Some further topics on the

Mathematics of Data Science

Fall 2024

Antoine Maillard
antoine.maillard@math.ethz.ch

Last update: November 20, 2024

Overview

These notes cover the material I covered during an advanced session at ETH in the fall of 2024, complementary to the course "Mathematics of Data Science" (MDS) taught by Afonso Bandeira. The present notes are focused on discussing more advanced topics, as well as open problems, in the different areas explored in the MDS lecture, as well as some others. Especially Sections 3 to 7 should complement the book [BSS23] on which the main course is based. Because some sessions were attached to the ones of the MDS course, the topics of these notes can greatly vary from section to section!

You will notice some extra questions along the way, usually stated as **Challenges**. They are well-defined mathematical questions, of varying level of difficulty. The solution to some of them is given, but in any case you are encouraged try them by yourself before reading it!

Important disclaimer – *This draft is subject to possible future changes, adds and removals. If you find any typos or mistakes, please let me know! This draft was last updated on November 20, 2024.*

Contents

1 A warmup: Feige's conjecture

In [Fei06], Feige conjectured the following

Conjecture 1.1 (*Feige*)

Let $n \geq 1$, and $X_1, \cdots, X_n$ be independent random variables such that $X_i \geq 0$ and $\mathbb{E}[X_i] = 1$ for all $i \in [n]$. Then

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i < n+1\right] \geq \frac{1}{e}.$$

This conjecture has applications in particular in computer science and combinatorics, such as the weighted max-cut problem, see [Alq+24] for details and further applications. Note that Markov's inequality yields the much weaker bound that goes to 0 as n grows:

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i < n+1\right] \geq 1 - \frac{n}{n+1} = \frac{1}{n+1}.$$

In [Fei06], Feige showed that Conjecture 1.1 holds when replacing $1/e$ by $1/13$ on the right-hand side, by a technical case-by-case analysis after reducing the support of each X_i to have at most two elements. This constant has been progressively improved [HZZ10; Gar20], the best-known bound is due to [Guo+20] and is 0.1798 (note that $1/e \simeq 0.3679$). The conjecture was also recently proven for discrete log-concave random variables in [Alq+24]. The following challenge shows that the constant $1/e$ in Conjecture 1.1 can not be improved.

Challenge 1.1. *Find a set of random variables $X_1, \cdots, X_n$ as in Conjecture 1.1, such that*

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i < n+1\right] \to_{n\to\infty} \frac{1}{e}.$$

Solution of Challenge 1.1 – Pick $X_0, \cdots, X_n$ i.i.d. realizations of a random variable $X \in \{0, n+1\}$, with $\mathbb{P}[X = n+1] = 1/(n+1)$. Then

$$\mathbb{P}\left[\sum_{i=0}^{n} X_i < n+1\right] = \mathbb{P}[X=0]^n = \left(1 - \frac{1}{n+1}\right)^n \sim_{n\to\infty} \frac{1}{e}.$$

□

2 Discrepancy theory

2.1 Some motivation

Imagine we are given a family of sets $\mathcal{S} \coloneqq \{S_1, \cdots, S_m\}$, with $S_i \subseteq \{1, \cdots, n\}$. A good example to keep in mind is a clinical trial: patients are labeled from 1 to n, and a set S_i contains the patients having a (binary) characteristic $i \in \{1, \cdots, m\}$ (older or younger than 40, sex at birth, patient having a cancer or not, ...).

Our goal is to find a *coloring*, i.e. a map $\chi : \{1, \cdots, n\} \to \{-1, 1\}$, that is a partition of $\{1, \cdots, n\}$ in two sets, such that for each $i \in [m]$, $\chi^{-1}(\{1\}) \cap S_i$ and $\chi^{-1}(\{-1\}) \cap S_i$ partition S_i into two sets of roughly of the same size. In the example above, this corresponds to splitting patients into two groups, such that each characteristic $i \in [m]$ is roughly balanced between the two groups, ensuring that in a clinical trial a significant difference of outcomes between the two groups would likely be due to the treatment. Informally, discrepancy is about splitting our elements in two groups which look "as similar as possible".

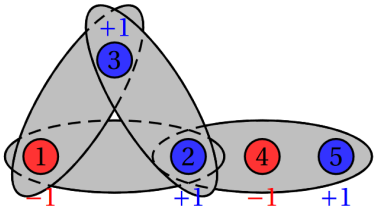
Let us formalize this. The *discrepancy* of the coloring is

$$\operatorname{disc}(\chi) \coloneqq \max_{i \in [m]} \left| \sum_{j \in S_i} \chi(j) \right|. \tag{1}$$

We define then the discrepancy of $\mathcal{S}$ as the minimal discrepancy of colorings, i.e.

$$\operatorname{disc}(\mathcal{S}) \coloneqq \min_{\chi : [n] \to \{-1,1\}} \max_{S \in \mathcal{S}} \left| \sum_{j \in S} \chi(j) \right|. \tag{2}$$

In the example below (taken from https://math.mit.edu/classes/18.095/lect6/notes.pdf), the set system is $\mathcal{S} = \{\{1,2\}, \{1,3\}, \{2,3\}, \{2,4,5\}\}$, and a minimal discrepancy coloring that we show has discrepancy 2.



Discrepancy of vectors – Notice that we can represent the sets S_i by their indicator vectors $a_i \coloneqq \mathbb{1}_{S_i} \in \{0,1\}^n$, and denote $\varepsilon_i \coloneqq \chi(i)$, with $\varepsilon \in \{\pm 1\}^n$. With these notations, we have

$$\operatorname{disc}(\mathcal{S}) = \min_{\varepsilon \in \{\pm 1\}^n} \max_{i \in [m]} |\langle \varepsilon, a_i \rangle|. \tag{3}$$

We can represent eq. (3) in an equivalent way, if we denote $u_1, \cdots, u_n \in \{0,1\}^m$ the vectors such that $(u_i)_j \coloneqq (a_j)_i$, as

$$\operatorname{disc}(\mathcal{S}) = \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_\infty. \tag{4}$$

These general forms motivate considering the discrepancy of more general sets of vectors, beyond indicator vectors. Generally, for $n, d \geq 1$ and $u_1, \cdots, u_n \in \mathbb{R}^d$, we will denote:

$$\operatorname{disc}(u_1, \cdots, u_n) \coloneqq \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_\infty. \tag{5}$$

2.2 Spencer's theorem

In our original motivation with set systems, the vectors u_i took value in $\{0,1\}$. As such, the following seminal result of discrepancy theory directly applies to this case.

Theorem 2.1 (*"Six Standard Deviations Suffice"*)

There exists a constant $C > 0$ such that the following holds. Let $n \geq 1$, and let $u_1, \cdots, u_n \in \mathbb{R}^n$ with $\|u_i\|_\infty \leq 1$. Then $\mathrm{disc}(u_1, \cdots, u_n) \leq C\sqrt{n}$.

Notice that one can show that $C \leq 6$ (hence the name "Six deviations suffice"), although the optimal value of C is not know. In [Ban10], Bansal showed that a discrepancy $\mathcal{O}(\sqrt{n})$ can be achieved with a polynomial-time algorithm.

Challenge 2.1. *Prove a weaker version of Theorem 2.1 where $\sqrt{n}$ is replaced with $\sqrt{n \log n}$.*

Solution to Challenge 2.1 – We use the so-called *probabilistic method*, in the following form. Let $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{\pm 1\})$ be random signs. Let $a_1, \cdots, a_n$ defined as $(a_j)_i \coloneqq (u_i)_j$. Then $\|a_i\|_\infty \leq 1$ for all $i \in [n]$, and $\|\sum_{i=1}^n \varepsilon_i u_i\|_\infty = \max_{j\in[n]} |\langle \varepsilon, a_j\rangle|$. Let us recall Hoeffding's inequality, which you have seen in class:

Theorem 2.2 (*Hoeffding's inequality, bounded random variables*)

Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely for $i = 1, \ldots, n$. Then, for any $t \geq 0$, it holds that

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}X_i\right| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

By Hoeffding's inequality, we have for any $j \in [n]$:

$$\mathbb{P}[|\langle \varepsilon, a_j\rangle| \geq t] \leq 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n (a_j)_i^2}\right\} \leq 2\exp\left\{-\frac{t^2}{2n}\right\}. \tag{6}$$

Using the union bound, we have

$$\mathbb{P}[\max_{j\in[n]} |\langle \varepsilon, a_j\rangle| \geq t] \leq 2n\exp\{-t^2/(2n)\}. \tag{7}$$

Therefore,

$$\mathbb{P}[\max_{j\in[n]} |\langle \varepsilon, a_j\rangle| < t] \geq 1 - 2n\exp\{-t^2/(2n)\}. \tag{8}$$

In particular, if $1 - 2n\exp\{-t^2/(2n)\} > 0$, then $\mathbb{P}[\max_{j\in[n]} |\langle \varepsilon, a_j\rangle| < t] > 0$ and therefore there *must exist* a signing $\varepsilon \in \{\pm 1\}^n$ with $\max_{j\in[n]} |\langle \varepsilon, a_j\rangle| < t$ (such arguments are called the probabilistic method). This implies that $\mathrm{disc}(u_1, \cdots, u_n) \leq \sqrt{2n\log(2n)}$. □

A great reference on the probabilistic method is the set of lectures by Spencer [Spe94].

Sketch of proof of Theorem 2.1 – We describe here a sketch of the proof of Spencer's theorem (with suboptimal constant C), which is due to Gluskin [Glu89] and relies on beautiful geometric ideas. The original proof of Spencer [Spe85] shares many ideas, but is less accessible. A great reference for this proof (which we follow here) is this series of blog posts of Raghu Meka: [Mek14a; Mek14c; Mek14b]. For convenience, we

will consider the equivalent problem as a function of $(a_i)_j = (u_j)_i$, see eq. (3), so that $\|\sum_{i=1}^n \varepsilon_i u_i\|_\infty = \max_{j\in[n]} |\langle \varepsilon, a_j\rangle|$.

1. Partial coloring is sufficient – First, a crucial idea is that it is enough to "color" (i.e. pick $\varepsilon_i \in \{\pm 1\}$) a positive fraction of $i \in [n]$. The remaining ε_i will be taken equal to 0.

Lemma 2.3 (*Partial coloring*)

For any $n, d \geq 1$ and any $a_1, \cdots, a_n \in \mathbb{R}^d$ with $\|a_i\|_\infty \leq 1$, there exists $\varepsilon \in \{-1, 0, 1\}^d$ such that:

(*i*) For all $j \in [d]$, $|\langle \varepsilon, a_j\rangle| \leq C\sqrt{d \max(1, \log \frac{n}{d})}$ (for some absolute constant $C > 0$).

(*ii*) $\#\{i \in [d] \,:\, \varepsilon_i \neq 0\} \geq \eta d$, for some absolute constant $\eta > 0$.

Let us show that Lemma 2.3 implies Theorem 2.1. We choose $\varepsilon \in \{-1, 0, 1\}^n$ given by Lemma 2.3 for $(a_1, \cdots, a_n)$, and denote $I \coloneqq \#\{i \in [n] \,:\, \varepsilon_i = 0\}$, so that $|I| \leq (1-\eta)n$. By the triangle inequality we have:

$$\operatorname{disc}(a_1, \cdots, a_n) \leq C\sqrt{n} + \operatorname{disc}(b_1, \cdots, b_n),$$

where $b_i \in \mathbb{R}^I$, with $b_i = (a_i)_{|I}$ is the restriction of the vector a_i to the index set I. Notice that $\|b_i\|_\infty \leq 1$, and $b_i \in \mathbb{R}^{n_1}$, with $n_1 \leq (1-\eta)n$. We can then apply Lemma 2.3 to b_i, and proceed like this iteratively. After k steps, we have

$$\operatorname{disc}(a_1, \cdots, a_n) \leq C \cdot \sqrt{n} \sum_{p=0}^{k-1} (1-\eta)^{p/2} \left[\max(1, -p\log(1-\eta))\right]^{1/2} + \operatorname{disc}(c_1, \cdots, c_n). \tag{9}$$

with $c_i \in \mathbb{R}^{n_k}$, $n_k \leq (1-\eta)^k n$, and $\|c_i\|_\infty \leq 1$. Taking $k = \mathcal{O}(\log n)$ is thus sufficient to make it such that $n_k = \mathcal{O}(1)$, in which case it is clear that $\operatorname{disc}(c_1, \cdots, c_n) = \mathcal{O}(1)$. Thanks to the geometric decrease of the bound in eq. (9), we get

$$\begin{aligned}\operatorname{disc}(a_1, \cdots, a_n) &\leq C \cdot \sqrt{n} \sum_{p=0}^{\infty} (1-\eta)^{p/2} \left[\max(1, -p\log(1-\eta))\right]^{1/2} + \mathcal{O}(1), \\ &\leq C(\eta)\sqrt{n},\end{aligned}$$

which ends the proof of Theorem 2.1.

2. Minkowski's theorem and the discrepancy body – Now that we saw that proving Lemma 2.3 is sufficient, we will show it holds by using a beautiful geometric argument. For $\Delta > 0$, we define the *discrepancy body* as:

$$\mathcal{K}_\Delta \coloneqq \{x \in \mathbb{R}^d \,:\, \max_{j\in[n]} |\langle x, a_j\rangle| \leq \Delta\}. \tag{10}$$

Notice that $\mathcal{K}_\Delta$ is a compact symmetric convex set. The main idea is to show that $\mathcal{K}_\Delta$ contains an integer point $\varepsilon \in \{-1, 0, 1\}^d$ (with a large support) in a non-constructive way, by *lower bounding its volume*. We will not prove the following lemma, we refer to [Mek14c].

Lemma 2.4 (*Minkowski's lemma*)

Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a symmetric convex set. Assume that, for $g \sim \mathcal{N}(0, \mathrm{I}_d)$:

$$\gamma_d(\mathcal{K}) \coloneqq \mathbb{P}[g \in \mathcal{K}] \geq \ell \cdot 2^{-d},$$

for some integer $\ell \geq 1$. Then $\mathcal{K}$ contains at least ℓ points in $\{-1,0,1\}^d$.

We notice that $\mathcal{K}_\Delta$ is the intersection of the n symmetric convex sets (called *slabs*) $\{x : |\langle x, a_j\rangle| \leq \Delta\}$. Sidak's lemma [Šid67] shows that for $g \sim \mathcal{N}(0, \mathrm{I}_d)$:

$$\mathbb{P}[g \in \mathcal{K}_\Delta] \geq \prod_{j=1}^{n} \mathbb{P}[|\langle g, a_j\rangle| \leq \Delta]. \tag{11}$$

Eq. (11) is now known as a special case of the *Gaussian Correlation Inequality*, a very general theorem on the Gaussian measure of intersection of symmetric convex sets, which was only proven a decade ago [Roy14; LM17a] (read the story about this result!). Since $\langle g, a_j\rangle$ is a Gaussian random variable with zero mean and variance $\|a_j\|_2^2 \leq \|a_j\|_\infty^2 \leq d$, it is easy to show that $\mathbb{P}[|\langle g, a_j\rangle| > \Delta] \leq 2\exp\{-\Delta^2/(2d)\}$. This yields from eq. (11):

$$\mathbb{P}[g \in \mathcal{K}_\Delta] \geq \left[1 - 2\exp\{-\Delta^2/(2d)\}\right]^n. \tag{12}$$

Therefore, if we pick $\Delta = C\sqrt{d\max(1, \log n/d)}$, we have:

$$\mathbb{P}[g \in \mathcal{K}_\Delta] \geq \left[1 - 2\exp\left\{-\frac{C^2}{2}\max(1, \log n/d)\right\}\right]^n. \tag{13}$$

One can easily check[1] that there is $C > 0$ large enough such that for all n, d:

$$\left[1 - 2\exp\left\{-\frac{C^2}{2}\max(1, \log n/d)\right\}\right]^{n/d} \geq \frac{3}{4}. \tag{14}$$

We then get:

$$\mathbb{P}[g \in \mathcal{K}_\Delta] \geq (3/4)^d. \tag{15}$$

Applying Lemma 2.4, we see that $\mathcal{K}_\Delta$ (for $\Delta = C\sqrt{d}$) has at least $(3/2)^d$ points in $\{-1,0,1\}^d$. Notice that, for any $\eta > 0$, there are most

$$L_\eta = \binom{d}{\eta d} 3^{\eta d}$$

points in $\{-1,0,1\}^d$ that are zero everywhere except possibly on a set of cardinality ηd. It is a simple exercise (do it!) to show that for $\eta > 0$ small enough (but independent of d), $L_\eta \leq (5/4)^d < (3/2)^d$. Therefore, $\mathcal{K}_\Delta$ must contain an element $\varepsilon \in \{-1,0,1\}^d$ which has more than ηd non-zero coordinates: this ends the proof of Lemma 2.3. □

Challenge 2.2. *Write Theorem 2.1 when the number n of vectors is different from their dimension d, i.e. we have $u_1, \cdots, u_n \in \mathbb{R}^d$ with $\|u_i\|_\infty \leq 1$.*

The scaling $\sqrt{n}$ in Theorem 2.1 can not be improved, as shown by the following example.

Challenge 2.3. *A Hadamard matrix of order n is a square $n \times n$ matrix with elements in $\{\pm 1\}$, and such that its rows are mutually orthogonal. It is known that if $n = 2^k$ (for $k \geq 1$), there exists a symmetric Hadamard matrix[2] H_n of order n. Denote $u_1, \cdots, u_n$ the rows of H_n. Show that* $\mathrm{disc}(u_1, \cdots, u_n) \geq \sqrt{n}$.

[1] Separate the cases $n \geq ed$ and $n \leq ed$.

[2] Whether a Hadamard matrix of order n exists for all $n = 4k$ with $k \geq 1$ is still an open problem, look at https://en.wikipedia.org/wiki/Hadamard_matrix!

Solution to Challenge 2.3 – Let $\varepsilon \in \{\pm 1\}^n$. We let again $(a_i)_j \coloneqq (u_j)_i$: $(a_1, \cdots, a_n)$ are still orthogonal, with elements in $\{\pm 1\}$. Since $a_1, \cdots, a_n$ form an orthogonal basis of $\mathbb{R}^n$, we have

$$\|\varepsilon\|^2 = \sum_{i=1}^n \frac{\langle \varepsilon, a_i\rangle^2}{\|a_i\|^2},$$

so that $\sum_{i=1}^n \langle \varepsilon, a_i\rangle^2 = n^2$. Therefore there must exists some $i \in [n]$ such that $|\langle \varepsilon, a_i\rangle| \geq \sqrt{n}$. □

2.3 Komlós conjecture

The following is one of the most important open problems in discrepancy theory.

Conjecture 2.5 (*Komlós*)

Denote

$$K(n) \coloneqq \inf\left\{\operatorname{disc}(u_1, \cdots, u_n) \,|\, u_1, \cdots, u_n \in \mathbb{R}^n \text{ s.t. } \|u_i\|_2 \leq 1 \text{ for all } i \in [n]\right\}.$$

There exists $K > 0$ such that $K(n) \leq K$ for all $n \geq 1$.

At the moment, the best known bound is due to [Ban98], which has showed $K(n) = \mathcal{O}(\sqrt{\log n})$. The partial coloring method used to prove Spencer's theorem, on the other hand, has only been showed to yield a bound $K(n) = \mathcal{O}(\log n)$ [BMZ22]. The following exercise shows that, while applying Spencer's theorem 2.1 to the Komlós setting yields $K(n) \leq C\sqrt{n}$ (since $\|u_i\|_\infty \leq \|u_i\|_2 \leq 1$), such a bound can already be obtained with much simpler arguments.

Challenge 2.4. *Prove that $K(n)$ in Conjecture 2.5 satisfies $K(n) \leq \sqrt{n}$.*

Solution to Challenge 2.4 – Let $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \operatorname{Unif}(\{\pm 1\})$. Then

$$\begin{aligned}
\operatorname{disc}(u_1, \cdots, u_n)^2 &\leq \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i u_i\right\|_\infty^2, \\
&\leq \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i u_i\right\|_2^2, \\
&= \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2]\|u_i\|_2^2 + \sum_{i\neq j} \underbrace{\mathbb{E}[\varepsilon_i\varepsilon_j]}_{=0}\langle u_i, u_j\rangle, \\
&\leq n.
\end{aligned}$$

□

3 The Matrix Spencer conjecture

We make here a short detour back to discrepancy theory, which we introduced in Section 2. Recall that we defined the discrepancy of a collection of vectors $u_1, \cdots, u_n \in \mathbb{R}^d$ as

$$\operatorname{disc}(u_1, \cdots, u_n) \coloneqq \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i u_i \right\|_\infty .$$

We can generalize this notion to a collection of symmetric matrices $A_1, \cdots, A_n \in \mathbb{R}^{d \times d}$ as:

$$\operatorname{disc}(A_1, \cdots, A_n) \coloneqq \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i A_i \right\|_{\mathrm{op}} . \tag{16}$$

You have seen in the main class [BSS23] the non-commutative Khintchine (NCK) inequality, which implies that for a random choice of $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \operatorname{Unif}(\{\pm 1\})$:

$$\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i A_i \right\|_{\mathrm{op}} \lesssim \sqrt{\log d} \left\| \sum_{i=1}^n A_i^2 \right\|_{\mathrm{op}} . \tag{17}$$

Eq. (16) already gives a bound on the discrepancy via the probabilistic method. Very much like Spencer's Theorem 2.1 showed that the discrepancy of a set of vectors is smaller by a log factor than the bound given by a random choice, the following open problem states that the same thing occurs in matrix discrepancy.

Conjecture 3.1 (*Matrix Spencer*)

Let $A_1, \cdots, A_n \in \mathbb{R}^{n \times n}$ be symmetric matrices, with $\|A_i\|_{\mathrm{op}} \leq 1$ for all $i \in [n]$. Then

$$\operatorname{disc}(A_1, \cdots, A_n) = \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i A_i \right\|_{\mathrm{op}} \leq C\sqrt{n},$$

for an absolute constant $C > 0$.

Like Spencer's Theorem 2.1, a version of this conjecture can be stated for the size d of the matrices not necessarily equal to their numbers n. Conjecture 3.1 first appeared in [Zou12], and is also discussed in the blog post [Mek14a] that we mentioned in Section 2. As we mentioned above, the non-commutative Khintchine inequality gives Conjecture 3.1 up to a $\sqrt{\log n}$ factor.

Commutativity plays a role – Note that Spencer's Theorem 2.1 can be seen as the special case of Conjecture 3.1 when the matrices A_i are pairwise commutative (convince yourself!). So clearly, we need to be able to make a better choice of signings than a random one if we wish to prove Conjecture 3.1. On the other hand, when the matrices behave "very non-commutatively", we expect that a random choice of signings will work: this is mentioned in the notes [BSS23] as *improvements over the NCK inequality*. The following example is arguably the simplest instance of "very non-commutative" matrices.

Challenge 3.1. *Let $A_1, \cdots, A_n \overset{\text{i.i.d.}}{\sim} \mathrm{GOE}(d)$ (see eq. (21)). Show that (with high probability) there exists $\varepsilon \in \{\pm 1\}^n$ such that $\operatorname{disc}(A_1, \cdots, A_n) \leq C\sqrt{n}$. (Try to prove that one can take $C = 2$).*

Actually, characterizing the optimal possible constant $C > 0$ in Challenge 3.1 is still an open problem [KZ23; Mai24]. If $X_{n,d} \coloneqq \mathrm{disc}(A_1, \cdots, A_n)$, it has been proven in [Mai24] Via the first and second moment methods (we will revisit these methods in Section 7 if you are not familiar with it) that for $n \ll d^2$, $X_{n,d}/\sqrt{n} \to 2$, and on the other hand that $X_{n,d} \ll \sqrt{n}$ if $n \gg d^2$. The same work also shows non-trivial bounds on $X_{n,d}$ when $n/d^2 = \Theta(1)$, but the actual limiting value of $X_{n,d}/\sqrt{n}$ is not known.

