm{GOE}(d)^{\otimes n}$. Define:*

$$\mathrm{disc}(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(n)}) := \min_{\boldsymbol{\epsilon}\in\{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i \mathbf{A}^{(i)} \right\|_{\mathsf{op}}.$$

Assume that $n/d^2 \to \tau$ as $d \to \infty$. Then there exists $c(\tau) > 0$ such that:

$$\frac{\mathbb{E}[\mathrm{disc}(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(n)}]}{\sqrt{\mathrm{Var}[\mathrm{disc}(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(n)})]}} \geq c(\tau)\sqrt{d}.$$

By continuity of $\bar{\tau}(\kappa)$, we take ϵ such that $\tau > \bar{\tau}(\kappa - \epsilon)$. With $\mathrm{x} = \mathrm{disc}(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(n)})$, we have $\mathbb{E}[\mathrm{x}] \leq 2\sqrt{n}$ since $\mathrm{x} \leq \| \sum_{i=1}^{n} \mathbf{A}^{(i)} \|_{\mathsf{op}}$. By lemma 11, we obtain:

$$\mathrm{Var}(\mathrm{x}) \leq \frac{4n}{c(\tau)^2 d}.$$

We apply the Paley-Zygmund inequality to $\mathbf{M}_{\kappa-\epsilon}$ to get a lower bound on the probability that the imbalance X is less than $(\kappa - \epsilon)\sqrt{n}$:

$$\mathbb{P}[\mathrm{x} \leq (\kappa - \epsilon)\sqrt{n}] \geq L^{-1} \cdot \left[1 - \frac{\bar{\tau}(\kappa - \epsilon)}{\tau}\right]^{1/2} + o_d(1).$$

Using Chebyshev's inequality and the variance of X, we show that:

$$\mathbb{P}[\mathrm{x} \geq \mathbb{E}[\mathrm{x}] - t] \geq 1 - \frac{4n}{c(\tau)^2 d t^2}.$$

Choosing $t = c(\tau)^{-1}\sqrt{8C(\tau, \kappa, \epsilon)n/d}$, we get:

$$\mathbb{P}[\mathrm{x} \geq \mathbb{E}[\mathrm{x}] - t] \geq 1 - (2C(\tau, \kappa, \epsilon))^{-1}.$$

Combining the previous results, we obtain an upper bound on the expectation of x:

$$\mathbb{E}[\mathrm{x}] \leq (\kappa - \epsilon)\sqrt{n} + c_2(\tau, \kappa, \epsilon)\sqrt{\frac{n}{d}}.$$

Applying Chebyshev's inequality again, we show that for $u = \epsilon\sqrt{n} - c_2\sqrt{n/d}$, we have:

$$\mathbb{P}[\mathrm{x} \leq \kappa\sqrt{n}] \geq 1 - \frac{4c_1(\tau)}{\epsilon^2 d} \to_{d\to\infty} 1.$$

This concludes the proof by showing that the probability that the imbalance is less than $\kappa\sqrt{n}$ tends to 1 as $d \to \infty$.