The "non-commutativity" of a set of matrices in general is hard to quantify, but several approaches have been taken in this direction. Using the improvements over NCK of [BBH23] and a nice decomposition argument, the best-known results on Conjecture 3.1 are due to [BJM23], which prove:

Theorem 3.2 (*[BJM23]*)

Conjecture 3.1 holds if we further assume that the matrices (A_i) also satisfy

$$\max_{i \in [n]} \mathrm{rk}(A_i) \leq \frac{n}{\log^3 n}.$$

In a nutshell, we have two extreme cases (commutative and very non-commutative matrices) in which two very different arguments (respectively Spencer's theorem 2.1 and a random choice, via improvements over the NCK inequality) yield Conjecture 3.1. The difference of nature of these arguments, as well as the fact that we do not know how to combine them to study general matrices A_i, are two reasons why Conjecture 3.1 is so interesting!

A group theoretic approach – Motivated by the role played by commutativity, [Ban+24] introduced the special case of Conjecture 3.1 when A_i are the regular representation of a finite group G: in this case, the commutativity of the matrices can be quantified by the commutativity of G.

Conjecture 3.3 (*Group Matrix Spencer*)

Let G be a finite group of size n, and $A_1, \cdots, A_n \in \mathbb{R}^{n \times n}$ the matrices corresponding to the left regular representation of G. Then

$$\mathrm{disc}(A_1, \cdots, A_n) = \min_{\varepsilon \in \{\pm 1\}^n} \left\| \sum_{i=1}^n \varepsilon_i A_i \right\|_{\mathrm{op}} \leq C\sqrt{n},$$

for some absolute constant $C > 0$.

[Ban+24] prove Conjecture 3.3 when G is a simple group (and Spencer's Theorem 2.1 proves it when G is abelian), but the general case is still open.

4 Concentration inequalities

Disclaimer – This section is largely inspired by Chapter 2 of [Ver18], and is essentially a (very) shortened version of it. This book is a must read for students interested in high-dimensional probability, and the mathematical aspects of data science!

4.1 Sub-Gaussian random variables, and general Hoeffding's inequality

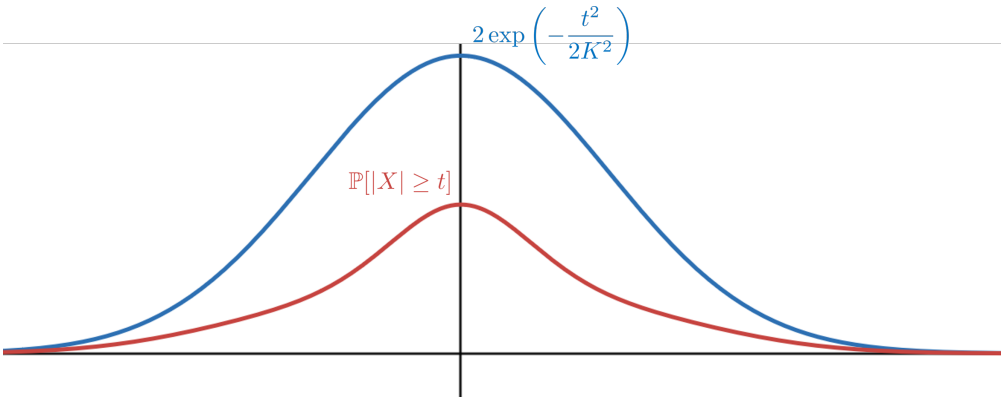
We have already seen Hoeffding's inequality (Theorem 2.2), which shows that if $X_1, \cdots, X_n$ are centered independent random variables, such that $|X_i| \leq K_i$ a.s., then the sum $S_n := \sum_{i=1}^n X_i$ satisfies the tail inequality, for any $t > 0$:

$$\mathbb{P}[|S_n| \geq t] \leq 2 \exp\left\{ - \frac{t^2}{2 \sum_{i=1}^n K_i^2} \right\}. \tag{18}$$

Can we hope to generalize a result such as eq. (18) to random variables X_i under weaker assumptions than boundedness? Clearly, if eq. (18) holds for all $n \geq 1$, then it must hold for $n = 1$, so any random variable X in the sum must satisfy:

$$\mathbb{P}[|X| \geq t] \leq 2 \exp\{-t^2/(2K^2)\}. \tag{19}$$

Eq. (19) defines what is called a **sub-Gaussian** random variable: it mathematically states that the tails of the random variable X are dominated by a Gaussian curve.



We can give several equivalent definitions of a sub-Gaussian random variable. To motivate them, we characterize then two other properties of Gaussian random variables (which are clearly sub-Gaussian!), via the moment generating function (MGF), and via the moments themselves.

Challenge 4.1. *Let $X \sim \mathcal{N}(0, \sigma^2)$. Show:*

(*i*) *(Tail) For all $t > 0$, $\mathbb{P}[|X| \geq t] \leq 2 \exp\{-t^2/(2\sigma^2)\}$.*

(*ii*) *(MGF) For all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp\{\lambda X\}] \leq \exp\{\lambda^2 \sigma^2 / 2\}$.*

(*iii*) *(Moments) For all $p \geq 1$, $\|X\|_p := [\mathbb{E}|X|^p]^{1/p} \leq C\sigma\sqrt{p}$, for some global constant $C > 0$.*

Challenge 4.1 motivates the following definition.

Definition 4.1 (*Sub-Gaussian random variable*)

A centered random variable X is sub-Gaussian if it satisfies one of the following three conditions.

(*i*) **(Tail)** For all $t > 0$, $\mathbb{P}[|X| \geq t] \leq 2\exp\{-t^2/(2K_1^2)\}$, for some $K_1 > 0$.

(*ii*) **(MGF)** For all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp\{\lambda X\}] \leq \exp\{\lambda^2 K_2^2/2\}$, for some $K_2 > 0$.

(*iii*) **(Moments)** For all $p \geq 1$, $\|X\|_p := [\mathbb{E}|X|^p]^{1/p} \leq K_3\sqrt{p}$, for some $K_3 > 0$.

We will say that X is σ-sub-Gaussian (or $\mathrm{SG}(\sigma)$) if $\mathbb{E}[\exp\{\lambda X\}] \leq \exp\{\lambda^2\sigma^2/2\}$ for all $\lambda \in \mathbb{R}$.

Challenge 4.2. *Check that the conditions (i), (ii), (iii) in Definition 4.1 are equivalent, and that (K_1, K_2, K_3) differ by at most an absolute multiplicative constant.*

The following exercise is a classical consequence of sub-Gaussianity.

Challenge 4.3. *Let $X_1, \cdots, X_n$ be independent $\mathrm{SG}(\sigma)$ random variables. Show that ($\lesssim$ means "up to a constant independent of σ and n"):*

$$\mathbb{E}\max_{i\in[n]}|X_i| \lesssim \sigma\sqrt{\log n}.$$

(Harder) Is the corresponding lower bound true for some sub-Gaussian random variables?

Finally, this challenge shows that bounded random variables are sub-Gaussian (which is not surprising, since bounded random variables have tails $\mathbb{P}(|X| \geq t) = 0$ for large enough t!).

Challenge 4.4. *Show that if $|X| \leq a$, then X is $\mathrm{SG}(Ka)$, for $K > 0$ an absolute constant.*

The careful reader will notice that proving Challenge 4.4 is suspiciously similar to the proof of Hoeffding's inequality (Theorem 2.2) done in [BSS23]... We are now ready to state the most general version of Hoeffding's inequality. It shows that sub-Gaussianity is both necessary and sufficient to generalize eq. (18).

Theorem 4.1 (*Hoeffding's inequality*)

Let $X_1, \cdots, X_n$ be independent and centered sub-Gaussian random variables, with sub-Gaussian parameters $\sigma_1, \cdots, \sigma_n$. Then for all $a \in \mathbb{R}^n$ and all $t > 0$:

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n a_i^2\sigma_i^2}\right\}.$$

Proof of Theorem 4.1 – We use again Chernoff's method, which reads

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \inf_{\lambda\geq 0}\left[e^{-\lambda t}\,\mathbb{E}\,e^{\sum_{i=1}^n a_i X_i}\right].$$

Using independence of the X_i's, and the definition of sub-Gaussianity, we get:

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \inf_{\lambda\geq 0}\left[e^{-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^n a_i^2\sigma_i^2}\right] = \exp\left[\inf_{\lambda\geq 0}\left\{-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^n a_i^2\sigma_i^2\right\}\right].$$

The minimum of this quadratic function is reached in $\lambda = t/(\sum_{i=1}^n a_i^2\sigma_i^2)$, which gives:

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^n a_i^2\sigma_i^2}\right\}.$$

Applying this inequality to $-X_i$ and using the union bound yields the statement of Theorem 4.1. □

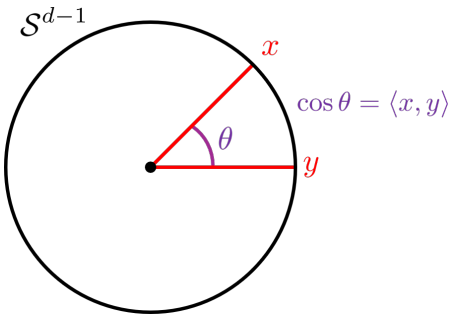
Remark I – Notice that the proof of Hoeffding's inequality is extremely simple once we have the correct notion of random variables to apply it to. Further, if you look again at the proof of Theorem 2.2 which you have seen in class [BSS23], you will see that the heart of it is to prove that bounded random variables are sub-Gaussian (see Challenge 4.4), which now makes sense!

Remark II – An equivalent way to state Hoeffding's inequality is that if $X_i \in \mathrm{SG}(\sigma_i)$ are independent, then $\sum_{i=1}^n X_i \in \mathrm{SG}(\sqrt{\sum_{i=1}^n \sigma_i^2})$. Remember that the sum of independent sub-Gaussian variables is also sub-Gaussian, and its sub-Gaussian parameter scales like the standard deviation of the sum!

Challenge 4.5 (Almost-orthogonal vectors in high dimension)**.** *Let $d \geq 1$ and $\varepsilon > 0$. How many points*[3] *$u_1, \cdots, u_n \in \mathcal{S}^{d-1}$ can we build such that*

$$(\forall i \neq j) \quad |\langle u_i, u_j \rangle| \leq \varepsilon \ ?$$

(Hint: build your points randomly.) (This challenge is also present in the main notes [BSS23].)



For $x, y \in \mathcal{S}^{d-1}$, the "overlap" $\langle x, y \rangle$ is a direct measure of the angle between the two vectors, and thus of their distance.

Solution of Challenge 4.5 – Let us draw $u_1, \cdots, u_n$ i.i.d. from $\mathrm{Unif}(\{\pm 1/\sqrt{d}\}^d)$ and apply the probabilistic method. Clearly, $u_i \in \mathcal{S}^{d-1}$. Applying Hoeffding's inequality, we get for any $i \neq j$:

$$\mathbb{P}(|\langle u_i, u_j \rangle| \geq \varepsilon) \leq 2\exp\{-d\varepsilon^2/2\}.$$

We can now use the union bound:

$$\begin{aligned}
\mathbb{P}(\max_{i \neq j} |\langle u_i, u_j \rangle| \geq \varepsilon) &\leq \sum_{i<j} \mathbb{P}(|\langle u_i, u_j \rangle| \geq \varepsilon), \\
&\leq \frac{n(n-1)}{2} \cdot 2\exp\{-d\varepsilon^2/2\}, \\
&\leq n^2 \exp\{-d\varepsilon^2/2\}.
\end{aligned}$$

Therefore, $\mathbb{P}(\max_{i \neq j} |\langle u_i, u_j \rangle| \leq \varepsilon) > 0$ as long as $n < \exp\{d\varepsilon^2/4\}$. By the probabilistic method, this means that there must exist at least $\lfloor \exp\{d\varepsilon^2/4\} - 1 \rfloor$ vectors with pairwise overlap at most ε! □

Challenge 4.5 is another example of how un-intuitive high-dimensional geometry is: we can build a number of almost-orthogonal vectors which is exponentially large in the dimension!

[3] Recall that $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d \ : \ \|x\|_2 = 1\}$ is the Euclidean unit sphere.

4.2 Sub-exponential tails, and Bernstein's inequality

Despite being a generic and powerful inequality, Hoeffding's inequality (Theorem 4.1) fails to control some random variables, for which the sum should still concentrate quite strongly. A good example is given as follows. Let $g = (g_1, \cdots, g_n) \sim \mathcal{N}(0, \mathrm{I}_n)$. Then

$$\|g\|_2^2 = \sum_{i=1}^{n} g_i^2$$

is the sum of n independent random variables, and by the central limit theorem we expect that, as $n \to \infty$, $\|g\|_2^2$ strongly concentrates on its average $\mathbb{E}[\|g\|_2^2] = n$. But unfortunately Hoeffding's inequality does not apply here, because $\mathbb{P}[g_i^2 \geq t] = \mathbb{P}[|g_i| \geq \sqrt{t}] \sim e^{-t/2}$ for large t, so clearly an inequality of the type of eq. (19) will not be true for g_i^2.

Instead, we try to generalize the approach we used to obtain Hoeffding's inequality to weaker tails, which behave like $\exp(-Ct)$.

Challenge 4.6. *Let X be a symmetric random variable such that $\mathbb{P}[|X| \geq t] = e^{-t}$. Show*

(i) $\mathbb{E}[\exp(\lambda X)] \leq \exp 2\lambda^2$ *for all* $|\lambda| \leq 1/2$.

(ii) $\|X\|_p = \mathbb{E}[|X|^p]^{1/p} \leq Cp$, *for some constant* $C > 0$, *and all* $p \geq 1$.

Like the properties of Gaussian variables motivated the definition of sub-Gaussian random variables, Challenge 4.6 motivates the definition of **sub-exponential** random variables.

Definition 4.2 (*Sub-exponential random variable*)

A centered random variable X is sub-exponential if it satisfies one of the following three conditions.

(i) **(Tail)** For all $t > 0$, $\mathbb{P}[|X| \geq t] \leq 2\exp\{-t/K_1\}$, for some $K_1 > 0$.

(ii) **(MGF)** For all $|\lambda| \leq 1/K_2$, $\mathbb{E}[\exp\{\lambda X\}] \leq \exp\{\lambda^2 K_2^2\}$, for some $K_2 > 0$.

(iii) **(Moments)** For all $p \geq 1$, $\|X\|_p := [\mathbb{E}|X|^p]^{1/p} \leq K_3 p$, for some $K_3 > 0$.

We will say that X is ν-sub-exponential (or $\mathrm{SE}(\nu)$) if $\mathbb{E}[\exp\{\lambda X\}] \leq \exp\{\lambda^2\nu^2\}$ for all $|\lambda| \leq 1/\nu$.

Challenge 4.7. *Check that the conditions (i), (ii), (iii) in Definition 4.2 are equivalent, and that (K_1, K_2, K_3) differ by at most an absolute multiplicative constant.*

Notice that the MGF condition for sub-exponential random variables is very similar to the one of sub-Gaussian ones, but (crucially) the bound only holds for λ in a neighborhood of zero. This also directly shows that a ν-sub-Gaussian random variable is also ν-sub-exponential.

Challenge 4.8. *Let X, Y be two sub-Gaussian random variables (not necessarily independent). Show that XY is sub-exponential.*

We are now ready to state our main concentration result for the sum of sub-exponential random variables. It is named Bernstein's inequality, although it is quite different from the version of Bernstein's inequality stated in [BSS23] (which is instead a refinement of Hoeffding's inequality for bounded random variables).

Theorem 4.2 (*Bernstein's inequality*)

Let $X_1, \cdots, X_n$ be independent and centered sub-exponential random variables, with sub-exponential parameters $\nu_1, \cdots, \nu_n$. Then, for all $a \in \mathbb{R}^n$, for all $t > 0$:

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2\exp\left\{-\min\left(\frac{t^2}{4\sum_{i=1}^n a_i^2\nu_i^2}, \frac{t}{2\max_{i\in[n]}(\nu_i|a_i|)}\right)\right\}.$$

Proof of Theorem 4.2 – We again apply the Chernoff method. Notice however that, for each $i \in [n]$, $\mathbb{E}[e^{\lambda a_i X_i}]$ only exists if $|\lambda| < 1/(\nu_i|a_i|)$. We can still apply the method for $0 \leq \lambda \leq 1/\max(|a_i|\nu_i)$, which gives:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) &\leq \inf_{0\leq\lambda\leq 1/\max(|a_i|\nu_i)}\left[e^{-\lambda t}\prod_{i=1}^n \mathbb{E}[e^{\lambda a_i X_i}]\right], \\ &\leq \exp\left\{\inf_{0\leq\lambda\leq 1/\max(|a_i|\nu_i)}\left[-\lambda t + \lambda^2\sum_{i=1}^n a_i^2\nu_i^2\right]\right\}\end{aligned}$$

We are now looking for the minimum of a degree-2 polynomial of the type $-\lambda t + a\lambda^2$ (with $a > 0$), over the interval $[0, b]$. It is easy to show that

$$\begin{aligned}\min_{\lambda\in[0,b]}[-\lambda t + a\lambda^2] &= \begin{cases} -\frac{t^2}{4a} & \text{if } t \leq 2ab, \\ -bt + ab^2 & \text{if } t \geq 2ab. \end{cases}, \\ &\leq -\min\left\{\frac{t^2}{4a}, \frac{bt}{2}\right\}.\end{aligned}$$

So we reach:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) &\leq \inf_{0\leq\lambda\leq 1/\max(|a_i|\nu_i)}\left[e^{-\lambda t}\prod_{i=1}^n \mathbb{E}[e^{\lambda a_i X_i}]\right], \\ &\leq \exp\left\{-\min\left(\frac{t^2}{4\sum_{i=1}^n a_i^2\nu_i^2}, \frac{t}{2\max_{i\in[n]}(|a_i|\nu_i)}\right)\right\}.\end{aligned}$$

Using as usual the argument on $-X$ and the union bound ends the proof. □

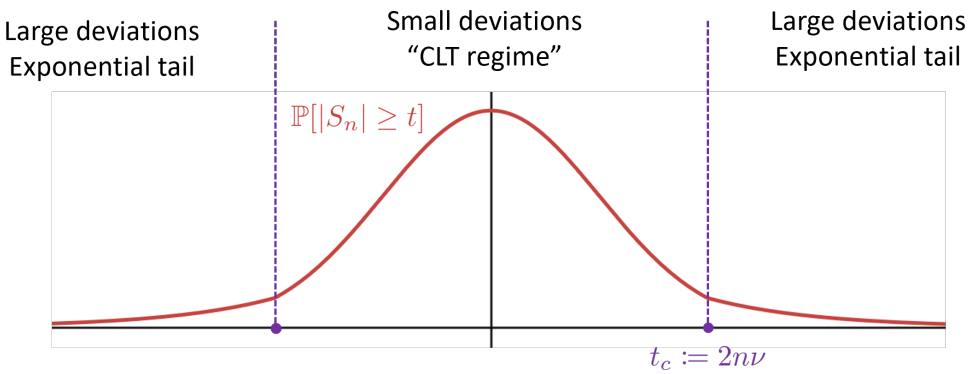
Challenge 4.9. *Apply Bernstein's inequality to obtain a concentration bound for $\|g\|_2^2$, the squared norm of a random vector with i.i.d. $\mathcal{N}(0,1)$ coordinates.*

4.3 The two-tailed behavior

The most striking feature of Bernstein's inequality is that it exhibits a "two-tails" behavior. As we saw in Hoeffding's inequality (Theorem 4.1), the sum of independent $\mathrm{SG}(\sigma)$ random variables has a purely sub-Gaussian tail. However, if we assume that $X_i \in \mathrm{SE}(\nu)$ are independent, and $S_n = \sum_{i=1}^n X_i$, then Bernstein's inequality gives instead

$$\mathbb{P}(|S_n| \geq t) \leq 2\exp\left(-\min\left\{\frac{t^2}{4n\nu^2}, \frac{t}{2\nu}\right\}\right).$$

Essentially, the tail of S_n is bounded as follows.



The central limit theorem – We know by the central limit theorem that $S_n/\sqrt{n}$ converges in law to a Gaussian random variable, so it is not surprising that $\mathbb{P}[|S_n| \geq t]$ is governed by a Gaussian tail for $t = \mathcal{O}(\sqrt{n})$. For sub-Gaussian random variables, Hoeffding's inequality shows that this is true for all $t > 0$ (even growing with n faster than $\sqrt{n}$), while Bernstein's inequality shows that this still holds all the way to $t \leq t_c := 2n\nu$, which is of order $\mathcal{O}(n)$!

Large deviations regime – For $t > t_c := 2n\nu$, the tail of S_n is upper bounded by $2\exp(-t/2\nu)$. Notice that if we consider a single X_i, then $\mathbb{P}[|X_i| \geq t]$ is bounded by $2\exp(-Ct/\nu)$ for large t. This is not a coincidence: the bound of S_n in the "large deviations" regime of Bernstein's inequality is driven by very rare[4] situations in which a *very small number of X_i's* are so large that they influence the total sum. On the other hand, while such events can occur for sub-Gaussian random variables, their probability is so low that they never dominate the large deviations of S_n.

4.4 Other tails and concentration inequalities

One can also build concentration inequalities for other tails than sub-exponentials, for instance tails of the type $\exp(-t^\alpha)$ for $\alpha \in (0, 2]$ (they are called *sub-Weibull tails*), see for instance Lemmas 3.5 and 3.6 of [Ada+11], and references therein. They are proved:

- Using the usual Chernoff method when $\alpha \geq 1$, so that the MGF $\mathbb{E}[\exp(\lambda X)]$ exists for λ close to 0.
- By using Markov's inequality on the moments of order p when $\alpha < 1$, and optimizing over p rather than λ (since $\alpha < 1$, the MGF $\mathbb{E}[\exp(\lambda X)]$ in general does not exist for any λ).

Finally, let us mention that [BLM13] is a great reference on concentration inequalities in general.

[4] Here $t \gtrsim n$, so we are only talking about events with exponentially small probability!

5 The BBP transition

Disclaimer – Much like [BSS23], this section will contain some mathematically informal steps. However, (*i*) some of the main definitions require a good knowledge of probability theory, although the main examples can still be understood without it, and (*ii*) we will give references to papers and textbooks where all steps are properly carried away (some of them being quite tedious, although the idea behind them will be very clear). A good mathematical reference containing all the theorems we will state in this section (and many more) is [BN11]. Regarding random matrix theory in general, [AGZ10] is an amazing mathematical textbook.

Our goal here will be to generalize the picture of the "BBP" transition (named after the authors of [BBP05]) that was introduced in class [BSS23]. We focus on the additive perturbation case.

5.1 The model

5.1.1 Formal definition

We (the statistician) are given an observation under the form of a symmetric matrix $\mathbf{Y}$, built as:

$$\mathbf{Y} = \mathbf{Z} + \theta \mathbf{x}\mathbf{x}^\top \in \mathbb{R}^{d\times d}. \tag{20}$$

- $\mathbf{Z} \in \mathbb{R}^{d\times d}$ is a matrix that will typically represent the "noise", and that we will assume to be random. The canonical example is just a matrix with i.i.d. elements drawn from a Gaussian distribution (more on that later).
- $\mathbf{x} \in \mathbb{R}^d$ is our "signal". We assume that $\|\mathbf{x}\|_2 = 1$, and our goal is to recover (or "infer") the value of $\mathbf{x}$ – or at least some partial information about $\mathbf{x}$ – from the observation of $\mathbf{Y}$.
- $\theta \geq 0$ is the "signal-to-noise ratio" (or SNR). We expect that, as θ increases, recovery of $\mathbf{x}$ becomes more and more feasible, since for $\theta \gg 1$, $\mathbf{Y}$ should approach the rank-one matrix $\theta\mathbf{x}\mathbf{x}^\top$.

Our main goal in Section 5 is to answer the following question:

Does the top eigenvector $v_{\max}(\mathbf{Y})$ contain information about $\mathbf{x}$?

In order to carry out our mathematical analysis, we make the following assumptions on the noise matrix $\mathbf{Z}$ (*warning:* these assumptions can be mathematically heavy, have a look at the examples given later on to get a good intuition!).

Model 5.1 (*Noise model*)

Denote $\mathrm{Sp}(\mathbf{Z}) = \{z_1, \cdots, z_d\}$, with $z_1 \geq \cdots \geq z_d$. We assume:

H.1 The empirical spectral distribution converges[5] to a limit probability measure μ:

$$\hat{\mu}_{\mathbf{Z}} := \frac{1}{d}\sum_{i=1}^{d} \delta_{z_i} \xrightarrow[d\to\infty]{\text{weakly}} \mu \quad \text{(a.s.)}$$

Moreover, we assume that μ has compact support, and that $\mathrm{supp}(\mu) = [a, b]$ for some $a < b$.

H.2 The top eigenvalue z_1 of $\mathbf{Z}$ converges to the right edge of the support of μ:

$$z_1 = \max_{i\in[d]} z_i \xrightarrow[d\to\infty]{} b \quad \text{(a.s.)}$$

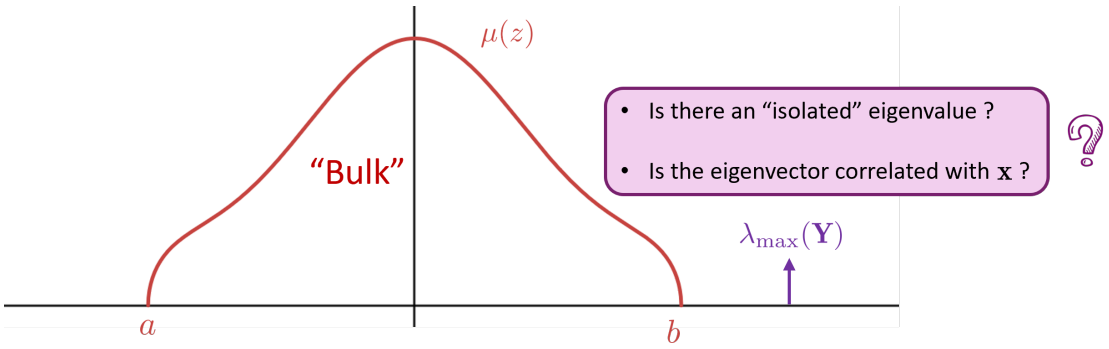


Figure 1: Very schematic view of the question we want to answer regarding the model of eq. (20).

H.3 We assume that the eigenvectors of $\mathbf{Z}$ are *completely delocalized*. Concretely, we assume that $\mathbf{Z}$ is *rotationally-invariant*: for any orthogonal matrix $\mathbf{O} \in \mathcal{O}(d)$, $\mathbf{Z} \overset{\mathrm{d}}{=} \mathbf{OZO}^\top$ ($\overset{\mathrm{d}}{=}$ means equality of distributions).

We give in Fig. 1 a schematic view of our problematic, having now introduced the main quantities of interest regarding the spectrum of $\mathbf{Z}$. Let us make some remarks on some of the (heavy) definitions of Model 5.1.

- H.1 is sometimes called the convergence of the "bulk": the spectrum of $\mathbf{Z}$ admits a limiting curve. The assumption that $\mathrm{supp}(\mu) = [a, b]$ is a single interval is not necessary, but it will help fix our ideas.
- H.2 is very important: indeed, it ensures that no eigenvalue of $\mathbf{Z}$ is "running away" from the support of the limiting distribution μ as d gets large. Therefore, if we see a single isolated eigenvalue away from $[a, b]$ (see Fig. 1), it must be present in $\mathbf{Y}$ but not in $\mathbf{Z}$!
- H.3 means that the eigenvectors of $\mathbf{Z}$ are not skewed towards any "special direction". It mathematically implies that the eigenvectors of $\mathbf{Z}$ can are uniformly sampled from the orthogonal group $\mathcal{O}(d)$. We stress that H.3 can be greatly relaxed, to much weaker assumptions than rotation invariance. Moreover, one can even completely drop randomness assumptions on the eigenvectors of $\mathbf{Z}$, by assuming instead that the signal $\mathbf{x}$ is randomly sampled (independently of $\mathbf{Z}$).

We refer to [BN11] for more details on all these points and further generalizations.

5.1.2 Examples of noise matrices

Model 5.1 is a quite generic assumption, and you have already seen two very important examples in class [BSS23].

Gaussian orthogonal ensemble – If $\mathbf{Z}$ is a centered symmetric $d \times d$ random matrix, with

$$\begin{cases} Z_{ij} \sim \mathcal{N}(0, 1/d) & \text{for } i < j, \\ Z_{ii} \sim \mathcal{N}(0, 2/d), \end{cases} \tag{21}$$

then we say that $\mathbf{Z}$ is drawn from the *Gaussian orthogonal ensemble* (or GOE(d)). The difference in normalization between diagonal and off-diagonal elements in eq. (21) is a

[5] Don't be confused by the mix of weak and almost sure convergence: the convergence happens almost surely, but the convergence itself is the weak convergence of measures.

convention, which implies the nice fact that the probability density of $\mathbf{Z}$ can be written (up to a constant) in the compact form:

$$\varphi(\mathbf{Z}) \propto \exp\left\{-\frac{d}{4}\mathrm{Tr}[\mathbf{Z}^2]\right\}.$$

The seminal work of Wigner [Wig55], that can be seen as the start of random matrix theory, proves that the GOE(d) ensemble satisfies Model 5.1 (more precisely H.1 and H.2), with μ given by the *semicircle* distribution:

$$\mu_{\text{s.c.}}(x) := \frac{\sqrt{4-x^2}}{2\pi}\mathbb{1}\{|x| \le 2\}. \tag{22}$$

Challenge 5.1. *Check that the* GOE(d) *distribution satisfies H.3 of Model 5.1.*

The GOE(d) distribution will be the driving example of the BBP transition, and is a good example to have in mind in what follows.

Wishart matrices – You have seen in class [BSS23] that so-called "Wishart" matrices [Wis28] of the type $\mathbf{Z} = (1/n)\mathbf{X}\mathbf{X}^\top$, where $\mathbf{X} \in \mathbb{R}^{d\times n}$ has i.i.d. $\mathcal{N}(0,1)$ elements, and where $d = \gamma n$ for $\gamma \in (0,1]$, satisfy Model 5.1, with μ given by the *Marchenko-Pastur* distribution [MP67]:

$$\mu_{\text{MP},\gamma}(x) := \frac{\sqrt{(\gamma_+ - x)(x-\gamma_-)}}{2\pi\gamma x}\mathbb{1}\{\gamma_- \le x \le \gamma_+\}, \tag{23}$$

where $\gamma_\pm := (1 \pm \sqrt{\gamma})^2$.

Challenge 5.2. *Check that the Wishart distribution satisfies H.3 of Model 5.1. (After understanding Challenge 5.1, the proof should take you one line!)*

Challenge 5.3. *Simulate numerically the* GOE(d) *and Wishart distributions, and compare the histogram of eigenvalues to the limiting curves we described.*

Generic densities – More generally, given any probability measure μ with support $[a,b]$, we can engineer a random matrix ensemble satisfying Model 5.1. Indeed, it suffices to draw $z_1, \cdots, z_d \overset{\text{i.i.d.}}{\sim} \mu$, and to build $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{O}\mathrm{Diag}(\{z_i\}_{i=1}^d)\mathbf{O}^\top, \tag{24}$$

where we draw $\mathbf{O}$ from the uniform measure (also called *Haar measure*) on the orthogonal group $\mathcal{O}(d)$, i.e. $\mathbf{O}$ is a random orthogonal matrix. The following can then be checked.

Challenge 5.4. *Prove that the model of eq. (24) satisfies Model 5.1, if μ has support $[a,b]$.*

5.2 The general BBP transition

5.2.1 The top eigenvalue

Let y be an eigenvalue of $\mathbf{Y}$, defined in eq. (20). Then

$$\det[y\mathrm{I}_d - (\mathbf{Z} + \theta\mathbf{x}\mathbf{x}^\top)] = 0.$$

If we assume that y is not an eigenvalue of $\mathbf{Z}$, we can multiply this by $\det[(y\mathrm{I}_d - \mathbf{Z})^{-1}]$, which yields;

$$\det[\mathrm{I}_d - \theta\mathbf{x}\mathbf{x}^\top(y\mathrm{I}_d - \mathbf{Z})^{-1}] = 0,$$

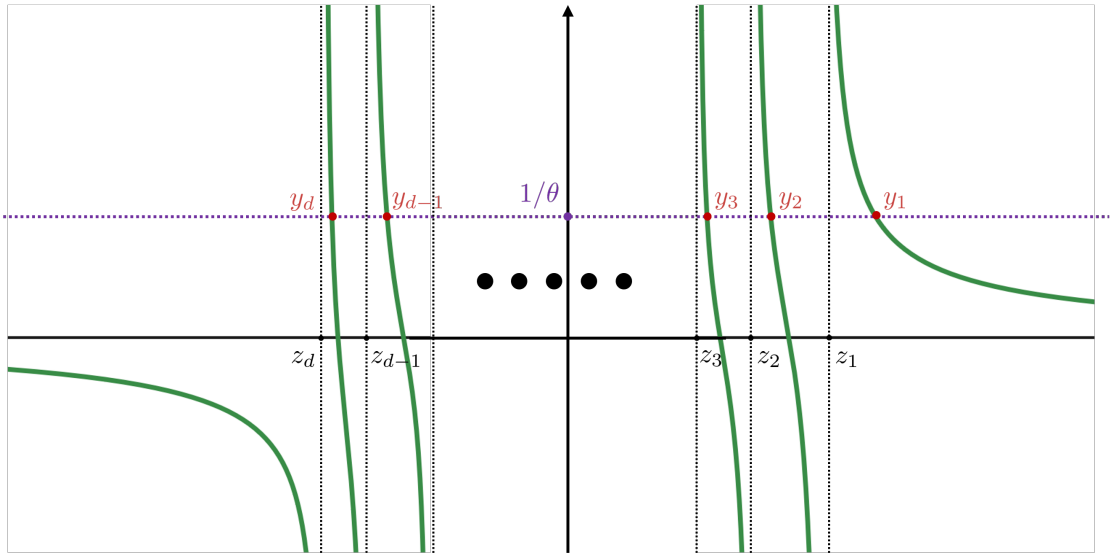
i.e. 1 is an eigenvalue of $\mathbf{M} \coloneqq \theta \mathbf{x}\mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-1}$. Notice that $\mathbf{M}$ is a rank-one matrix, so it has a single non-zero eigenvalue, which is also equal to $\mathrm{Tr}[\mathbf{M}] = \theta \mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-1}\mathbf{x}$. In the end, we get that $y \in \mathrm{Sp}(\mathbf{Y}) \backslash \mathrm{Sp}(\mathbf{Z})$ if and only if

$$\frac{1}{\theta} = \mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-1}\mathbf{x}. \tag{25}$$

We denote $\mathbf{w}_1, \cdots, \mathbf{w}_d$ the eigenvectors of $\mathbf{Z}$, and $\mathbf{x} = \sum_{i=1}^d \alpha_i \mathbf{w}_i$. Because $\|\mathbf{x}\|_2 = 1$ and we have H.3, it is easy to see that $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)$ is uniformly sampled from the unit sphere $\mathcal{S}^{d-1}$, and is independent of the eigenvalues $z_1 \leq \cdots \leq z_d$ of $\mathbf{Z}$. Moreover, eq. (25) reads:

$$\frac{1}{\theta} = \sum_{i=1}^d \frac{\alpha_i^2}{y - z_i}. \tag{26}$$

It is easy to see that if all $\alpha_i \neq 0$ (which happens with high probability since $\boldsymbol{\alpha} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$), then eq. (26) has exactly d solutions $y_1 \geq \cdots \geq y_d$, and they satisfy $y_i \geq z_i \geq y_{i+1}$ for all $i \in [d]$ (with $y_{d+1} \coloneqq -\infty$). We draw the behavior of the right hand side of eq. (26) in green in the figure that follows, highlighting the solutions $(y_1, \cdots, y_d)$ to this equation.



Since we have d solutions, this means we have characterized all the eigenvalues of $\mathbf{Y}$!

Remark – Notice that one could also have deduced that the eigenvalues of $\mathbf{Y}$ are positively shifted with respect to the ones of $\mathbf{Z}$ from Weyl's interlacing inequalities, although it would not have given eq. (26).

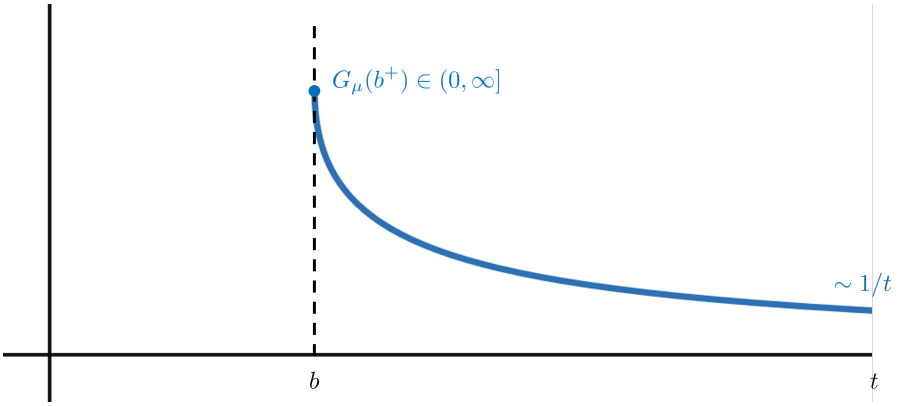
We are interested here in the behavior of the top eigenvalue of $\mathbf{Y}$, and we now give a heuristic argument for the value of its limit. The mathematically rigorous approach follows the same line, and can be found e.g. in [BN11]. We assume that the largest eigenvalue y_1 of $\mathbf{Y}$ approaches, as $d \to \infty$, a deterministic value λ, and thus $\lambda \geq b$ since $\lambda > z_1$ and $z_1 \to b$ by H.2. Since $\boldsymbol{\alpha} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$, for any $\lambda > b$, eq. (26) concentrates, as $d \to \infty$, around its average:

$$\frac{1}{\theta} \simeq \mathbb{E} \sum_{i=1}^d \frac{\alpha_i^2}{\lambda - z_i} = \frac{1}{d} \sum_{i=1}^d \frac{1}{\lambda - z_i} = \int \hat{\mu}_{\mathbf{Z}}(\mathrm{d}z) \frac{1}{\lambda - z} \underset{\text{(a)}}{\simeq} \int_a^b \mu(\mathrm{d}z) \frac{1}{\lambda - z}.$$

We used H.1 in (a), and $\simeq$ means here "equality up to $o_d(1)$" terms. In the end, we reach the following equation characterizing the limit λ, if we have $\lambda > b$:

$$G_\mu(\lambda) \coloneqq \int \mu(\mathrm{d}z) \frac{1}{\lambda - z} = \frac{1}{\theta}. \tag{27}$$

The function $G_\mu(t)$ (for $t > b$) defined in eq. (27) is called the *Stieltjes (or Cauchy) transform* of μ. It is strictly decreasing, and behaves like $1/t$ for large $t \gg 1$.



The conclusion of the argument above is that the top eigenvalue of $\mathbf{Y}$ approaches a value $\lambda > b$ if and only if $G_\mu(\lambda) = 1/\theta$, i.e. if $\theta > \theta_c := 1/G_\mu(b^+)$. Notice that $\theta_c \geq 0$, and we can have $\theta_c = 0$ if $G_\mu(b^+) = \infty$: in this case, for any $\theta > 0$ there is an outlier eigenvalue in the spectrum of $\mathbf{Y}$, as $d \to \infty$. On the other hand, if $\theta \leq \theta_c$, there is no solution $\lambda > b$ to eq. (27), and therefore the top eigenvalue y_1 of $\mathbf{Y}$ approaches b as $d \to \infty$ (since $y_1 \geq z_1$ and $z_1 \to b$). We can summarize this in the following theorem.

Theorem 5.1 (*General "BBP" transition for the top eigenvalue*)

Assume we are considering a noise matrix $\mathbf{Z}$ satisfying Model 5.1, and recall that $\mathbf{Y} = \mathbf{Z} + \theta \mathbf{x}\mathbf{x}^\top$, and $\|\mathbf{x}\|_2 = 1$. Denote $y_1 \geq \cdots \geq y_d$ the eigenvalues of $\mathbf{Y}$. Then:

- If $\theta > \theta_c := 1/G_\mu(b^+)$, then $y_1 \xrightarrow[d\to\infty]{} G_\mu^{-1}(1/\theta)$ (a.s.) and $y_2 \xrightarrow[d\to\infty]{} b$ (a.s.)
- If $\theta \leq \theta_c := 1/G_\mu(b^+)$, then $y_1 \xrightarrow[d\to\infty]{} b$ (a.s.)

Theorem 5.1 informs us on the limit of the top eigenvalue of $\mathbf{Y}$, and answers the first question laid out in Fig. 1. We will apply Theorem 5.1 to a concrete example in Section 5.2.3.

5.2.2 Correlation with the signal

We now consider the correlation of the associated eigenvector with the signal $\mathbf{x}$. Let $(\mathbf{v}, y)$ be an eigenvector-eigenvalue pair for $\mathbf{Y}$, such that $y \in \mathrm{Sp}(\mathbf{Y}) \backslash \mathrm{Sp}(\mathbf{Z})$. The relation $\mathbf{Y}\mathbf{v} = y\mathbf{v}$ yields:

$$(y\mathrm{I}_d - \mathbf{Z})\mathbf{v} = \theta \langle \mathbf{x}, \mathbf{v} \rangle \mathbf{x}.$$

Multiplying this equation by $(y\mathrm{I}_d - \mathbf{Z})^{-1}$:

$$\mathbf{v} = \theta \langle \mathbf{x}, \mathbf{v} \rangle (y\mathrm{I}_d - \mathbf{Z})^{-1}\mathbf{x}.$$

This equation still involves $\mathbf{v}$ on both sides. However since $\mathbf{v}$ is normalized, we can write:

$$\mathbf{v} = \pm \frac{(y\mathrm{I}_d - \mathbf{Z})^{-1}\mathbf{x}}{\sqrt{\mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-2}\mathbf{x}}}.$$

And in particular the correlation $\langle \mathbf{x}, \mathbf{v} \rangle$ satisfies:

$$\langle \mathbf{x}, \mathbf{v} \rangle^2 = \frac{\left(\mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-1}\mathbf{x}\right)^2}{\mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-2}\mathbf{x}}.$$

By eq. (25), we can further simplify it into:

$$\langle \mathbf{x}, \mathbf{v} \rangle^2 = \left(\theta^2 \mathbf{x}^\top (y \mathrm{I}_d - \mathbf{Z})^{-2} \mathbf{x} \right)^{-1}. \tag{28}$$

We can analyze the limit of eq. (28) in a very similar way to what we did to analyze the limit of $\mathbf{x}^\top (y\mathrm{I}_d - \mathbf{Z})^{-1}\mathbf{x}$ in eq. (25), which yielded eq. (27) as $d \to \infty$. Concretely, we get (convince yourself!) that, if $\lambda > b$, $\mathbf{x}^\top (\lambda \mathrm{I}_d - \mathbf{Z})^{-2}\mathbf{x} \simeq_{d\to\infty} -G'_\mu(\lambda)$. Moreover, if we assume that $\theta \leq \theta_c$ and that $G'_\mu(b^+) = -\infty$, then we get that $\langle \mathbf{v}, \mathbf{x} \rangle^2 \to 0$. We can summarize this in the following theorem.

Theorem 5.2 (*General "BBP" transition for the top eigenvector*)

Assume we are considering a noise matrix $\mathbf{Z}$ satisfying Model 5.1, and recall that $\mathbf{Y} = \mathbf{Z} + \theta \mathbf{x}\mathbf{x}^\top$, and $\|\mathbf{x}\|_2 = 1$. Denote $\mathbf{v}_{\max}$ the eigenvector of $\mathbf{Y}$ associated to its largest eigenvalue y_1. Then:

- If $\theta > \theta_c := 1/G_\mu(b^+)$, then $\langle \mathbf{v}_{\max}, \mathbf{x} \rangle^2 \xrightarrow[d\to\infty]{} (-\theta^2 G'_\mu(\lambda))^{-1}$ (a.s.), where $\lambda = G_\mu^{-1}(1/\theta)$ is the limit of the top eigenvalue, cf. Theorem 5.1.
- If $\theta \leq \theta_c := 1/G_\mu(b^+)$ **and** $G'_\mu(b^+) = -\infty$, then $\langle \mathbf{v}_{\max}, \mathbf{x} \rangle^2 \xrightarrow[d\to\infty]{} 0$ (a.s.)

Notice that for $\theta \to \infty$, since $G_\mu(t) \sim 1/t$, we have $\lambda \sim \theta$. Thus $G'_\mu(\lambda) \sim G'_\mu(\theta) \sim -1/\theta^2$, so we obtain that for $\theta \to \infty$, the correlation $\langle \mathbf{v}_{\max}, \mathbf{x} \rangle^2$ approaches 1: for very large SNR we recover the signal almost perfectly!

5.2.3 The canonical example: GOE matrices

Let us now apply Theorems 5.1 and 5.2 on concrete examples. The canonical example in this problem, which you should always have in mind, is when $\mathbf{Z}$ is drawn from the GOE(d) distribution, see eq. (21). We saw that then the empirical spectrum converges to the semicircle law, and we can analytically compute its Stieltjes transform (the solution to this challenge is given in Section 5.4).

Challenge 5.5. *Show that if $\mu(\mathrm{d}x) = [\sqrt{4 - x^2}/(2\pi)]\mathbb{1}\{|x| \leq 2\}$ is the semicircle law, its Stieltjes transform satisfies, for all $t > 2$:*

$$G_\mu(t) = \frac{t - \sqrt{t^2 - 4}}{2}. \tag{29}$$

(Hint: try to write it as an integral over the complex unit circle, and use the residue theorem)

In particular, $\theta_c = 1/G_\mu(2^+) = 1$. Moreover $G_\mu^{-1}(s) = s + 1/s$ for $s \leq 1$. Applying Theorem 5.1, we get:

$$\lim_{d\to\infty} \lambda_{\max}(\mathbf{Y}) = \begin{cases} 2 & \text{if } \theta \leq 1, \\ \theta + \dfrac{1}{\theta} & \text{if } \theta > 1. \end{cases} \tag{30}$$

Let us now look at the correlation of the top eigenvector, i.e. Theorem 5.2. Since $G'_\mu(t) = (1/2) - t/[2\sqrt{t^2 - 4}]$, we have $G'_\mu(2^+) = -\infty$. You can then check the following easily from Theorem 5.2.

$$\lim_{d\to\infty} \langle \mathbf{v}_{\max}, \mathbf{x} \rangle^2 = \begin{cases} 0 & \text{if } \theta \leq 1, \\ 1 - \dfrac{1}{\theta^2} & \text{if } \theta > 1. \end{cases} \tag{31}$$

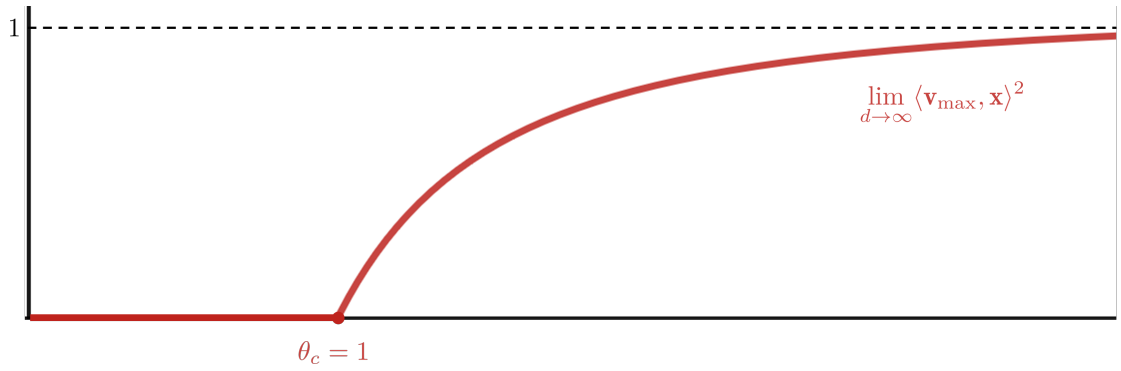


Figure 2: The limiting overlap in the BBP transition, cf. eq. (31).

For $\theta < \theta_c = 1$, the "spike" $\theta \mathbf{x}\mathbf{x}^\top$ has no influence on the spectrum of $\mathbf{Y}$, and its top eigenvector is not correlated at all with $\mathbf{x}$. As θ grows and crosses the critical value $\theta_c = 1$, an eigenvalue pops out of the spectrum (cf Fig. 1) and its associated eigenvector becomes positively correlated with $\mathbf{x}$! This phenomenon is what is usually referred to as the "BBP transition".

Challenge 5.6. *Simulate numerically the BBP transition for* $\mathbf{Z}$ *a* GOE(d) *matrix, and check the validity of eqs.* (30) *and* (31).

5.2.4 The behavior of the density at the edge

If we assume that $\mu(x) \sim c(b-x)^\alpha$ for $x \uparrow b$, for some $c > 0$ and $\alpha > -1$ (so that μ is integrable), then the fact that $G_\mu(b^+)$ and $G'_\mu(b^+)$ are finite solely depends on α. By a straightforward computation, we have in this case:

(*i*) $G_\mu(b^+) < -\infty \Leftrightarrow \alpha > 0$. So $\theta_c > 0 \Leftrightarrow \alpha > 0$. And for $\alpha \neq 0$, there is no transition: as soon as $\theta > 0$, an isolated eigenvalue pops out of the spectrum!

(*ii*) $G'_\mu(b^+) = -\infty \Leftrightarrow \alpha \leq 1$. In this case, there is a sharp transition for the correlation, according to Theorem 5.2: the correlation is 0 for $\theta \leq \theta_c$, and positive for $\theta > \theta_c$.

Notice that the two examples (GOE and Wishart matrices) we saw in Section 5.1.2 satisfy this condition with $\alpha = 1/2$: for these random matrix models, $\theta_c > 0$, and there is a sharp transition for the correlation!

5.3 Generalization to multiple spikes

All the picture we described can be generalized to the case of "multi-spike" models, i.e. we consider instead

$$\mathbf{Y} = \mathbf{Z} + \sum_{i=1}^{r} \theta_i \mathbf{x}_i \mathbf{x}_i^\top, \tag{32}$$

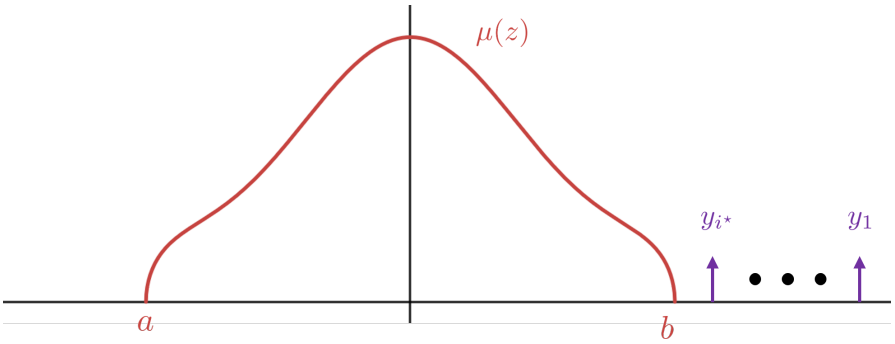
for some $r \geq 1$ (fixed as $d \to \infty$), and we can assume without loss of generality that the $\mathbf{x}_i$'s are orthonormal vectors and that $\theta_1 \geq \cdots \geq \theta_r$. We get the following result.

Theorem 5.3 (*"Multi-spike" BBP transition, eigenvalues*)

Assume we are considering a noise matrix $\mathbf{Z}$ satisfying Model 5.1, and recall that $\mathbf{Y}$ is generated from eq. (32), and has eigenvalues $y_1 \geq \cdots \geq y_d$. Let $\theta_c := 1/G_\mu(b^+)$, and let $i^\star \in \{0, \cdots, r\}$ such that $\theta_{i^\star} > \theta_c \geq \theta_{i^\star+1}$. Then:

- For all $i \in \{1, \cdots, i^\star\}$, $y_i \xrightarrow[d\to\infty]{} G_\mu^{-1}(1/\theta_i)$ (a.s.)

- For all $i \in \{i^\star, \cdots r\}$, $y_i \xrightarrow[d\to\infty]{} b$ (a.s.)



Informally, everything happens as if the different spikes in eq. (32) each had its own independent BBP transition! One can formulate similar results for the eigenvectors correlations: things are a bit more technical there if there is degeneracy in the spiked matrix, i.e. if $\theta_i = \theta_j$, and we refer to [BN11] for a precise theorem on the limit of the correlations of eigenvectors of $\mathbf{Y}$ to the signals in the multi-spike case.

5.4 Solution to challenges

Solution of Challenge 5.5 – Let $t > 2$. Changing variables to $x = 2\cos\theta$ we get:

$$\begin{aligned} G_\mu(t) &= \frac{2}{\pi}\int_0^\pi \frac{\sin^2\theta}{t - 2\cos\theta}\mathrm{d}\theta, \\ &= \frac{1}{\pi}\int_{-\pi}^\pi \frac{\sin^2\theta}{t - 2\cos\theta}\mathrm{d}\theta. \end{aligned}$$

Writing $\zeta = e^{i\theta}$, this can be written as:

$$\begin{aligned} G_\mu(t) &= \frac{1}{\pi}\oint_{|\zeta|=1}\left(\frac{\zeta - \zeta^{-1}}{2i}\right)^2 \frac{1}{t - (\zeta + \zeta^{-1})}\frac{\mathrm{d}\zeta}{i\zeta}, \\ &= \frac{1}{4i\pi}\oint_{|\zeta|=1}\frac{(\zeta^2-1)^2}{\zeta^2(\zeta^2 - t\zeta + 1)}\mathrm{d}\zeta. \end{aligned} \tag{33}$$

The integrand in eq. (33) has three poles, in $\zeta \in \{0, (t \pm \sqrt{t^2-4})/2\}$. Since $t > 2$, the only two poles inside the unit circle are 0 and $(t - \sqrt{t^2-4})/2$, and they respectively have residues

$$\begin{aligned} \mathrm{Res}_0\left[\frac{(\zeta^2-1)^2}{\zeta^2(\zeta^2 - t\zeta + 1)}\right] &= t, \\ \mathrm{Res}_{(t-\sqrt{t^2-4})/2}\left[\frac{(\zeta^2-1)^2}{\zeta^2(\zeta^2 - t\zeta + 1)}\right] &= -\sqrt{t^2-4}. \end{aligned}$$

Using the residue theorem in eq. (33), we finally find

$$G_\mu(t) = \frac{t - \sqrt{t^2-4}}{2},$$

which ends the proof. □

6 Spiked matrix models: hypothesis testing and contiguity

6.1 Motivation

In this part, we are still motivated by the problem of Section 5:

$$\text{Given } \mathbf{Y} = \mathbf{Z} + \theta \mathbf{x}\mathbf{x}^\top, \text{ can we recover information on } \mathbf{x} \in \mathbb{R}^d, \text{ when } d \gg 1?$$

Important: In this section, we will only consider the Gaussian noise case: here $\mathbf{Z} \sim \mathrm{GOE}(d)$, see eq. (21).

In Section 5 we completely characterized the performance of one estimator for $\mathbf{x}$, that we call the PCA estimator $\hat{\mathbf{x}}_{\mathrm{PCA}}(\mathbf{Y}) \coloneqq \mathbf{v}_{\max}(\mathbf{Y})$ (the top eigenvector): this is the BBP transition for the eigenvector, given in eq. (31). However it is not clear *a priori* that taking the largest eigenvector is the best thing to do to recover $\mathbf{x}$ from $\mathbf{Y}$, and this is precisely the question we ask here:

(Q) *Is it possible to have $\hat{\mathbf{x}}(\mathbf{Y})$ that recovers information about $\mathbf{x}$ when the PCA estimator fails?*

As we will see, the answer to this question greatly depends on the prior knowledge we (as a statistician trying to infer $\mathbf{x}$) have on the signal $\mathbf{x}$. A good example to keep in mind is if $\mathbf{x} \in \{\pm 1/\sqrt{d}\}^d$: the vector $\mathbf{x}$ then belongs to a very specific subset of the unit sphere, and perhaps one can hope to leverage this information to improve over the PCA estimator!

Weak recovery – The success of an estimator will be measured by its correlation: informally, if $\lim_{d\to\infty} \langle \hat{\mathbf{x}}(\mathbf{Y}), \mathbf{x}\rangle^2 > 0$, we say that the estimator $\hat{\mathbf{x}}(\mathbf{Y})$ *weakly recovers* the signal $\mathbf{x}$.

The goal of this section – In this section, we will mainly focus on *lower bounds*, i.e. we will try to prove that (in some regimes to specify), it is *impossible to recover $\mathbf{x}$ with any procedure*. On the other hand, there is a very rich line of work on designing practical algorithms that can improve over the PCA estimator: we will mention a few later on, but it is not the topic here.

A very good reference on the topics of this section is [Per+18], and it also contains a thorough description of the rich literature on spiked matrix models (at the time of its writing).

6.2 Distinguishability, contiguity, and the likelihood ratio

Clearly, a necessary condition to weakly recover $\mathbf{x}$ is to be able to *distinguish* the matrix $\mathbf{Y} = \mathbf{Z} + \theta\mathbf{x}\mathbf{x}^\top$ from a pure sample of the noise $\mathbf{Y} = \mathbf{Z}$. In this section, we will study sufficient conditions for two probability distributions to be *indistinguishable*, which will yield sufficient conditions for the weak recovery of $\mathbf{x}$ to be impossible. To formulate this theory, we use a greater level of generality, beyond spiked matrix models.

Definition 6.1 (*Statistical distinguishability*)

We consider $(\mathbb{P}_d, \mathbb{Q}_d)$ two sequences of probability distributions on a common measurable space $(\mathcal{S}_d, \mathcal{F}_d)$, where $\mathcal{S}_d \subseteq \mathbb{R}^d$. We say that a measurable function $f : \mathcal{S}_d \to \{0, 1\}$ *strongly distinguishes* $\mathbb{P}_d$ and $\mathbb{Q}_d$ if:

$$\begin{cases} \mathbb{P}_d[\{f(\mathbf{Y}) = 1\}] &= 1 - o_{d\to\infty}(1), \\ \mathbb{Q}_d[\{f(\mathbf{Y}) = 0\}] &= 1 - o_{d\to\infty}(1). \end{cases} \tag{34}$$

If no such f exists, we say that $\mathbb{P}_d$ and $\mathbb{Q}_d$ are *statistically indistinguishable*.

We think of f as a test function: given a sample $\mathbf{Y}$, it tries to categorize it either as coming from the distribution $\mathbb{P}_d$ (then it sets $f(\mathbf{Y}) = 1$) or from the distribution $\mathbb{Q}_d$ (then $f(\mathbf{Y}) = 0$). In our motivating example, $\mathbb{Q}_d$ is the law of the pure noise, while $\mathbb{P}_d$ is the law of the signal + noise matrix:

$$\mathbb{Q}_d : \mathbf{Y} = \mathbf{Z},$$
$$\mathbb{P}_d : \mathbf{Y} = \mathbf{Z} + \theta \mathbf{x}\mathbf{x}^\top.$$

We now introduce another definition on sequences of probability distributions, which easily implies statistical indistinguishability.

Definition 6.2 (*Contiguity*)

We say that $\mathbb{P}_d$ is contiguous to $\mathbb{Q}_d$ (denoted $\mathbb{P}_d \lhd \mathbb{Q}_d$) if, for all sequences of events $A_d \in \mathcal{F}_d$:

$$\lim_{d\to\infty} \mathbb{Q}_d(A_d) = 0 \Rightarrow \lim_{d\to\infty} \mathbb{P}_d(A_d) = 0.$$

Informally, $\mathbb{P}_d \lhd \mathbb{Q}_d$ means that all events which happen under $\mathbb{Q}_d$ with high probability must also have high probability under $\mathbb{P}_d$.

Theorem 6.1

Assume that $\mathbb{P}_d \lhd \mathbb{Q}_d$ or $\mathbb{Q}_d \lhd \mathbb{P}_d$. Then $\mathbb{P}_d$ and $\mathbb{Q}_d$ are statistically indistinguishable.

Proof of Theorem 6.1 – We assume that $\mathbb{P}_d \lhd \mathbb{Q}_d$, the other case will follow by symmetry. Assume that there exists a test f strongly distinguishing $\mathbb{P}_d$ and $\mathbb{Q}_d$. Then, if $A_d \coloneqq \{f(\mathbf{Y}) = 1\}$, by definition $\mathbb{Q}(A_d) \to 0$ as $d \to \infty$. Since $\mathbb{P}_d \lhd \mathbb{Q}_d$, we must have $\mathbb{P}(A_d)$, which contradicts the notion of strong distinguishability. □

An extremely useful sufficient condition for contiguity arises from the study of the *likelihood ratio*. It is defined as[6]

$$L_d(\mathbf{Y}) \coloneqq \frac{\mathrm{d}\mathbb{P}_d}{\mathrm{d}\mathbb{Q}_d}(\mathbf{Y}) \tag{35}$$

The likelihood ratio has many interesting properties. Notably, L_d can be used (via a simple thresholding procedure) to build an optimal statistical test to distinguish between $\mathbb{P}_d$ and $\mathbb{Q}_d$: this result is usually called the *Neyman-Pearson* lemma [NP33], see [KWB19] for more details. Here, we will focus on another property of the likelihood ratio which is related to the contiguity of distributions.

Lemma 6.2 (*Second moment method for continguity*)

Let $\|L_d\|^2_{\mathbb{Q}} \coloneqq \mathbb{E}_{\mathbf{Y}\sim\mathbb{Q}_d}[L_d(\mathbf{Y})^2]$. If $\|L_d\|^2_{\mathbb{Q}} = \mathcal{O}_{d\to\infty}(1)$, then $\mathbb{P}_d \lhd \mathbb{Q}_d$.

Lemma 6.2 is an extremely strong tool to prove contiguity, and thus statistical indistinguishability: it suffices to compute $\|L_d\|^2_{\mathbb{Q}}$, and to study its large-d behavior! Before proving Lemma 6.2 let us make a couple of remarks:

(*i*) Since we will use Lemma 6.2 to prove that $\mathbb{Q}_d$ and $\mathbb{P}_d$ are indistinguishable, another sufficient condition is that $\mathrm{d}\mathbb{Q}_d/\mathrm{d}\mathbb{P}_d$ has bounded $L^2(\mathbb{P}_d)$-norm. However, we usually take the convention that $\mathbb{Q}_d$ is the "noise" distribution, which will typically be simple (for instance with independent elements): for this reason, computing the likelihood ratio of eq. (35) is usually much easier than computing $\mathrm{d}\mathbb{Q}_d/\mathrm{d}\mathbb{P}_d$.

[6] We assume that $\mathbb{P}_d$ is absolutely continuous with respect to $\mathbb{Q}_d$, which will always be the case in our applications.

(*ii*) $\|L_d\|^2_{\mathbb{Q}} = \mathcal{O}_{d\to\infty}(1)$ is not a necessary condition for contiguity: there might be events which have probability $o_d(1)$ under both $\mathbb{P}_d$ and $\mathbb{Q}_d$ (so they do not impact contiguity), however if on these events the ratio $\mathrm{d}\mathbb{P}_d/\mathrm{d}\mathbb{Q}_d$ is extremely large, it could cause the second moment $\|L_d\|^2_{\mathbb{Q}}$ to diverge. One idea can then be to condition away from these rare "bad" events, and study the likelihood ratio of the conditioned distributions: see [Per+18] (Lemma 2.4) for details.

Proof of Lemma 6.2 – The proof is surprisingly simple. Let A_d be a sequence of events such that $\mathbb{Q}_d(A_d) \to 0$ as $d \to \infty$. Then

$$\begin{aligned}\mathbb{P}_d(A_d) &= \mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_d}[\mathbb{1}_{A_d}(\mathbf{Y})],\\ &= \mathbb{E}_{\mathbf{Y}\sim\mathbb{Q}_d}\left[\frac{\mathrm{d}\mathbb{P}_d}{\mathrm{d}\mathbb{Q}_d}(\mathbf{Y})\mathbb{1}_{A_d}(\mathbf{Y})\right],\\ &\overset{(a)}{\leq} (\mathbb{E}_{\mathbf{Y}\sim\mathbb{Q}_d}[L_d(\mathbf{Y})^2])^{1/2}(\mathbb{Q}_d(A_d))^{1/2} \overset{(b)}{=} o_d(1).\end{aligned}$$

We just used the Cauchy-Schwarz inequality in (a), and the assumption that $\|L_d\|^2_{\mathbb{Q}}$ is bounded in (b). □

6.3 Application in spiked matrix models

We will now leverage the theoretical framework we developed in Section 6.2 to tackle the spiked matrix model. Recall that we defined the two distributions

$$\begin{cases} \mathbb{Q}_d \,:\, \mathbf{Y} = \mathbf{Z}, \\ \mathbb{P}_d \,:\, \mathbf{Y} = \mathbf{Z} + \theta\mathbf{x}\mathbf{x}^\top, \end{cases} \tag{36}$$

in which $\mathbf{Z} \sim \mathrm{GOE}(d)$ (see eq. (21)), $\theta > 0$ is the signal-to-noise ratio, and $\mathbf{x} \in \mathbb{R}^d$ is the signal. Here we assume that $\mathbf{x} \sim \mathcal{X}_d$, a probability distribution on $\mathbb{R}^d$. The properties of $\mathcal{X}_d$ will quantify the "structure" present in $\mathbf{x}$: our motivating example is $\mathcal{X}_d = \mathrm{Unif}(\{\pm 1/\sqrt{d}\}^d)$. Using Lemma 6.2, we wish to study when one can not improve over the PCA estimator. We thus compute the norm of the likelihood ration.

Theorem 6.3 (*Norm of the likelihood ratio in the spiked matrix model*)

For the distributions of eq. (36):

$$\|L_d\|^2_{\mathbb{Q}} = \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mathcal{X}_d} \exp\left\{\frac{d\theta^2}{2}\langle \mathbf{x}, \mathbf{x}'\rangle^2\right\}.$$

Proof of theorem 6.3 – $\mathbb{P}_d$ and $\mathbb{Q}_d$ are both absolutely continuous with respect to the Lebesgue measure. We have

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbf{Y}}(\mathrm{d}\mathbf{Y}) = \frac{1}{N_d}\prod_{i<j} e^{-d\frac{Y_{ij}^2}{2}} \prod_{i=1}^d e^{-d\frac{Y_{ii}^2}{4}} = \frac{e^{-\frac{d}{4}\mathrm{Tr}[\mathbf{Y}^2]}}{N_d}.$$

For any smooth function f, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_d}[f(\mathbf{Y})] &= \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d}\mathbb{E}_{\mathbf{Z}\sim\mathrm{GOE}(d)}[f(\mathbf{Z}+\theta\mathbf{x}\mathbf{x}^\top)],\\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d}\frac{1}{N_d}\int \mathrm{d}\mathbf{Z} e^{-\frac{d}{4}\mathrm{Tr}[\mathbf{Z}^2]} f(\mathbf{Z}+\theta\mathbf{x}\mathbf{x}^\top),\end{aligned}$$

$$= \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d} \frac{1}{N_d} \int \mathrm{d}\mathbf{Y} e^{-\frac{d}{4}\mathrm{Tr}[(\mathbf{Y}-\theta\mathbf{x}\mathbf{x}^\top)^2]} f(\mathbf{Y}).$$

So:

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbf{Y}}(\mathbf{Y}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d} \frac{1}{N_d} e^{-\frac{d}{4}\mathrm{Tr}[(\mathbf{Y}-\theta\mathbf{x}\mathbf{x}^\top)^2]}.$$

We reach that

$$L_d(\mathbf{Y}) \coloneqq \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\mathbf{Y}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d}\left[e^{-\frac{d\theta^2}{4}\|\mathbf{x}\|^4 + \frac{d\theta}{2}\mathbf{x}^\top\mathbf{Y}\mathbf{x}}\right].$$

Using Fubini's theorem:

$$\begin{aligned}\|L_d\|_{\mathbb{Q}}^2 &= \mathbb{E}_{\mathbf{Y}\sim\mathrm{GOE}(d)}\left[\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d}\left[e^{-\frac{d\theta^2}{4}\|\mathbf{x}\|^4 + \frac{d\theta}{2}\mathbf{x}^\top\mathbf{Y}\mathbf{x}}\right]\right)^2\right], \\ &= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mathcal{X}_d}\left[e^{-\frac{d\theta^2}{4}(\|\mathbf{x}\|^4+\|\mathbf{x}'\|^4)}\mathbb{E}_{\mathbf{Y}\sim\mathrm{GOE}(d)}\left[e^{\frac{d\theta}{2}\mathrm{Tr}[\mathbf{Y}(\mathbf{x}\mathbf{x}^\top+\mathbf{x}'(\mathbf{x}')^\top)]}\right]\right].\end{aligned}$$

By Gaussian integration, it is easy to show that for any symmetric matrix $\mathbf{M}$:

$$\mathbb{E}_{\mathbf{Y}\sim\mathrm{GOE}(d)} \exp\{d\mathrm{Tr}[\mathbf{M}\mathbf{Y}]\} = \exp\{d\mathrm{Tr}[\mathbf{M}^2]\}.$$

Applying this to our computation, we get:

$$\begin{aligned}\|L_d\|_{\mathbb{Q}}^2 &= \mathbb{E}_{\mathbf{Y}\sim\mathrm{GOE}(d)}\left[\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{X}_d}\left[e^{-\frac{d\theta^2}{4}\|\mathbf{x}\|^4 + \frac{d\theta}{2}\mathbf{x}^\top\mathbf{Y}\mathbf{x}}\right]\right)^2\right], \\ &= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mathcal{X}_d}\left[e^{-\frac{d\theta^2}{4}(\|\mathbf{x}\|^4+\|\mathbf{x}'\|^4)+\frac{d\theta^2}{4}\mathrm{Tr}[(\mathbf{x}\mathbf{x}^\top+\mathbf{x}'\mathbf{x}'^\top)^2]}\right], \\ &= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mathcal{X}_d}\left[e^{\frac{d\theta^2}{2}\langle\mathbf{x},\mathbf{x}'\rangle^2}\right],\end{aligned}$$

which ends the proof. □

Theorem 6.3 expresses $\|L_d\|_{\mathbb{Q}}^2$ as a function of the "overlap" $\langle\mathbf{x},\mathbf{x}'\rangle$ of two independent samples under $\mathcal{X}_d$. We now look for conditions on $\mathcal{X}_d$ to show that $\|L_d\|^2 = \mathcal{O}(1)$, to apply Lemma 6.2.

Sub-Gaussian random vectors – Recall the definition of a sub-Gaussian $\mathrm{SG}(\sigma)$ random variable in Definition 4.1. By analogy, we sat that a vector $\mathbf{x}\in\mathbb{R}^d$ has a sub-Gaussian $\mathrm{SG}(\sigma)$ distribution if, for any $\mathbf{v}\in\mathcal{S}^{d-1}$, the projection $\langle\mathbf{v},\mathbf{x}\rangle$ is $\mathrm{SG}(\sigma)$:

$$\sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \mathbb{E}[\exp\{\lambda\langle\mathbf{v},\mathbf{x}\rangle\}] \le \exp\{\lambda^2\sigma^2/2\}, \ (\forall\lambda\in\mathbb{R}). \tag{37}$$

Examples of sub-Gaussian distributions include $\mathcal{N}(0,\mathrm{I}_d)\in\mathrm{SG}(1)$, and $\mathrm{Unif}(\{\pm1/\sqrt{d}\}^d)$ is $\mathrm{SG}(1/\sqrt{d})$ as well. More generally, any i.i.d. random vector with $\mathrm{SG}(\sigma)$ coordinates is also $\mathrm{SG}(\sigma)$. But the notion of sub-Gaussian random vectors allow for non-iid distributions: for instance the uniform distribution on the sphere $\mathrm{Unif}(\mathcal{S}^{d-1})$ is $\mathrm{SG}(C/\sqrt{d})$ for some $C>0$. We refer to [Ver18] for more details on sub-Gaussian random vectors, and proofs of these properties. The main result of this section is the following theorem, due to [Per+18].

Theorem 6.4 (*[Per+18]*)

Assume that $\mathbf{x}\sim\mathcal{X}_d$, and that $\mathcal{X}_d$ is $(\sigma/\sqrt{d})$-subGaussian. If $\theta < 1/\sigma$, the two

distributions of eq. (36) satisfy $\mathbb{P}_d \lhd \mathbb{Q}_d$, and are thus statistically indistinguishable.

Notice that Theorem 6.4 only gives a sufficient condition for contiguity: for instance, when $\mathcal{X}_d = \mathrm{Unif}[\mathcal{S}^{d-1}]$, it can be shown by refining the proof of this theorem that $\mathbb{P}_d \lhd \mathbb{Q}_d$ for all $\theta < 1$, see Corollary 3.14 of [Per+18].

Proof of Theorem 6.4 – We prove Theorem 6.4 when $\mathcal{X}_d = \mathrm{Unif}(\{\pm 1/\sqrt{d}\})$ (in this case $\sigma = 1$). The general case can be found in [Per+18], Proposition 3.9.

Notice that here $\langle \mathbf{x}, \mathbf{x}' \rangle = \sum_{i=1}^d x_i x_i'$ is the sum of d i.i.d. random variables, which are bounded as $|x_i x_i'| \le 1/d$. Thus, by Hoeffding's inequality (Theorem 2.2):

$$\mathbb{P}(|\langle \mathbf{x}, \mathbf{x}' \rangle| \ge t) \le 2\exp\{-dt^2/2\}. \tag{38}$$

We now compute:

$$\begin{aligned}
\|L_d\|_{\mathbb{Q}}^2 &\overset{(a)}{=} \mathbb{E}_{\mathbf{x},\mathbf{x}' \sim \mathcal{X}_d}\left[e^{\frac{d\theta^2}{2}\langle \mathbf{x},\mathbf{x}'\rangle^2}\right], \\
&= \int_0^\infty \mathbb{P}\left[e^{\frac{d\theta^2}{2}\langle \mathbf{x},\mathbf{x}'\rangle^2} \ge u\right] \mathrm{d}u, \\
&\overset{(b)}{=} 1 + \int_1^\infty \mathbb{P}\left[e^{\frac{d\theta^2}{2}\langle \mathbf{x},\mathbf{x}'\rangle^2} \ge u\right] \mathrm{d}u, \\
&= 1 + \int_1^\infty \mathbb{P}\left[|\langle \mathbf{x}, \mathbf{x}'\rangle| \ge \frac{1}{\theta}\sqrt{\frac{2\log u}{d}}\right] \mathrm{d}u, \\
&\overset{(c)}{\le} 1 + 2\int_1^\infty \exp\left\{-\frac{\log u}{\theta^2}\right\} \mathrm{d}u, \\
&= 1 + 2\int_1^\infty u^{-1/\theta^2} \mathrm{d}u.
\end{aligned}$$

In (a) we used theorem 6.3, in (b) that $\exp(u) \ge 1$ for $u \ge 0$, and in (c) eq. (38). We have shown that for all $d \ge 1$,

$$\|L_d\|_{\mathbb{Q}}^2 \le 1 + 2\int_1^\infty u^{-1/\theta^2} \mathrm{d}u,$$

and the right-hand side of this equality is clearly $< \infty$ as long as $\theta < 1$, which ends the proof. □

Consequence – So, if $\mathcal{X}_d$ is $1/\sqrt{d}$-sub-Gaussian (e.g. if $\mathcal{X}_d = \mathrm{Unif}(\{\pm 1/\sqrt{d}\}^d)$, but also if $\mathcal{X}_d = \mathrm{Unif}[\mathcal{S}^{d-1}]$ as we mentioned):

- For $\theta < 1$, *no procedure* can recover $\mathbf{x}$ from the observation of $\mathbf{Y} \sim \mathbb{P}_d$.
- For $\theta > 1$, the PCA estimator recovers partially $\mathbf{x}$, and we can characterize the limiting overlap $\lim_{d\to\infty} \langle \hat{\mathbf{x}}_{\mathrm{PCA}}, \mathbf{x}\rangle^2 > 0$ via the BBP transition, see Section 5.

For such signal distributions, we can say that the spectral PCA algorithm achieves the optimal *weak recovery* (i.e. partial recovery) θ-threshold. This answers in the negative our original question **(Q)**! However, there are some distributions $\mathcal{X}_d$ for which PCA is provably not optimal: we will investigate an example in Section 6.4. Before that, let us make two final remarks.

Remark I: contiguity of spectra – It turns out that $\theta = 1$ is not only the weak recovery threshold for PCA, but for any estimator based on the spectrum of the matrix $\mathbf{Y}$, for very general classes of distributions $\mathcal{X}_d$. Indeed, the following theorem, proven in [Per+18], shows the *contiguity of the spectra* under $\mathbb{P}_d$ and $\mathbb{Q}_d$.

Theorem 6.5 (*Contiguity of spectral distributions [Per+18]*)

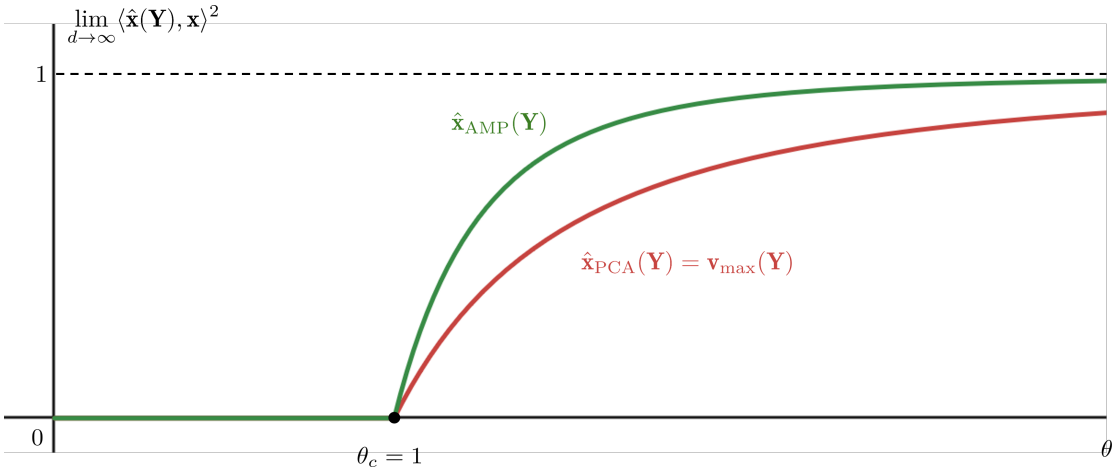
Let $\mathcal{X}_d$ be such that, if $\mathbf{x} \sim \mathcal{X}_d$, then $\|\mathbf{x}\|_2 \to 1$ in probability as $d \to \infty$.

- $\widetilde{\mathbb{Q}}_d$ be the joint law of the eigenvalues of $\mathbf{Y} \sim \mathbb{Q}_d$ of eq. (36).
- $\widetilde{\mathbb{P}}_d$ be the joint law of the eigenvalues of $\mathbf{Y} \sim \mathbb{P}_d$ of eq. (36).

If $\theta < 1$, then $\widetilde{\mathbb{P}}_d \lhd \widetilde{\mathbb{Q}}_d$.

The condition $\|\mathbf{x}\|_2 \to 1$ is very weak, and essentially amounts to say that $\theta > 0$ is the well-defined "signal-to-noise" ratio of the problem.

Remark II: Beyond weak recovery – However, even if $\mathcal{X}_d$ is $(1/\sqrt{d})$-sub-Gaussian, Theorem 6.4 does not inform us about what is the optimal performance achievable for $\theta > 1$. For instance, for $\mathbf{x} \sim \mathrm{Unif}(\{1/\sqrt{d}\})$, one might still expect to be able to leverage the structure of $\mathbf{x}$ to enhance the limiting correlation with respect to the PCA estimator. This turns out to true, and the optimal known algorithms for this belong to the class of *approximate message-passing* (AMP), and have connections with statistical physics [LKZ15b]. The (informal) picture looks as follows.



The red curve is the same one as in Fig. 2! The green curve can be analytically computed by analyzing the high-dimensional limit of AMP algorithms: this was first done using non-rigorous methods of statistical physics [LKZ15b], and is now rigorously established in many settings, including the spiked matrix model, see e.g. [GB23].

6.4 Spiked matrix model with sparse signals

What happens to our results if we assume that the signal vector $\mathbf{x}$ has a lot of structure that can be exploited by algorithms? We discuss here some important results related to this question, which allow to give a positive answer to **(Q)** for some distributions $\mathcal{X}_d$.

A particularly important type of structure (encountered as well in the lecture [BSS23]) is *sparsity*. We consider the distribution $\mathcal{X}_d := \pi_\rho^{\otimes d}$ (i.e. $x_i \overset{\text{i.i.d.}}{\sim} \pi_\rho$), where $\rho \in (0, 1]$ is the sparsity parameter:

$$\pi_\rho := (1-\rho)\delta_0 + \frac{\rho}{2}\left[\delta_{(1/\sqrt{d\rho})} + \delta_{(-1/\sqrt{d\rho})}\right]. \tag{39}$$

Typically, the vector $\mathbf{x} \sim \mathcal{X}_d$ will have a fraction ρ of non-zero coordinates. The following results were shown in the series of works [LKZ15b; LKZ15a; Dia+16; KXZ16; LM17b].

- Notice that $\mathbb{E}_{\pi_\rho}[x^2] = 1/d$, and more generally one can easily show that $\|\mathbf{x}\|_2 \to 1$ in probability as $d \to \infty$. By Theorem 6.5, this implies that PCA will not succeed in recovering $\mathbf{x}$ (even partially) for any $\theta < 1$. The PCA performance curve is shown in red in Fig. 3.
- For large enough ρ, the optimal estimator fails to weakly recover the signal for $\theta < 1$. However, for small ρ (i.e. very sparse vectors), there is a critical value $\rho^\star \simeq 0.09$ (for which an explicit formula exists) such that, for $\rho < \rho^\star$, partial recovery of $\mathbf{x}$ is possible for any $\theta > \theta_c(\rho)$, and that $\theta_c(\rho) < 1$! The optimal estimator in this case, denoted $\hat{\mathbf{x}}_{\text{opt.}}(Y)$ and shown in green in Fig. 3, can be naively computed in time $\exp\{\Theta(d)\}$.
- Finally, the best-known *polynomial-time* algorithms for this problem are still the AMP algorithms we mentioned above. However, they fail to weakly recovery the signal $\mathbf{x}$ for any $\theta < 1$, even for $\theta > \theta_c(\rho)$! On the other hand, for $\theta > 1$, their performance matches the one of $\hat{\mathbf{x}}_{\text{opt.}}(\mathbf{Y})$: it is shown in purple in Fig. 3.

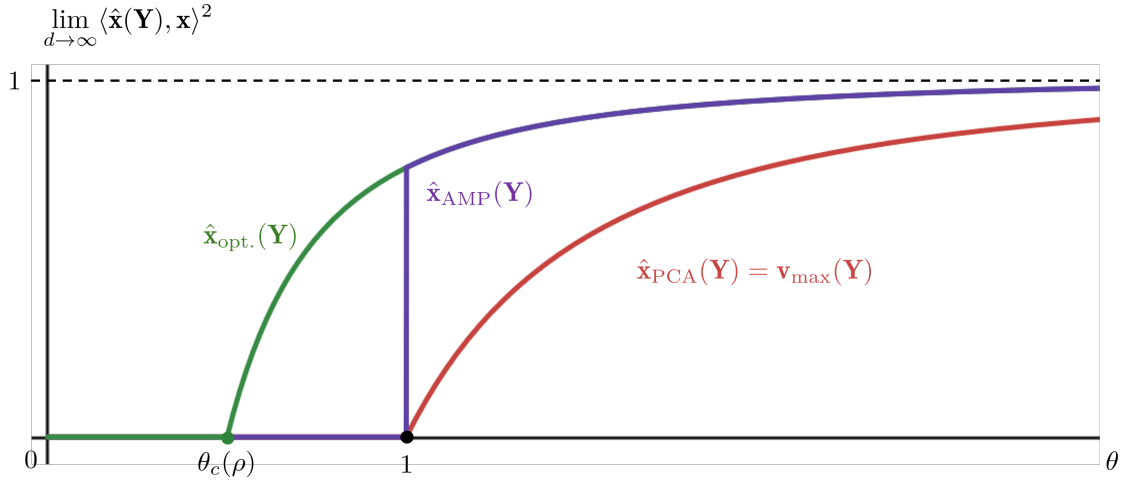


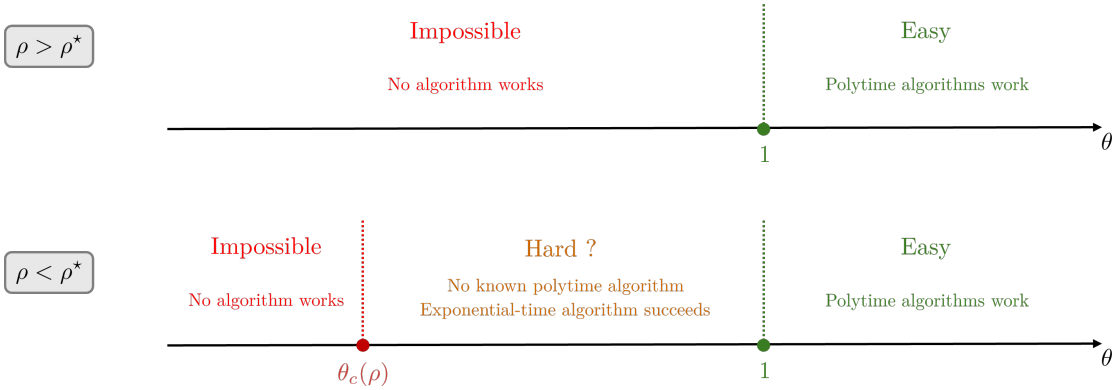
Figure 3: Schematic view of the different limiting performances in the spiked matrix model with a sparse signal, for $\rho < \rho^\star$.

The picture above leads to conjecture that sparse PCA presents a *computational-to-statistical gap* (also called *hard phase*).

Conjecture 6.6 (*Hardness of sparse PCA, informal*)

For any $\rho \in (0, 1]$ and $\theta < 1$, no *polynomial-time* algorithm can achieve (with high probability) positive correlation with $\mathbf{x}$ as $d \to \infty$.

The informal picture for weak recovery (i.e. achieving $\lim_{d\to\infty} \langle \hat{\mathbf{x}}(\mathbf{Y}), \mathbf{x}\rangle^2 > 0$) is:



Towards Conjecture 6.6 – Unfortunately we lack a theory similar to the one of NP-hardness for random problems that we want to solve with high probability. For this reason, the current progress on such hardness conjectures usually focuses on showing the failure of large classes of algorithms, which are strong indications of the hardness of the problem. This includes the AMP algorithm mentioned above, but also low-degree polynomials, or algorithms based on the sum-of-squares hierarchy for instance: take a look at the very nice surveys [KWB19] (on low-degree polynomials) and [GMZ22] (more focused on AMP algorithms and the SoS hierarchy) to learn the ideas behind these different approaches!

7 Finding cliques in random graphs

In this chapter we will consider some questions related to finding cliques in random graph ensembles, and in particular introduce one of the most famous problems for which it is conjectured that a computational-to-statistical gap (similarly to what we saw for sparse PCA in Section 6) exists.

Recall that for an unweighted and undirected graph $G = (V, E)$, a *clique* $K \subseteq V$ is a set of edges such that all nodes in K are connected: $\forall i, j \in K,\ i \neq j \Rightarrow (i, j) \in E$.

7.1 Cliques in unstructured random graphs: sharp asymptotics

Here we consider a *random* graph from the Erdős–Rényi distribution $G \sim \mathrm{ER}(n, p)$: i.e. G is a graph on n nodes, and for each pair (i, j) (with $i < j$) we connect i and j with probability p. For simplicity we will consider $p = 1/2$ here, but all can be generalized to any fixed $p \in (0, 1)$, see Challenge 7.3. We ask the following question, for large $n \gg 1$:

What is the size of the largest clique in G?

This question was solved more than 50 years ago (see e.g. [Mat70; GM75; BE76]), and its answer is given in the following theorem.

Theorem 7.1 (*Larges clique in a random graph*)

Let $G \sim \mathrm{ER}(n, 1/2)$, and denote X_n the size of the largest clique in G. Then

$$\frac{X_n}{2\log_2 n} \xrightarrow[n\to\infty]{\text{(p)}} 1.$$

Here, (p) denotes the limit in probability. Let us first prove Theorem 7.1. Later, in Section 7.2 we will highlight an important open problem related to this result.

Sketch of the proof – The proof of Theorem 7.1 is based on the *first and second moment methods*. These are classical tools in high-dimensional probability (in particular regarding random constraint satisfaction problems). The first thing is to notice that $X_n = \max\{k \in [n] \,:\, Z_k \geq 1\}$, in which Z_k is the number of cliques in G containing exactly k vertices. The first moment method then amounts to study the expectation of Z_k, and we will show[7].

Lemma 7.2

For $k^\star(n) = 2\log_2 n$, we have for any $\varepsilon > 0$:

- **(First moment method):** $\lim_{n\to\infty} \mathbb{E}[Z_{(1+\varepsilon)k^\star}] = 0$, and so $\mathbb{P}[Z_{(1+\varepsilon)k^\star} > 0] \to 0$ as $n \to \infty$.
- $\lim_{n\to\infty} \mathbb{E}[Z_{(1-\varepsilon)k^\star}] = \infty$.

We prove Lemma 7.2 in Section 7.1.1. The reason why the first moment method is tractable is the *linearity of expectation*, a seemingly trivial result which has very non-trivial consequences: while $Z_k = \sum_{S\subseteq[n]} \mathbb{1}\{|S| = k \text{ and } S \text{ is a clique}\}$ is the sum of many correlated random variables, its expectation is still the sum of expectations, and can therefore be computed quite easily! Notice that the second statement in the first

[7] For simplicity we write $Z_{(1-\varepsilon)k}$ instead of $Z_{\lfloor(1-\varepsilon)k\rfloor}$ (and similarly for other quantities): this will allow to lighten the presentations, and will make no difference in the $n \to \infty$ limit.

moment method is a direct consequence of Markov's inequality (since Z_k is integer-valued), and it implies that $\mathbb{P}[Z_{(1+\varepsilon)k^\star} > 0] = \mathbb{P}[X_n \geq (1+\varepsilon)k^\star] \to 0$ for any $\varepsilon > 0$. This already gives us the upper bound for Theorem 7.1!

The corresponding lower-bound follows from the *second moment method*. The idea is to now study the variance of Z_k, and to use the following elementary inequality.

Lemma 7.3 (*Paley-Zygmund inequality*)

Let Y be a non-negative random variable. Then, for any $\theta \in (0,1)$:

$$\mathbb{P}[Y \geq \theta \mathbb{E}Y] \geq (1-\theta)^2 \frac{(\mathbb{E}Y)^2}{\mathbb{E}[Y^2]}.$$

(Second moment method, general) If Y is not identically zero:

$$\mathbb{P}[Y > 0] \geq \frac{(\mathbb{E}Y)^2}{\mathbb{E}[Y^2]} = 1 - \frac{\mathrm{Var}(Y)}{\mathbb{E}[Y^2]}.$$

Challenge 7.1. *Prove Lemma 7.3 using the Cauchy-Schwarz inequality.*

Essentially, Lemma 7.3 states that if $\mathbb{E}(Y) > 0$ and Y is concentrated close to its average (which is quantified as $\mathrm{Var}(Y) \ll \mathbb{E}[Y^2]$), then Y is positive with large probability. This should be very reminiscent of Chebyshev's inequality. Our study of $\mathbb{E}[Z_k^2]$ is done in Section 7.1.2, where we prove:

Lemma 7.4

For $k^\star(n) = 2\log_2 n$, we have for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \frac{\mathbb{E}[Z^2_{(1-\varepsilon)k^\star}]}{(\mathbb{E}Z_{(1-\varepsilon)k^\star})^2} = 1.$$

Combining Lemma 7.4 with the second moment method in Lemma 7.3, we get that $\mathbb{P}[Z_{(1-\varepsilon)k^\star} > 0] = \mathbb{P}[X_n \geq (1-\varepsilon)k^\star] \to 1$, and we obtain the matching lower bound to conclude our proof of Theorem 7.1!

7.1.1 First moment method: proof of Lemma 7.2

As said above, we use linearity of expectation:

$$\begin{aligned}
\mathbb{E}[Z_k] &= \mathbb{E}\left[\sum_{S\subseteq[n]} \mathbb{1}\{|S| = k \text{ and } S \text{ is a clique}\}\right], \\
&= \sum_{S\subseteq[n]} \mathbb{P}\{|S| = k \text{ and } S \text{ is a clique}\}, \\
&= \#\{S \subseteq [n] \,:\, |S| = k\} \times \mathbb{P}[\{1,\cdots,k\} \text{ is a clique}],
\end{aligned}$$

where the last equality follows from the symmetry of the problem. The probability that $\{1,\cdots,k\}$ is a clique is easy to compute, and we get:

$$\mathbb{E}[Z_k] = \binom{n}{k} \left(\frac{1}{2}\right)^{\binom{k}{2}}. \tag{40}$$

Upper bound – From the inequality $\binom{n}{k} \leq n^k$, we have

$$\mathbb{E}[Z_k] \leq 2^{k\left[\log_2 n - \frac{k-1}{2}\right]}.$$

Therefore, if we assume that $k \geq 2(1+\varepsilon)\log_2 n$ (for a fixed $\varepsilon > 0$), we obtain that $\mathbb{E}[Z_k] \to 0$.

Lower bound – It is also a classical inequality (prove it!) that $\binom{n}{k} \geq (n/k)^k$ for any $k \in \{1, \cdots, n\}$. From this and eq. (40) we get:

$$\mathbb{E}[Z_k] \geq 2^{k\left[\log_2 n - \frac{k-1}{2} - \log_2 k\right]}.$$

Again, from this expression, it is clear that if $k \leq 2(1-\varepsilon)\log_2 n$ (for a fixed $\varepsilon > 0$), we have that $\mathbb{E}[Z_k] \gtrsim 2^{(\varepsilon k \log_2 n)/2} \to \infty$, which ends the proof. □

7.1.2 Second moment method: proof of Lemma 7.4

The second moment computation involves more complicated combinatorics than the first moment, but the idea is similar. A good reference for this computation, as well as other applications of the second moment method, is [AS16]. For $S \subseteq [n]$, we define $I_S \coloneqq \mathbb{1}\{S \text{ is a clique}\}$. Then

$$\begin{aligned}\mathbb{E}[Z_k^2] &= \mathbb{E}\left[\left(\sum_{\substack{S\subseteq[n]\\|S|=k}} I_S\right)^2\right],\\ &= \sum_{\substack{S,T\subseteq[n]\\|S|=k,|T|=k}} \mathbb{E}[I_S I_T].\end{aligned}$$

Notice that $\mathbb{E}[I_S I_T] = \mathbb{P}[S \text{ and } T \text{ are cliques}]$, and for the Erdős–Rényi model, $\mathbb{E}[I_S I_T] = \mathbb{E}[I_S]\mathbb{E}[I_T]$ if $|S \cap T| \leq 1$. Moreover, recall that $\mathbb{E}[Z_k] = \sum_{|S|=k} \mathbb{E}[I_S]$. So we have

$$\frac{\mathbb{E}[Z_k^2]}{\mathbb{E}[Z_k]^2} = \frac{\sum_{|S|=k,|T|=k} \mathbb{1}\{|S\cap T| \leq 1\}\mathbb{E}[I_S]\mathbb{E}[I_T]}{\sum_{|S|=k,|T|=k} \mathbb{E}[I_S]\mathbb{E}[I_T]} + \frac{\sum_{|S|=k,|T|=k} \mathbb{1}\{|S\cap T| \geq 2\}\mathbb{E}[I_S I_T]}{\sum_{|S|=k,|T|=k} \mathbb{E}[I_S]\mathbb{E}[I_T]}.$$

And so

$$\frac{\mathbb{E}[Z_k^2]}{\mathbb{E}[Z_k]^2} - 1 = \frac{\sum_{|S|=k,|T|=k} \mathbb{1}\{|S\cap T| \geq 2\}(\mathbb{E}[I_S I_T] - \mathbb{E}[I_S]\mathbb{E}[I_T])}{\sum_{|S|=k,|T|=k} \mathbb{E}[I_S]\mathbb{E}[I_T]}. \tag{41}$$

Since $\mathbb{E}[Z_k^2] \geq \mathbb{E}[Z_k]^2$, we even have

$$0 \leq \frac{\mathbb{E}[Z_k^2]}{\mathbb{E}[Z_k]^2} - 1 \leq \frac{\sum_{|S|=k,|T|=k} \mathbb{1}\{|S\cap T| \geq 2\}\mathbb{E}[I_S I_T]}{\sum_{|S|=k,|T|=k} \mathbb{E}[I_S]\mathbb{E}[I_T]}. \tag{42}$$

To simplify notations, we denote $S \sim T$ if $|S| = |T| = k$, and $|S \cap T| \geq 2$. We now show that the right-hand side of eq. (42) goes to 0 as $n \to \infty$, for $k = (1-\varepsilon)k^\star$. The denominator is simply $\mathbb{E}[Z_k]^2$, which we have computed in eq. (40). The numerator can be computed using the conditional expectation, and the symmetry of the problem:

$$\begin{aligned}\sum_{S\sim T} \mathbb{E}[I_S I_T] &= \sum_{S\sim T} \mathbb{P}[I_S = 1 \wedge I_T = 1],\\ &= \sum_{S\sim T} \mathbb{P}[I_S = 1]\mathbb{P}[I_T = 1 | I_S = 1],\end{aligned}$$

$$= \sum_{|S|=k} \mathbb{P}[I_S = 1] \left(\sum_{T:T\sim S} \mathbb{P}[I_T = 1 | I_S = 1] \right).$$

By symmetry, the term $\sum_{T:T\sim S} \mathbb{P}[I_T = 1|I_S = 1]$ is independent of the choice of S: if we let $S_0 \coloneqq \{1, \cdots, k\}$, we therefore have:

$$\begin{aligned}\sum_{S\sim T} \mathbb{E}[I_S I_T] &= \left(\sum_{|S|=k} \mathbb{P}[I_S = 1] \right) \left(\sum_{T:T\sim S_0} \mathbb{P}[I_T = 1|I_{S_0} = 1] \right), \\ &= \mathbb{E}[Z_k] \left(\sum_{T:T\sim S_0} \mathbb{P}[I_T = 1|I_{S_0} = 1] \right).\end{aligned}$$

We have thus simplified eq. (42) as:

$$0 \leq \frac{\mathbb{E}[Z_k^2]}{\mathbb{E}[Z_k]^2} - 1 \leq \frac{\sum_{T:T\sim S_0} \mathbb{P}[I_T = 1|I_{S_0} = 1]}{\sum_{|S|=k} \mathbb{E}[I_S]}. \tag{43}$$

We must now do the combinatorics. Denoting $i = |T \cap S_0|$, we have

$$\sum_{T:T\sim S_0} \mathbb{P}[I_T = 1|I_{S_0} = 1] = \sum_{i=2}^{k} \underbrace{\binom{k}{i}\binom{n-k}{k-i}}_{\text{choice of the vertices}} \times \underbrace{\left(\frac{1}{2}\right)^{\left[\binom{k}{2}-\binom{i}{2}\right]}}_{T \text{ is a clique}}.$$

Therefore:

$$\begin{aligned}\frac{\sum_{T:T\sim S_0} \mathbb{P}[I_T = 1|I_{S_0} = 1]}{\sum_{|S|=k} \mathbb{E}[I_S]} &= \frac{2^{\binom{k}{2}}}{\binom{n}{k}} \sum_{i=2}^{k} \binom{k}{i}\binom{n-k}{k-i} 2^{-\left[\binom{k}{2}-\binom{i}{2}\right]}, \\ &= \sum_{i=2}^{k} \frac{\binom{k}{i}\binom{n-k}{k-i} 2^{\binom{i}{2}}}{\binom{n}{k}}, \\ &\leq \frac{k}{\binom{n}{k}} \max_{2\leq i\leq k} \left[\binom{k}{i}\binom{n-k}{k-i} 2^{\binom{i}{2}} \right]. \end{aligned} \tag{44}$$

From eq. (43), we now have to prove that the right-hand side of eq. (44) goes to 0 for $k = 2(1-\varepsilon)\log_2 n$ to finish the proof of Lemma 7.4. We look first at the term $i = 2$:

$$\frac{k}{\binom{n}{k}} \cdot \binom{k}{2}\binom{n-k}{k-2} \cdot 2 \sim \frac{k^5}{n^2} \to 0, \tag{45}$$

where we implicitly used Stirling's approximation for $k = 2(1-\varepsilon)\log_2 n \ll \sqrt{n}$. Now we notice that the maximum in eq. (44) is reached in $i = 2$ because for any $i \in \{2, \cdots, k\}$:

$$\begin{aligned}\frac{\binom{k}{i}\binom{n-k}{k-i}2^{\binom{i}{2}}}{\binom{k}{2}\binom{n-k}{k-2}2^{\binom{2}{2}}} &= \left(\frac{(k-2)!}{(k-i)!}\right)^2 \cdot \frac{(n-2k+2)!}{(n-2k+i)!} \cdot \frac{2^{i(i-1)/2}}{i!}, \\ &\leq k^{2(i-2)} \cdot \left(\frac{1}{n-2k}\right)^{i-2} \cdot 2^{(i+1)(i-2)/2} \cdot \frac{2}{i!}, \\ &\leq \left(\frac{k^2}{n-2k} 2^{(i+1)/2}\right)^{i-2}, \\ &\leq \left(\frac{k^2}{n-2k} 2^{(k+1)/2}\right)^{i-2}.\end{aligned}$$

Since $k = 2(1-\varepsilon)\log_2 n$, $2^{k/2} = n^{1-\varepsilon}$, and so for sufficiently large n, we have

$$\frac{\binom{k}{i}\binom{n-k}{k-i}2^{\binom{i}{2}}}{\binom{k}{2}\binom{n-k}{k-2}2^{\binom{2}{2}}} \leq 1,$$

which (together with eq. (45)) shows that the right-hand side of eq. (44) goes to 0. □

Remark I– It's quite typical that second moment computation involve two independent copies of the same randomness (here S and T, see eq. (41)), and involves statistics of their so-called "overlap", here $|S \cap T|$. For problems involving random vectors rather than random sets, this overlap will often be $|\langle \mathbf{x}, \mathbf{y} \rangle|$.

Remark II – Note that since we have shown that $\mathrm{Var}(Z_{(1-\varepsilon)k^\star}) \ll \mathbb{E}[Z_{(1-\varepsilon)k^\star}]^2$, by Chebyshev's inequality, with high probability $Z_{(1-\varepsilon)k^\star} \geq (1 - o_n(1))\mathbb{E}[Z_{(1-\varepsilon)k^\star}]$. Moreover, $\mathbb{E}[Z_{(1-\varepsilon)k^\star}] \to \infty$ by Lemma 7.2. All in all, we have shown that there exists (with high probability) not only one, but an infinite number of cliques of size $2(1-\varepsilon)\log_2 n$.

7.2 A simple algorithm, and an open problem

Now that we have established the size of the largest clique to be around $2\log_2 n$ by Theorem 7.1, we can ask

Can we find large cliques in G?

The key work here is *find*: we want to produce an algorithm that outputs a clique. Note that since we know the size of the largest clique, we can simply do an exhaustive search over all possible subsets of large enough size: this gives an algorithm which runs in roughly

$$\binom{n}{2\log_2 n} \sim n^{2\log_2 n}$$

operations. While not exponentially large in n, this is still going to infinity faster than any polynomial, and we would like to have an efficient (i.e. polynomial-time) algorithm for this task.

An easily-analyzable such algorithm is a greedy search: we start by $S = \{v\}$ for a random vertex v, and then at each step we choose a vertex $i \in V \backslash S$ such that $S \cup \{i\}$ is a clique. Once such a vertex does not exist, we stop and return S. This algorithm can be stated in a slightly more convenient form for analysis as follows, due to [GM75].

Algorithm 1: Greedy algorithm

Result: A clique S
Input: A graph $G = (V, E)$. We label the vertices $V = \{v_1, \cdots, v_n\}$ arbitrarily;
Initialize $S = \emptyset$;
for $i = 1, \cdots, n$ **do**
  If $S \cup \{v_i\}$ is a clique, $S \leftarrow S \cup \{v_i\}$;
end

Informally, we expect that if the clique S has size p, then the probability that a new vertex is connected to all elements in S is 2^{-p}, and so we will have to see (on average) 2^p vertices before Algorithm 1 is able to increase the size of the clique. This suggests that the size k of the clique returned by Algorithm 1 satisfies $\sum_{i=1}^{k} 2^i = 2^{k+1} - 1 \simeq n$, so that $k \simeq \log_2 n$. This argument can be made rigorous as follows.

Theorem 7.5

For any $\varepsilon > 0$, with probability $1 - o_n(1)$, the output of Algorithm 1 is a clique of size at least $(1-\varepsilon)\log_2 n$.

Proof of Theorem 7.5 – In order to make the analysis simpler, we embed G into a graph on infinitely many nodes, indexed by $\{1, 2, \cdots\}$, in which each edge appears with probability $1/2$. In this way, we can run the procedure of Algorithm 1 for an infinite time: the output of Algorithm 1 is simply the clique after n steps. Denote S_i the clique of Algorithm 1 after step i. Let T_p (for $p \geq 1$) be the number of iterations for the clique S to grow from size $p-1$ to p. By definition, T_p has a geometric distribution with parameter $2^{-(p-1)}$, i.e. for any $l \geq 1$:

$$\mathbb{P}[T_p \leq l] = 1 - (1 - 2^{-(p-1)})^l,$$

and $\mathbb{E}[T_p] = 2^{p-1}$. Moreover, $(T_p)_{p\geq 1}$ are independent random variables. Therefore, for any k:

$$\begin{aligned}
\mathbb{P}(|S_n| \geq k) &= \mathbb{P}(T_1 + \cdots + T_k \leq n), \\
&\geq \prod_{p=1}^{k} \mathbb{P}\left(T_p \leq \frac{n}{k}\right), \\
&= \prod_{p=1}^{k} \left[1 - \left(1 - 2^{-(p-1)}\right)^{n/k}\right], \\
&\overset{(a)}{\geq} \left[1 - \left(1 - 2^{-k}\right)^{n/k}\right]^k, \\
&\overset{(b)}{\geq} 1 - k\left(1 - 2^{-k}\right)^{n/k}.
\end{aligned}$$

In (a) we used that $2^{-(p-1)} \geq 2^{-k}$, and in (b) that $(1-x)^k \geq 1 - kx$ for $x \in [0,1]$ (by convexity). If $k = (1-\varepsilon)\log_2 n$, we have:

$$\begin{aligned}
1 - \mathbb{P}(|S_n| \geq k) &\leq k \cdot \left(1 - n^{-(1-\varepsilon)}\right)^{n/k}, \\
&\sim k \exp\left\{-\frac{n^\varepsilon}{k}\right\} \to 0.
\end{aligned}$$

This ends the proof. □

Challenge 7.2. *Compute the average time complexity of Algorithm 1.*

An important open problem – Algorithm 1 is a naive greedy algorithm that can provably return cliques of size $(1-\varepsilon)\log_2 n$, for any $\varepsilon > 0$. On the other hand, we know (Theorem 7.1) that the largest clique is of size around $2\log_2 n$ with high probability. This raises the following open problem.

Open Problem 7.1

Does there exists a polynomial-time algorithm that returns, for $G \sim \mathrm{ER}(n, 1/2)$, a clique of size $(1+\varepsilon)\log_2 n$ with high probability?

Remarkably, this question (which can be tracked to [Kar76]) is still unanswered. A positive or negative answer to this question would be a breakthrough, and would likely deepen our understanding of computational hardness in average-case scenarios.

Remark – Finally, let us mention that all we said can be generalized to other probabilities than $1/2$ in the Erdős–Rényi distribution.

Challenge 7.3. *Generalize all the results above (Theorem 7.1 and the greedy algorithm and its analysis) to the case $p \in (0,1)$: as you will see, all this extends there up to replacing $\log_2 n$ by $\log_{1/p} n$ in some places.*

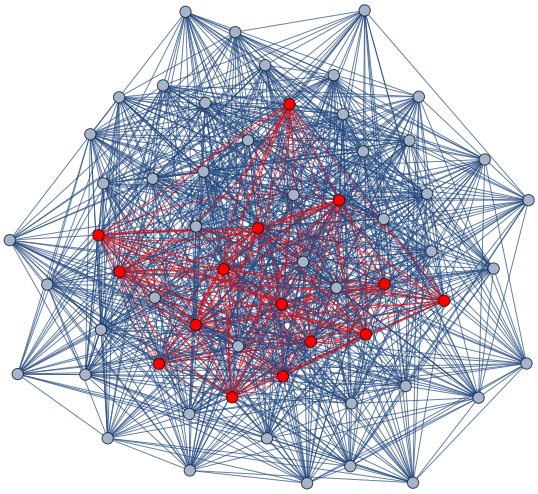
7.3 The planted clique problem

The *planted clique* model was introduced in [Jer92]: it is also related to finding cliques in random graphs, but is slightly different from the question we studied above. Given $G \sim \mathrm{ER}(n, 1/2)$, we will *plant* a clique inside G by forcing some vertices to be connected. We now ask: given the knowledge of G, can a statistician recover the planted clique?

More formally, we choose a set $K \subseteq [n]$ with $|K| = k$ uniformly at random, and we then draw the edges of G according to the distribution:

$$\mathbb{P}[i \sim j] = \begin{cases} 1 & \text{if } (i,j) \in K, \\ 1/2 & \text{otherwise .} \end{cases} \tag{46}$$

The graph might then look as follows, where the planted clique is shown in red.



One way to formulate the planted clique question is in the form of *hypothesis testing* (see Section 6): can we distinguish a graph G drawn from the distribution of eq. (46) from a graph $G \sim \mathrm{ER}(n, 1/2)$? Here, we focus rather on the *recovery* question: can we recover the clique, or at least a fraction of the clique?

7.3.1 Information-theoretic thresholds

- If $k \leq (2-\varepsilon)\log_2 n$, we have shown in Section 7.1 that there is an infinite number of cliques of size k (with high probability). This implies that recovering the planted clique with high probability is **impossible** if it has a size $k \leq (2-\varepsilon)\log_2 n$.
- On the other hand, if $k \geq (2+\varepsilon)\log_2 n$, then similar arguments than the ones used to prove Theorem 7.1 imply that the planted clique is (whp) the only clique of this size! In particular, we can do an exhaustive (very inefficient) search over all subsets, and we will be assured to recover the clique.

Challenge 7.4. *Given a node not in the planted clique K, give a high-probability bound on the number of its neighbors which are in K. Deduce from this and Theorem 7.1 that if $k \geq C\log_2 n$ (for some large enough constant $C > 0$), the planted clique is the unique clique of size k in the graph.*

The following challenge is harder but shows the desired result.

Challenge 7.5 (Uniqueness of the planted clique)**.** *Using combinatorial arguments similar to the ones used in the proof of Theorem 7.1, show that if $k \geq (2+\varepsilon)\log_2 n$ (for any $\varepsilon > 0$), the planted clique is the unique clique of size k in the graph.*

The points above can be summarized by saying that $2\log_2 n$ is the **information-theoretic threshold** for recovering (and also detecting) the planted clique.

Finally, it is possible to show that when the planted clique is the unique maximum clique, then one can find it in super-polynomial time.

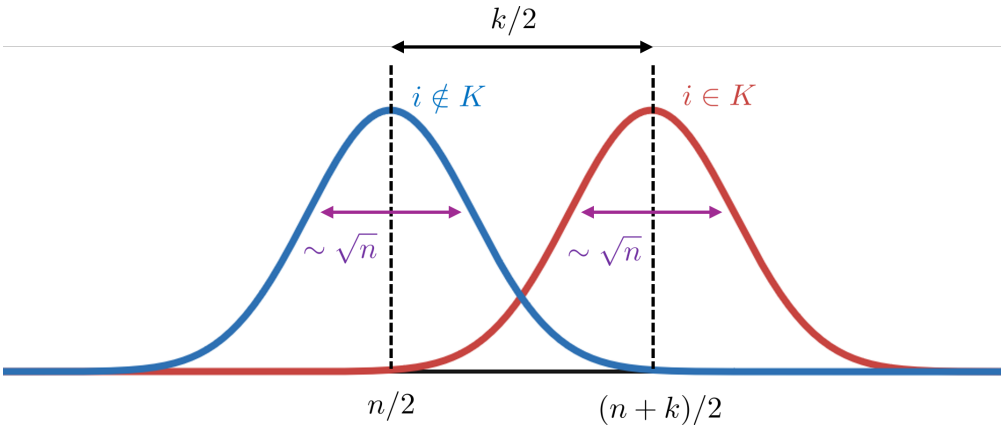
Challenge 7.6. *For $k \geq (2+\varepsilon)\log_2 n$, design an algorithm that runs in time $n^{\Theta(\log_2 n)}$ (quasi-polynomial time) that recovers the planted clique with high probability.*

These different (hard) challenges are solved at the end of this section.

7.3.2 Efficient algorithms, and computational hardness

But what about efficient (polynomial-time) algorithms? Can they recover the planted clique? A first naive approach is based on the following "degree test" informal argument.

Degree test – By the central limit theorem, the degree of any given vertex not in the planted clique is a random variable whose distribution approaches a Gaussian distribution, with mean $(n-1)/2 \simeq n/2$, and variance $\mathcal{O}(n)$. On the other hand, a vertex in the planted clique has a degree distribution with mean $(n+k)/2$, and variance $\mathcal{O}(n)$ as well.



Therefore, if $k \gg \sqrt{n}$, the two distributions are mostly separated, and we expect to be able to distinguish vertices in the planted clique simply by their degrees. Based on this idea, one can show that simply collecting the vertices with highest degree recovers the planted clique with high probability when $k = \Omega(\sqrt{n \log n})$ (the additional $\sqrt{\log n}$ factor arises from a union bound argument).

Challenge 7.7. *Prove that if $k \geq C\sqrt{n \log n}$ (for some constant $C > 0$), and if S is the set of k vertices with the highest degree, then $\mathbb{P}[S = K] \to 1$ as $n \to \infty$.*

This argument goes back to [Kuč95]. The proof is based on classical concentration inequalities, and (if you are stuck) you can find a solution in [WX23]. Still, $\Theta(\sqrt{n \log n})$ is very far from the information-theoretic threshold $2\log_2 n$!

Spectral methods – Another (more clever) algorithm uses the idea of *spectral clustering*, which you have been introduced to in class. Here we consider it for the hypothesis testing problem, but it can be extended to recovery [AKS98]. It manages to detect the planted clique as soon as $k \geq C\sqrt{n}$, for some (large enough) constant $C > 0$, as follows from the following challenge.

Challenge 7.8 (Spectral method for the planted clique problem)**.** *For a given graph G, we denote $B(G)$ the signed adjacency matrix, i.e. $B_{ij} = 1$ if $(i,j) \in E$, and $B_{ij} = -1$ otherwise.*

(*i*) *(Hard question) Let $G \sim \mathrm{ER}(n, 1/2)$. Show that, with probability $1 - o_n(1)$, $\|B(G)\|_{\mathrm{op}} \leq c\sqrt{n}$ for some $c > 0$.*

(*ii*) *Assume now that G has a planted clique of size k. Show that $\|B(G)\|_{\mathrm{op}} \geq k$.*

(*iii*) *Conclude that when $k > C\sqrt{n}$ (for some $C > 0$), the largest eigenvalue of $B(G)$ can be used to detect a planted clique, with high probability.*

This algorithm (and its variant for recovery) was analyzed in [AKS98], where the authors further showed that a modification of the algorithm allows it to recover the planted clique for $k > c\sqrt{n}$, for arbitrarily small (but fixed as $n \to \infty$) constant $c > 0$, with a complexity that now grows as $n^{\mathcal{O}(\log(1/c))}$.

The failure of these algorithms (as well as many others!) beyond the regime $k \gtrsim \sqrt{n}$ motivates the folllowing famous conjecture.

Conjecture 7.6 (*Hardness of planted clique*)

If $k = o(\sqrt{n})$, then there is no polynomial-time algorithm that can (with high probability) recover the planted clique.

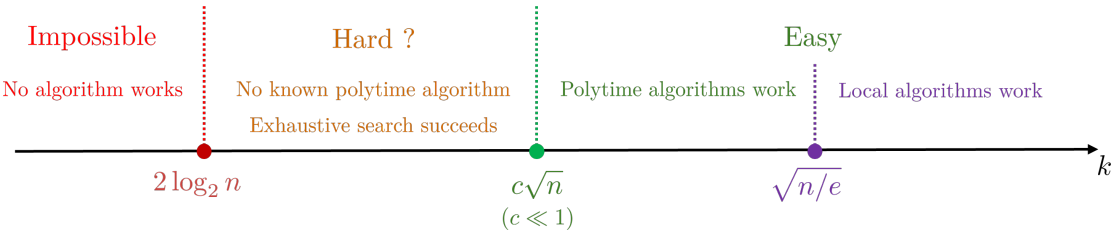


Figure 4: A schematic view of Conjecture 7.6.

Belief propagation – Beyond spectral methods, many other algorithms have been studied in the planted clique problem. A notable example are variants of *belief propagation* (BP), which aim at approximating averages under the Bayesian *posterior probability distribution*, and has origins in statistical physics. In many average-case problems, algorithms derived from statistical physics ideas (notably BP) have been showed to achieve the optimal polynomial-time performance among a large class of "local" algorithms[8], see [BPW18]. As an example, in the planted clique problem, a method derived from belief propagation was shown to recover the planted clique for $k \geq \sqrt{n/e}$ in quasi-linear time (i.e. time $\mathcal{O}(n \log^p n)$ for some $p \geq 1$) [DM15]. Extending these results, it is conjectured that $\sqrt{n/e}$ is the onset of hardness for "local" algorithms [Mon15].

Towards Conjecture 7.6 – In such average-case models, proving the non-existence of a good algorithm is a problem that we essentially do not know how to attack. Instead, our best evidence for the difficulty of the problem usually comes from showing that increasingly powerful classes of algorithms fail. In the planted clique problem, this was shown to be the case when $k/\sqrt{n} \to 0$ for Belief Propagation algorithms [DM15],

[8] Here this essentially corresponds to algorithms that, for each node i, estimate if i is in the planted clique or not only based on the knowledge of a local neighborhood of i in the graph. For instance, the degree test mentioned above is clearly local.

for Sum-of-Squares algorithms [Bar+19] (which you will encounter later in class), for Metropolis processes [Jer92], statistical query algorithms [Fel+17], and geometric arguments based on ideas from statistical physics also support Conjecture 7.6 [GZ19].

Average-case reductions – The planted clique conjecture is one of the most famous and well-studied examples of a *computational-to-statistical gap*: we already saw another example with sparse PCA in Section 6. In analogy with the worst-case theory of NP-hardness, there is a rich line of work which aims at showing that other problems that we believe to be hard to solve can be reduced to the planted clique conjecture (in the sense that if we prove that Conjecture 7.6 holds, then a similar statement will directly hold for these problems). This is the case for instance of the sparse PCA problem we discussed above, see e.g. [BR13], and many others (submatrix localization, compressed sensing, biclustering, planted dense subgraph problems, ...)

Solutions to Challenges of Section 7.3.1

Solution to Challenge 7.4

Part I – For a node $i \in V$, we denote ∂i the set of its neighbors. We show that

$$\mathbb{P}\left[\forall i \in V \backslash K \,:\, |\partial i \cap K| \leq \frac{k}{2} + \sqrt{k \log n}\right] = 1 - o_n(1). \tag{47}$$

To prove this, we use the union bound:

$$\begin{aligned}\mathbb{P}\left[\exists i \in V \backslash K \,:\, |\partial i \cap K| > \frac{k}{2} + \sqrt{k \log n}\right] &\leq n\mathbb{P}\left[|\partial i \cap K| > \frac{k}{2} + \sqrt{k \log n}\right], \\ &\leq n\mathbb{P}\left[\sum_{j=1}^{k} (x_j - \mathbb{E}[x_j]) > \sqrt{k \log n}\right],\end{aligned}$$

where $x_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{0,1\})$. Using Hoeffding's inequality (Theorem 2.2: notice that we don't need the factor 2 in Hoeffding's inequality if we look at one-sided tail bounds) we get:

$$\mathbb{P}\left[\exists i \in V \backslash K \,:\, |\partial i \cap K| > \frac{k}{2} + \sqrt{k \log n}\right] \leq n e^{-2(k \log n)/k} = \frac{1}{n}.$$

This proves eq. (47). In particular,

Part II – Assume now that $k \geq C \log_2 n$. We take $C > 0$ a large enough constant such that

$$\frac{k}{2} + \sqrt{k \log n} = k\left[\frac{1}{2} + \sqrt{\frac{\log n}{k}}\right] \leq k\left[\frac{1}{2} + \sqrt{\frac{\log 2}{C}}\right] \leq \frac{3k}{5}.$$

Therefore by eq. (47):

$$\mathbb{P}\left[\exists i \in V \backslash K \,:\, |\partial i \cap K| > \frac{3k}{5}\right] = o_n(1). \tag{48}$$

Denote $G_0 \sim \mathrm{ER}(n, 1/2)$ the graph without the planted clique, and G the graph G_0 to which we added the planted clique K. Let $T \subseteq V$ be a clique of size k in G. Clearly, $T \backslash K \subseteq T$ is then a clique in G_0. Therefore, by Theorem 7.1 (with high probability):

$$|T \backslash K| \leq 3 \log_2 n$$

This implies that

$$|T \cap K| \geq k\left[1 - \frac{3\log_2 n}{k}\right] \geq k\left[1 - \frac{3}{C}\right] \geq \frac{2k}{3}, \tag{49}$$

taking again $C > 0$ large enough. Notice that all nodes in $T\backslash K$ have at least $|T \cap K|$ neighbors inside the planted clique, eq. (48) and eq. (49) show that $T\backslash K = \emptyset$, and thus $T = K$. $\square$

Solution to Challenge 7.5

Here we only assume $k \geq (2+\varepsilon)\log_2 n$. For any $S \subseteq V$ with $|S| = k$, if we denote $l \coloneqq |S\backslash K|$, we have

$$\begin{aligned}
&\mathbb{P}(S \text{ is a clique}) \\
&= \mathbb{P}(S \cap K \text{ is a clique}) \cdot \mathbb{P}(\text{All nodes in } S \cap K \text{ and } S\backslash K \text{ are connected}), \\
&= \left(\frac{1}{2}\right)^{\binom{l}{2}} \cdot \left(\frac{1}{2}\right)^{l(k-l)}.
\end{aligned}$$

Then we use the union bound. Let A be the event "there exists $S \subseteq V$ with $|S| = k$, $S \neq K$, and S is a clique". Then

$$\begin{aligned}
\mathbb{P}[A] &\leq \sum_{l=1}^{k} \binom{n-k}{l}\binom{k}{k-l}\left(\frac{1}{2}\right)^{\binom{l}{2}} \cdot \left(\frac{1}{2}\right)^{l(k-l)}, \\
&\leq \sum_{l=1}^{k} \underbrace{\binom{n-k}{l}\binom{k}{k-l} 2^{-lk+\frac{l(l+1)}{2}}}_{=:T_l}.
\end{aligned}$$

We now show that for $k \geq (2+\varepsilon)\log_2 n$:

$$nT_1 = o_n(1), \tag{50}$$

and:

$$\max_{2 \leq l \leq n} \left(\frac{T_l}{T_1}\right) \leq 1. \tag{51}$$

Clearly, the combination of eqs. (50) and (51) imply that $\mathbb{P}[A] = o_n(1)$, which ends the proof.

Proof of eq. (50) – Notice that

$$nT_1 = n \cdot (n-k) \cdot k \cdot 2^{-k+1} \lesssim n^2 \cdot k \cdot 2^{-k} = 2^{2\log_2 n + \log_2 k - k}.$$

Clearly, $2\log_2 n + \log_2 k - k \to -\infty$ as $n \to \infty$ if $k \geq (2+\varepsilon)\log_2 n$.

Proof of eq. (51) – We have:

$$\begin{aligned}
\frac{T_l}{T_1} &= \frac{\binom{n-k}{l}\binom{k}{k-l} \cdot 2^{-lk+\frac{l(l+1)}{2}}}{k(n-k)2^{-k+1}}, \\
&= \frac{(n-k-1)!}{l!(n-k-l)!} \cdot \frac{(k-1)!}{l!(k-l)!} \cdot 2^{\frac{l(l+1)}{2}-1-(l-1)k}, \\
&\leq \frac{n^{l-1}k^{l-1}}{(l!)^2} \cdot 2^{\frac{(l-1)(l+2)}{2}-(l-1)k},
\end{aligned}$$

$$\leq \frac{1}{(l!)^2}\left(nk \cdot 2^{\frac{(l+2)}{2}-k}\right)^{l-1}.$$

Because $l \leq k$, we get from this:

$$\begin{aligned}\frac{T_l}{T_1} &\leq \left(2nk \cdot 2^{-\frac{k}{2}}\right)^{l-1}, \\ &\leq \left(2^{1+\log_2 n+\log_2 k-\frac{k}{2}}\right)^{l-1}.\end{aligned}$$

Since $k \geq (2+\varepsilon)\log_2 n$, we have $\log_2 n + \log_2 k - \frac{k}{2} \to -\infty$ as $n \to \infty$. This implies (for large enough n): $T_l \leq T_1$ for all $l \geq 2$, which ends the proof. □

Solution to Challenge 7.6

If $k = \Theta(\log_2 n)$, an exhaustive search algorithm will recover the planted clique K in time $\binom{n}{k} \sim n^k = n^{\Theta(\log_2 n)}$.

We design here an improvement that allows to recover the planted clique (with high probability) for any $k \geq C\log_2 n$ in time $n^{\Theta(\log_2 n)}$, where $C > 0$ is a large enough constant. The steps are the following.

Step 1 – By the exhaustive search algorithm, we can find T a clique, with $|T| = C\log_2 n$, in time $n^{\Theta(\log_2 n)}$. For any $\varepsilon > 0$, it holds that $|T \cap K| \geq (C - 2 - \varepsilon)\log_2 n$ with high probability. Indeed, $T \backslash K$ is a clique that has no vertex in common with the planted clique, so its size is upper bounded by $(2+\varepsilon)\log_2 n$ by Theorem 7.1, see also the solution to Challenge 7.4.

Step 2 – We now "expand" our first guess T, by picking S_T to be the set of vertices that have at least $3|T|/4$ neighbors in T. By Step 1, any $i \in K$ has at least $(C-2-\varepsilon)\log_2 n$ neighbors in T, and we can pick C large enough so that $(2+\varepsilon)/C \leq \frac{1}{4}$. All in all we have shown $\mathbb{P}[\forall T \text{ clique of size } C\log_2 n,\ K \subseteq S_T] \to 1$ as $n \to \infty$.

Step 3 – Now that we have built a set that contains K with high probability, we clean it. We let $G[S_T]$ be the induced subgraph on S_T, and we define $\hat{K}_T$ as the k vertices in S_T with the highest degree in $G[S_T]$. We will show that $\mathbb{P}[\forall T,\ \hat{K}_T = K] \to 1$ as $n \to \infty$. This is done in two substeps.

(*i*) For any given T and any node $i \in V \backslash K$,

$$\mathbb{P}(i \in S_T) = \mathbb{P}\left[\sum_{j=1}^{|T|}(x_j - \mathbb{E}[x_j]) \geq \frac{|T|}{4}\right],$$

where $x_j \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{0,1\})$. Notice that for $i \in T$ the sum actually runs only up to $|T|-1$, but this will not make any difference to the argument. Using Hoeffdings's inequality we get:

$$\mathbb{P}(i \in S_T) \leq \exp\left\{-\frac{|T|}{8}\right\} = n^{-C\log(2)/8}.$$

Since, for a given T, all the events $\{i \in S_T\}_{i \in V\backslash K}$ are i.i.d. we have, for any $r \geq 0$:

$$\mathbb{P}(|S_T \backslash K| = r) = \binom{n-k}{r}\mathbb{P}(i \in S_T)^r[1 - \mathbb{P}(i \in S_T)]^{n-k-r} \leq n^{r[1-C\log(2)/8]}.$$

Using the geometric series we even have, for large enough $C > 0$:

$$\begin{aligned}\mathbb{P}(|S_T \backslash K| \geq r) &\leq \sum_{l=r}^{n-k} n^{l[1-C\log(2)/8]}, \\ &\leq n^{r[1-C\log(2)/8]} \frac{1}{1-n^{1-C\log(2)/8}}, \\ &\leq 2n^{r[1-C\log(2)/8]}.\end{aligned}$$

Now we can crudely union bound over the choice of T, since $|T| = C\log_2 n$:

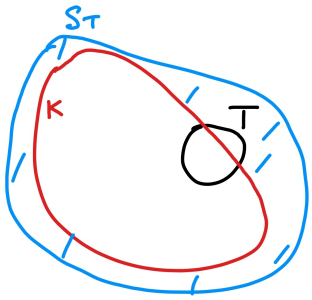
$$\begin{aligned}\mathbb{P}(\exists T \text{ a clique} : |T| = C\log_2 n \text{ and } |S_T \backslash K| \geq r) &\leq 2\binom{n}{C\log_2 n} n^{r[1-C\log(2)/8]}, \\ &\leq 2n^{C\log_2 n - r[C\log(2)/8-1]}.\end{aligned}$$

The exponent in the right-hand side goes to $-\infty$ for $n \to \infty$ as long as

$$r \geq \left[\frac{8C}{C\log(2)-8} + \varepsilon\right] \log_2 n,$$

for some $\varepsilon > 0$. Since $8/\log(2) \simeq 11.54 < 12$, we have shown that for a large enough constant $C > 0$:

$$\mathbb{P}(\exists T \text{ a clique} : |T| = C\log_2 n \text{ and } |S_T \backslash K| \geq 12\log_2 n) = o_n(1). \tag{52}$$



(*ii*) Now that we have shown that $|S_T \backslash K|$ is small, we will show that this implies that (whp), in the induced subgraph $G[S_T]$, there will be no node of high degree in $S_T \backslash K$, while all nodes in $S_T \cap K$ will have high degree because they belong to the planted clique. For any $i \in S_T$, we denote $d(i)$ its degree *in the subgraph* $G[S_T]$. We clearly have

$$\begin{aligned}\max_{i \in S_T \backslash K} d(i) &\leq |S_T \backslash K| + \max_{i \in S_T \backslash K} \#\{j \in K : i \sim j\}, \\ &\leq |S_T \backslash K| + \max_{i \in V \backslash K} \#\{j \in K : i \sim j\}.\end{aligned}$$

However, we know (cf eq. (47), which is proven using Hoeffding's inequality and the union bound) that, with probability $1 - o_n(1)$:

$$\max_{i \in V \backslash K} \#\{j \in K : i \sim j\} \leq \frac{k}{2} + \sqrt{k \log n}.$$

Combining this and eq. (52) we get

$$\mathbb{P}\left(\forall T, \max_{i \in S_T \backslash K} d(i) \leq 12\log_2 n + \frac{k}{2} + \sqrt{k \log n}\right) = 1 - o_n(1).$$

Since $k \geq C \log_2 n$, for large enough C, we get:

$$\mathbb{P}\left(\forall T,\ \max_{i \in S_T \setminus K} d(i) \leq \frac{3k}{4}\right) = 1 - o_n(1).$$

On the other hand, for any $i \in K$, $d(i) \geq k-1$ since $K \subseteq S_T$ with high probability. Therefore, $\mathbb{P}[\forall T,\ \hat{K}_T = K] = 1 - o_n(1)$, which ends the proof. $\square$

8 Optimization of random high-dimensional functions and the Kac-Rice formula

Disclaimer – An important class of problems in high-dimensional statistics and data science can be framed in the language of *optimization*. Optimization is a rich field of mathematics and theoretical computer science in its own right: here we will discuss some very high-level motivation, and introduce the simplest model of optimization of a "typical" high-dimensional function that one can imagine. As it turns out, understanding how to minimize this simple function is a daunting task, and it took decades of works of probabilists, theoretical computer scientists and theoretical physicists (as this object was introduced in the statistical physics literature) to obtain a clear answer.

8.1 Loss landscapes in statistics and machine learning

A broad motivation – In statistical learning, and more precisely when doing *supervised learning*, one is given access to a set of data samples $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathbb{R}^d$ (think of them as images), and corresponding labels $(y_1, \cdots, y_n)$, see Fig. 5. These labels can be discrete (in classification) or continuous (in regression).

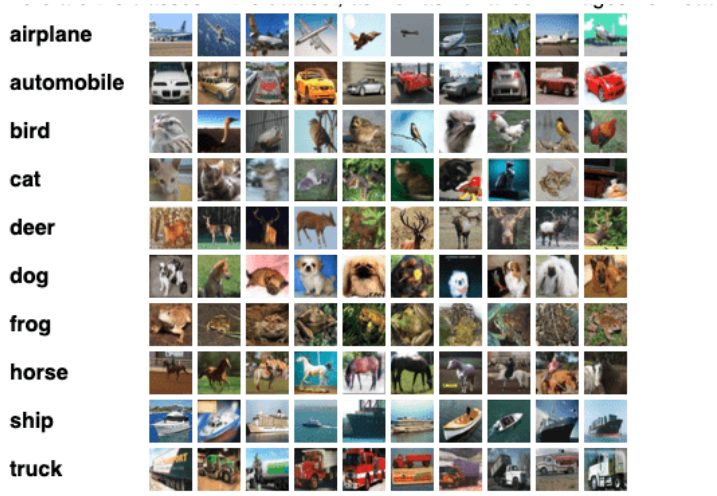


Figure 5: An example of a series of labeled images (taken here from the CIFAR-10 dataset).

The statistician then constructs a learning model (for instance a complex neural network) which is a *parametric function*: one can write that $\hat{y} = f_{\boldsymbol{\theta}}(\mathbf{x})$, where $\mathbf{x}$ is a data sample, and $\hat{y}$ is the label predicted by the model with parameters $\boldsymbol{\theta}$. Fig. 6 and 7 give examples of two models of functions $f_{\boldsymbol{\theta}}(\mathbf{x})$. *Learning from data* amounts to solve the following question:

> Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, how to find the best possible parameters $\boldsymbol{\theta}$ to most accurately predict the label of a new data sample $\mathbf{x}_{\text{new}}$?

Mathematical formulation – Let us try to make this question mathematically formal. For this we imagine that there is an underlying data distribution $\mathbb{P}_d$ from which $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are drawn independently. For instance $\mathbb{P}_d$ can be the law of a random image of an animal: importantly, we do not require that we know what this distribution is,

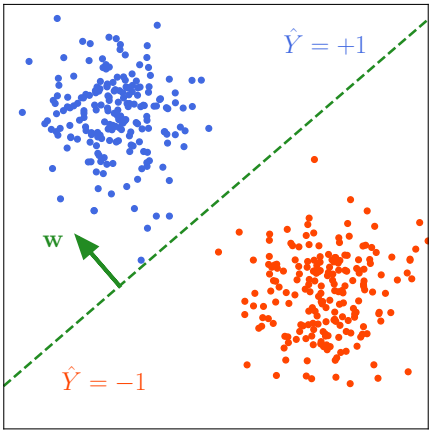


Figure 6: A linear classifier (a so-called *perceptron* model), where the decision boundary is a hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle = b$ for some $\mathbf{w} \in \mathcal{S}^{d-1}$ and $b \in \mathbb{R}$. The prediction is $\hat{y}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - b)$. The parameters $\boldsymbol{\theta} \coloneqq (\mathbf{w}, b)$ are adjusted to better classify the data.

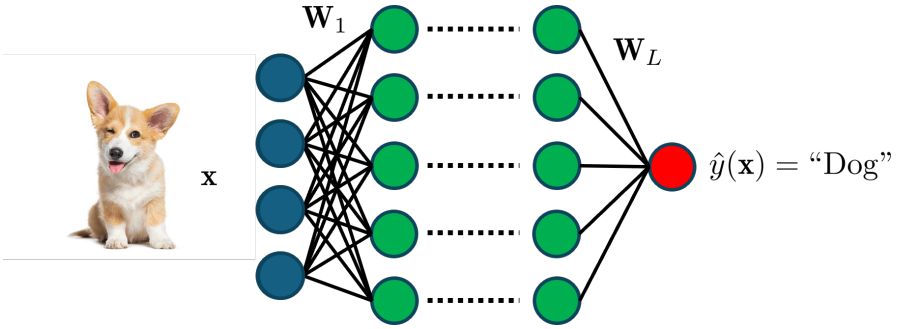


Figure 7: An example of a simple neural network model. Given the input $\mathbf{x} \in \mathbb{R}^d$, the output is predicted as $\hat{y} = \sigma_L(\mathbf{W}_L \sigma_{L-1}(\cdots \sigma_1(\mathbf{W}_1 \mathbf{x})))$, where $(\sigma_1, \cdots, \sigma_L)$ are real functions applied entrywise. The parameters are the weights $\boldsymbol{\theta} = (\mathbf{W}_1, \cdots, \mathbf{W}_L)$: they are matrices that the statistician adjusts to correctly classify the data.

just that it exists! In the same way, we imagine that there exists an ideal estimator $y(\mathbf{x})$ that, given a point $\mathbf{x}$, always returns the correct label.

If $\boldsymbol{\theta} \in \mathbb{R}^p$, in order to find the best[9] possible values of the parameters $\boldsymbol{\theta}$ for our model we would ideally like to compute:

$$\boldsymbol{\theta}^\star \coloneqq \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_d} \left[\left(y(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{x}) \right)^2 \right]. \tag{53}$$

The quantity on the right-hand side of eq. (53) is called the *population risk* (or population loss):

$$R(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_d} \left[\left(y(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{x}) \right)^2 \right]. \tag{54}$$

Unfortunately, in general we can not compute the population risk (nor its minimizer $\boldsymbol{\theta}^\star$), because we do not have access to the data distribution $\mathbb{P}_d$ or to the ideal estimator $y(\mathbf{x})$ (this is precisely what we want to approximate).

Empirical risk – This is where our data samples can help us. If we have enough data points $(\mathbf{x}_i, y_i)_{i=1}^n$, we can hope that their empirical distribution approximates well

[9] This depends on what we mean by "best". In particular one can consider other performance measurements than the square loss $(y - f_{\boldsymbol{\theta}})^2$.

enough the true data distribution. In mathematical terms, we can consider the *empirical risk*

$$\hat{R}_{\mathcal{D}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)^2 \tag{55}$$

as an approximation of the population risk of eq. (54). This is an example of more general *M-estimators*, which are functions of the type:

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) := \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell[\boldsymbol{\theta}; (\mathbf{x}_i, y_i)]. \tag{56}$$

In this more general context, the population loss minimizer would read

$$\boldsymbol{\theta}^\star := \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_d} \ell[\boldsymbol{\theta}; (\mathbf{x}, y(\mathbf{x}))]. \tag{57}$$

While the estimator of eq. (56) now only depends on the data samples we observed, and so we can in theory compute it, there remains several very important questions.

(Q1) Does $f_{\boldsymbol{\theta}^\star}(\mathbf{x})$ provide a good approximation of $y(\mathbf{x})$? In other words, is $R(\boldsymbol{\theta}^\star)$ small?

(Q2) Is the estimator $\hat{\boldsymbol{\theta}}(\mathcal{D})$ a good approximation of $\boldsymbol{\theta}^\star$? More generally, when does the empirical risk approximate well the population risk?

(Q3) Can the estimator of eq. (56) be computed efficiently?

(Q1) is solely a question about our choice of learning model: it amounts to understand if our parametric model is "expressive enough" to learn a good approximation to $y(\mathbf{x})$. On the other hand (Q2) depends on the data samples we receive (in particular on the number n of data samples, as we expect the empirical risk to approach the population risk when the number of data samples gets large enough), its study is a rich mathematical field. When $d = \mathcal{O}(1)$ and $n \to \infty$ this is one of main topics of classical statistics, and (Q3) is related to low-dimensional optimization.

Motivated by recent progress in learning algorithms, we consider here a *high-dimensional regime* where *both* $d, n \to \infty$. (Q2) has been studied as well in this case provided n grows sufficiently fast with respect to d, see for instance [MBM18]. In this lecture we will put our focus on understanding (Q3) in this high-dimensional regime. Notice that in general the empirical risk is a non-convex function of θ (even if the population risk is convex[10]): computing $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is thus a non-trivial optimization problem.

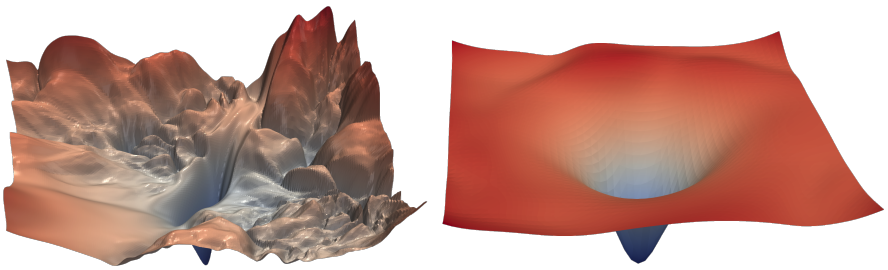
8.2 A (very) simplified model: Gaussian processes on the sphere

For the reasons we described above, local optimization of empirical risk functions is one of the workhorses of modern machine learning. As we have discussed above however, these functions are in general non-convex, so characterizing the performance of these procedures is a daunting task. Here we will simplify the model to make this amenable to a mathematical treatment. The empirical risk function of eq. (55) or eq. (56) has two crucial properties:

(*i*) It is a *random* function, because the data samples $(\mathbf{x}_i, y_i)$ are drawn randomly.

(*ii*) In the context of modern machine learning and large neural networks, it is a *high-dimensional* function, since these models have a very large number of parameters.

[10] Note that if the population risk is convex, then an answer to (Q2) can sometimes inform us on (Q3).



(a) A "rough" optimization landscape, with a great number of local minima and critical points.

(b) A smooth landscape, intuitively easier to optimize via local algorithms.

Figure 8: "Naive" representation of a rough (left) and smooth (right) optimization landscape. Note that these naive pictures do not take into account the high-dimensionality of the problem, and should be taken with caution! Pictures taken from [Li+18].

We now define one of the simplest models that retains these two key properties. We consider a unit-norm vector $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$ for some $d \geq 1$ (and we will analyze the limit $d \to \infty$), and a function $f(\boldsymbol{\theta})$ that is a centered Gaussian field: essentially this is a random function such that all $(f(\boldsymbol{\theta}))_{\boldsymbol{\theta} \in \mathcal{S}^{d-1}}$ are jointly Gaussian with zero mean. The distribution of such functions is thus uniquely characterized by their covariance structure, which we assume is given by

$$\mathbb{E}[f(\boldsymbol{\theta}) f(\boldsymbol{\theta}')] = \xi(\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle), \tag{58}$$

for some real and continuous function ξ. Notice that this is equivalent to saying that the covariance is a function of the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. The following theorem allows us to express all functions satisfying eq. (58) in the monomial basis.

Theorem 8.1 (*Schoenberg [Sch42]*)

Any centered Gaussian field that satisfies eq. (58) for a continuous function ξ can be written as

$$f(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} a_k \sum_{1 \leq i_1, \cdots, i_k \leq d} J_{i_1 \cdots i_k} \theta_{i_1} \cdots \theta_{i_k}, \qquad (\boldsymbol{\theta} \in \mathcal{S}^{d-1})$$

where $(a_k)_{k \geq 1}$ are real coefficients, and for each $k \geq 2$, $J_{i_1 \cdots i_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Moreover

$$\xi(t) = \sum_{k \geq 1} a_k^2 t^k.$$

Simplifying further, we will consider the case in which $a_k = \delta_{kp}$ for some $p \geq 3$ (and we change notations from $\boldsymbol{\theta}$ to $\mathbf{x}$):

$$f_p(\mathbf{x}) := \sum_{1 \leq i_1, \cdots, i_p \leq d} J_{i_1 \cdots i_p} x_{i_1} \cdots x_{i_p} \qquad (\mathbf{x} \in \mathcal{S}^{d-1}), \tag{59}$$

in which $J_{i_1 \cdots i_p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The function of eq. (59) is sometimes called the pure spherical p-spin model, as it originated under this name in the statistical physics literature [CS92].

Our goal – We will try to characterize the optimization landscape of $f_p(\mathbf{x})$, to understand if it is easy to optimize or not. One way to do this is to count the number of critical

points, and of local minima, in different regions of the landscape, see Fig. 8. As one can easily be convinced by a picture, f having many such critical points is generically a sign of a complex topology of its sublevel sets $\{f(\mathbf{x}) \leq u\}$, (see also Fig. 9 which is introduced later), which we can expect will make optimization algorithms struggle. This intuition can be made mathematically precise, and here we will focus on understanding the number of critical points (especially local minima) of f_p. As we will see in the end, this will allow us to very precisely probe the topology of the optimization landscape, and the performance of local optimization algorithms.

Challenge 8.1. *Consider the function of eq. (59) for $p \in \{1, 2\}$. What do the global minima of f_p correspond to? What can you say about the critical points of f_p? How many are there? (Be careful, you are looking at critical points on the sphere!)*

8.3 The Kac-Rice formula

We introduce here the Kac-Rice formula, which will be our main tool to compute the number of critical points of a random function. As critical points are zeros of the gradient, we consider equivalently:

Given a random smooth function $g : \mathbb{R}^d \to \mathbb{R}^d$, can we count the number of zeros of g?

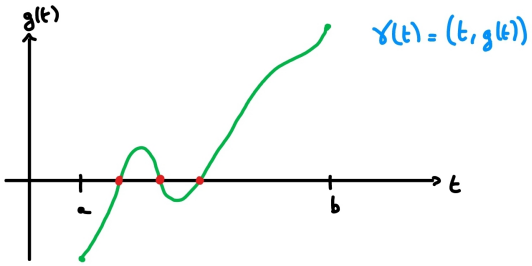
Keep in mind that this number of zeros it itself random: more precisely we will focus here on computing its expected value.

8.3.1 Intuition and formula in the case $d = 1$

Let $g : [a, b] \to \mathbb{R}$ be a real and smooth function (so far we do not consider its randomness). We define Z_g as the number of zeros of g. Since $[a, b]$ is uncountable, we can not write a formula of the type:

$$Z_g = \sum_{x \in [a,b]} \mathbb{1}\{g(x) = 0\}.$$

On the other hand, since g is smooth we can compute Z_g as



$$Z_g = \oint_\gamma \delta(y) \mathrm{d}\omega.$$

Applying naively the formula for the line integral along the path parametrized by $t \in [a, b] \mapsto (t, g(t))$, we get

$$Z_g = \int_a^b \delta(g(t))|g'(t)|\mathrm{d}t. \tag{60}$$

This derivation was quite heuristic, since we applied the line integral formula with an argument which is a distribution rather than a smooth function. This intuition can however be made rigorous, as you are encouraged to show!

Challenge 8.2 (Kac's counting formula)**.** *Let $g : [a, b] \to \mathbb{R}$ be a $\mathcal{C}^1$ function with $g(a) \cdot g(b) \neq 0$, and such that for all $x \in [a, b]$, $g(x) = 0 \Rightarrow g'(x) \neq 0$. Then*

$$Z_g = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_a^b \mathbb{1}\{|g(x)| < \varepsilon\} |g'(x)| \mathrm{d}x.$$

- Notice that, in the sense of distributions, $\mathbb{1}\{|x| < \varepsilon\}/(2\varepsilon) \to_{\varepsilon \downarrow 0} \delta(x)$, so we essentially recover eq. (60) from Kac's counting formula.
- The condition $g(x) = 0 \Rightarrow g'(x) \neq 0$ ensures that the zeros of g are isolated, and therefore that there is only a finite number of them, see Challenge 8.3. In a sense, this condition is essential already to make sense of the quantity Z_g!

Challenge 8.3. *Let $d \geq 1$. Show that if $K \subseteq \mathbb{R}^d$ is compact, and $g : \mathbb{R}^d \to \mathbb{R}^d$ is $\mathcal{C}^1$ and such that $g(x) = 0 \Rightarrow \det[Dg(x)] \neq 0$ for all $x \in K$ (where $Dg(x)$ is the Jacobian of g at x), then $Z_g := \#\{x \in K : g(x) = 0\} < \infty$.*

`(Hint: what can you say about the set` $g^{-1}(\{0\})$`?)`

So far we have not considered the randomness of g. Imagine that g is now a random function, which almost surely satisfies the hypotheses of Challenge 8.2. Then we can write eq. (60) in expectation, again heuristically:

$$\begin{aligned}
\mathbb{E}[Z_g] &= \int_a^b \mathbb{E}[\delta(g(t))|g'(t)|]\mathrm{d}t \\
&\overset{\text{(a)}}{=} \int_a^b \left(\int \mathrm{d}y \varphi_{g(t)}(y) \delta(y) \mathbb{E}[|g'(t)||g(t) = y] \right) \mathrm{d}t, \\
&= \int_a^b \varphi_{g(t)}(0) \mathbb{E}[|g'(t)||g(t) = 0] \mathrm{d}t.
\end{aligned}$$

In (a) we conditioned on $g(t) = y$, and used the law of the conditional expectation. Here, $\varphi_{g(t)}(y)$ is the density of the random variable $g(t)$, evaluated in y. While this derivation is heuristic, its result is mathematically correct and is called the *Kac-Rice formula*. We state it in what follows, keeping some technicalities under the rug.

Theorem 8.2 (*Kac-Rice, 1D, informal*)

Let $g : [a, b] \to \mathbb{R}$ be a random function. Assume that (almost surely) for all $x \in [a, b]$, $g(x) = 0 \Rightarrow g'(x) \neq 0$, and that g is "regular enough": Then

$$\mathbb{E}[Z_g] = \int_a^b \varphi_{g(t)}(0) \mathbb{E}[|g'(t)||g(t) = 0] \mathrm{d}t.$$

Remark – The rigorous derivation of the Kac-Rice formula is quite involved, as one must take the limit $\varepsilon \to 0$ in Kac's counting formula, after taking expectations. Some regularity conditions are needed there, such as e.g. the existence and continuity of the density $\varphi_{g(t)}(u)$ close to $u = 0$. These conditions are not detailed here, but they can be found (alongside a proof of the Kac-Rice formula and of its generalization to the multidimensional setting) in the textbooks [AT09; AW09]. In practice, for our choice of Gaussian random functions these regularity conditions will be verified, and we will take them for granted.

8.3.2 The Kac-Rice formula for zeros of random functions in arbitrary dimension

The Kac-Rice formula can be generalized to arbitrary dimensions in a very natural fashion. One still needs to assume that the function g has almost surely isolated zeros, see the remark below Theorem 8.2, as well as similar technical regularity assumptions. It can also be generalized to the case of a function on a smooth manifold: as we will consider in what follows the unit sphere $\mathcal{S}^{d-1}$, it is useful to state this generalization. Again, a detailed proof can be found in [AT09; AW09][11].

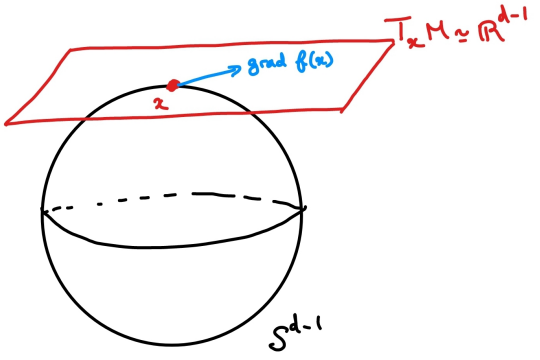
Theorem 8.3 (*Kac-Rice formula for zeros, informal*)

Let $\mathcal{M}$ be a smooth compact Riemannian manifold of dimension p, with volume measure $\mu_{\mathcal{M}}$. Let $g : \mathcal{M} \to \mathbb{R}^p$ a random function such that (almost surely) for all $x \in \mathcal{M}$, $g(x) = 0 \Rightarrow \det[Dg(x)] \neq 0$, where $Dg(x)$ is the Riemannian derivative of g in x. Assume further that g satisfies some technical regularity conditions. Then

$$\mathbb{E}[Z_g] = \int_{\mathcal{M}} \mathrm{d}\mu_{\mathcal{M}}(x)\varphi_{g(x)}(0)\mathbb{E}[|\det Dg(x)||g(x) = 0].$$

8.3.3 The Kac-Rice formula for critical points of random functions

Let us come back to our original motivation: counting the critical points of a random function $f : \mathcal{M} \to \mathbb{R}$, defined on a smooth compact manifold $\mathcal{M}$. We denote grad and Hess the Riemannian gradient and Hessian on $\mathcal{M}$. Recall that $\mathrm{grad} f(x) \in T_x\mathcal{M}$ lives in the tangent space, and $\mathrm{Hess} f(x)$ is a linear map on this tangent space. If $\mathcal{M}$ has dimension p, then one can identify $T_x\mathcal{M}$ with $\mathbb{R}^p$. In all our applications we will take $\mathcal{M} = \mathcal{S}^{d-1}$, and we will detail below how to compute the Riemannian derivatives in this case.



Denote $\varphi_x(0)$ the density of $\mathrm{grad} f(x)$ with respect to the Lebesgue measure on $\mathbb{R}^p$, taken at 0. One can apply the Kac-Rice formula (Theorem 8.3) to the gradient of the function, and obtain the following.

$$\mathbb{E}[\mathrm{Crit}_f] = \int_{\mathcal{M}} \mathrm{d}\mu_{\mathcal{M}}(x)\varphi_x(0)\mathbb{E}[|\det \mathrm{Hess} f(x)||\mathrm{grad} f(x) = 0]. \tag{61}$$

Here Crit_f is the number of critical points of f, i.e. of points such that $\mathrm{grad} f(x) = 0$. Importantly, in order to apply the Kac-Rice formula to $\mathrm{grad} f$, we had to assume that $\mathrm{grad} f(x) = 0 \Rightarrow \det[\mathrm{Hess} f(x)] \neq 0$, i.e. that the critical points of f are non-degenerate.

We can further generalize the formula to counting only certain types of critical points:

(*i*) We can impose that $f(x) \in B$, for a fixed set $B \subseteq \mathbb{R}$.

(*ii*) We can impose that $i(\mathrm{Hess} f(x)) = k$, where the *index* $i(M)$ of a symmetric matrix M is the number of negative eigenvalues of M. This characterizes the "saddleness" of the critical point. For instance, local minima have index $i(\mathrm{Hess} f(x)) = 0$.

[11] It is now based on a generalization of Kac's counting formula in higher dimensions, which is called the *area formula*.

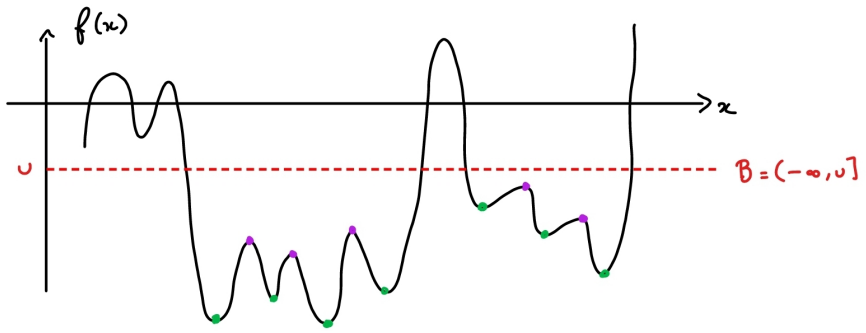


Figure 9: We count only critical points which are "low enough" in the landscape (i.e. $f(x) \leq u$), and we separate between local minima (green) and saddle points (purple).

If we impose these constraints when counting critical points we are led to define (this discrete sum is well-defined since f having non-degenerate critical points implies that there is a finite number of them, see Remark (R1) below)

$$\mathrm{Crit}_f(k, B) := \sum_{\mathrm{grad} f(x)=0} \mathbb{1}\{f(x) \in B \text{ and } i(\mathrm{Hess} f(x)) = k\}. \tag{62}$$

We then obtain the following natural generalization of the Kac-Rice formula.

Corollary 8.4 (*Kac-Rice formula for critical points, informal*)

Let $\mathcal{M}$ be a smooth compact Riemannian manifold of dimension p, with volume measure $\mu_{\mathcal{M}}$. Let $f : \mathcal{M} \to \mathbb{R}$ be a random function such that (almost surely) all the critical points of f are non-degenerate. Assume further that f satisfies some technical regularity conditions. Denote $\varphi_x(0)$ the density of $\mathrm{grad} f(x)$ with respect to the Lebesgue measure on $\mathbb{R}^p$, taken at 0. Then:

$$\mathbb{E}[\mathrm{Crit}_f(k, B)] = \int_{\mathcal{M}} \mathrm{d}\mu_{\mathcal{M}}(x) \varphi_x(0) \mathbb{E}[|\det \mathrm{Hess} f(x)| \mathbb{1}\{f(x) \in B\,;\, i(\mathrm{Hess} f(x)) = k\} | \mathrm{grad} f(x) = 0].$$

8.3.4 Remarks and generalizations

(R1) Smooth functions whose critical points are non-degenerate (i.e. for any x, $\mathrm{grad} f(x) = 0 \Rightarrow \det[\mathrm{Hess} f(x)] \neq 0$) are called *Morse functions*. Since $\mathcal{M}$ is compact, one can deduce that the number of critical points of a Morse function is finite, see Challenge 8.3. Moreover, the numbers of critical points of different Hessian indices $i(\mathrm{Hess} H(x))$ of a Morse function are constrained by the topology of $\mathcal{M}$ by the Morse inequalities [Mil63].

(R2) In general, conditional expectations of non-Gaussian random variables are intractable, making the Kac-Rice formula effectively useless since one has to know the law of the Hessian conditioned on the gradient being zero. Beyond the technical conditions mentioned above, this is a second reason why the Kac-Rice formula has been mostly applied to Gaussian random functions.

(R3) The Kac-Rice formula transforms a random differential geometry problem into a *random matrix theory* problem, since the Hessian of f is a random matrix. The main difficulty in evaluating the Kac-Rice formula comes from the distribution of the Hessian conditioned by the gradient being zero: even for Gaussian random fields, this is in general a heavily correlated Gaussian random matrix, for which very few results exist.

(R4) The Kac-Rice formula can be generalized to compute higher moments of $\mathbb{E}[\mathrm{Crit}_f(k,B)^p]$ as well. Via Morse's theory, it can even be used to compute the average Euler characteristic of the level sets of f, see [AB13] for an example.

8.4 Application to our model

We now see how to apply Corollary 8.4 for the spherical pure p-spin model, i.e. the function of eq. (59).

Precisely, for any $B \subseteq \mathbb{R}$, we wish to compute the large-d limit of the expectation of the number of critical points and of local minima of f_p, such that[12] $f_p(\mathbf{x}) \in \sqrt{d}B$.

$$\mathrm{Crit}_{f_p}(B) \coloneqq \sum_{\mathrm{grad} f_p(\mathbf{x})=0} \mathbb{1}\{f_p(\mathbf{x}) \in \sqrt{d}B\},$$
$$\mathrm{Crit}^0_{f_p}(B) \coloneqq \sum_{\mathrm{grad} f_p(\mathbf{x})=0} \mathbb{1}\{f_p(\mathbf{x}) \in \sqrt{d}B \text{ and } \mathrm{Hess} f_p(\mathbf{x}) \succeq 0\}.$$

In what follows we focus on the study of $\mathrm{Crit}_{f_p}(B)$. We will discuss later (see Section 8.5.1) the generalization of our calculation to the study of $\mathrm{Crit}^0_{f_p}(B)$. We now apply the Kac-Rice formula[13] of Corollary 8.4:

$$\mathbb{E}\,\mathrm{Crit}_{f_p}(B) = \int_{\mathcal{S}^{d-1}} \mathrm{d}\mu(\mathbf{x}) \varphi_{\mathbf{x}}(0)\, \mathbb{E}[|\det \mathrm{Hess} f_p(\mathbf{x})| \mathbb{1}\{f_p(\mathbf{x}) \in \sqrt{d}B\} | \mathrm{grad} f_p(\mathbf{x}) = 0], \tag{63}$$

in which μ is the usual surface measure on $\mathcal{S}^{d-1}$. Keep well in mind that grad and Hess stand for the *Riemannian* gradient and Hessian on the sphere, while we will denote ∇, ∇^2 the *Euclidean* gradient and Hessian.

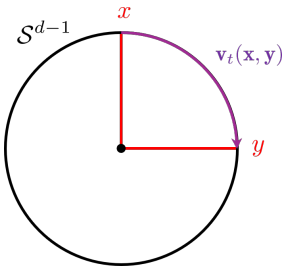
8.4.1 The joint distribution of $(f_p(\mathbf{x}), \mathrm{grad}\, f_p(\mathbf{x}), \mathrm{Hess}\, f_p(\mathbf{x}))$

Deriving the joint law of $(f_p(\mathbf{x}), \mathrm{grad}\, f_p(\mathbf{x}), \mathrm{Hess}\, f_p(\mathbf{x}))$ is a necessary first step in the Kac-Rice method, as these three random variables appear in the conditional expectation.

Gradient and Hessian on the sphere – The Riemannian gradient of a function $g : \mathcal{S}^{d-1} \to \mathbb{R}$ is simply the projection of the Euclidean gradient on the tangent plane to the sphere at $\mathbf{x}$, i.e. the hyperplane orthogonal to $\mathbf{x}$:

$$\mathrm{grad}\, g(\mathbf{x}) = P^{\perp}_{\mathbf{x}} \nabla g(\mathbf{x}), \tag{64}$$

where $P_{\mathbf{x}} \coloneqq (\mathrm{I}_d - \mathbf{x}\mathbf{x}^\top)$ is the rank-$(d-1)$ projection on $T_{\mathbf{x}}\mathcal{S}^{d-1}$. We now informally "guess" the correct expression of the Hessian $\mathrm{Hess}\, g(\mathbf{x})$. We consider $\mathbf{x} = (1, 0, \cdots, 0)$ and $\mathbf{y} = (0, 1, 0, \cdots, 0)$, and we compute the perturbative expansion as $\varepsilon \to 0$ of $g(\mathbf{v}_\varepsilon(\mathbf{x}, \mathbf{y}))$, where $(\mathbf{v}_t(\mathbf{x}, \mathbf{y}))_{0 \le t \le 1}$ is the circular arc between $\mathbf{x}$ and $\mathbf{y}$.



[12] While $\mathbb{E}[f_p(\mathbf{x})^2] = 1$ for a fixed $\mathbf{x}$, we will see that the minimal value of f_p scales as $\mathcal{O}(\sqrt{d})$.
[13] The proof that f_p is a.s. a Morse function can be found in [ABČ13].

Given the simple structure of the sphere we have

$$\mathbf{v}_t(\mathbf{x},\mathbf{y}) = \left(\sqrt{1-t^2}, t, 0, \cdots, 0\right) = \left(1 - \frac{t^2}{2}, t, 0, \cdots, 0\right) + \mathcal{O}(t^4).$$

Therefore (with $\nabla_i^k \coloneqq \partial^k/\partial x_i^k$ the Euclidean derivatives)

$$\begin{aligned} g(\mathbf{v}_\varepsilon(\mathbf{x},\mathbf{y})) &= g(\mathbf{x}) + \varepsilon\nabla_2 g(\mathbf{x}) + \frac{\varepsilon^2}{2}\nabla_2^2 g(\mathbf{x}) - \frac{\varepsilon^2}{2}\nabla_1 g(\mathbf{x}) + \mathcal{O}(\varepsilon^4), \\ &= g(\mathbf{x}) + \varepsilon\langle \mathbf{y}, \nabla g(\mathbf{x})\rangle + \frac{\varepsilon^2}{2}\mathbf{y}^\top[\nabla^2 g(\mathbf{x}) - \langle \mathbf{x}, \nabla g(\mathbf{x})\mathrm{I}_d\rangle]\mathbf{y} + \mathcal{O}(\varepsilon^4), \\ &= g(\mathbf{x}) + \varepsilon\langle \mathbf{y}, P_{\mathbf{x}}^\perp \nabla g(\mathbf{x})\rangle + \frac{\varepsilon^2}{2}\mathbf{y}^\top[P_{\mathbf{x}}^\perp \nabla^2 g(\mathbf{x}) P_{\mathbf{x}}^\perp - \langle \mathbf{x}, \nabla g(\mathbf{x})\rangle P_{\mathbf{x}}^\perp]\mathbf{y} + \mathcal{O}(\varepsilon^4). \end{aligned}$$

The final result is that the Riemannian Hessian on the sphere is:

$$\operatorname{Hess} g(\mathbf{x}) = P_{\mathbf{x}}^\perp \nabla^2 g(\mathbf{x}) P_{\mathbf{x}}^\perp - \langle \mathbf{x}, \nabla g(\mathbf{x})\rangle P_{\mathbf{x}}^\perp. \tag{65}$$

Challenge 8.4. *Using tools of differential geometry, prove eqs. (64) and (65).*

Remark – Notice that the "correction" $-\langle \mathbf{x}, \nabla g(\mathbf{x})\rangle P_{\mathbf{x}}^\perp$ is necessary: for instance if $g(\mathbf{x}) \coloneqq \|\mathbf{x}\|_2^2 - 1$, then $g(\mathbf{x}) = 0$ on the sphere, but $\nabla^2 g(\mathbf{x}) = 2\mathrm{I}_d$. This corrective term is due to the curvature of the sphere.

Computation in our case – For the function of eq. (59), we can prove:

Proposition 8.5 (*Joint law of the function, gradient, and Hessian*)

For any $\mathbf{x} \in \mathcal{S}^{d-1}$, identifying $T_{\mathbf{x}}\mathcal{S}^{d-1} \simeq \mathbb{R}^{d-1}$, the joint law of $(f_p(\mathbf{x}), \operatorname{grad} f_p(\mathbf{x}), \operatorname{Hess} f_p(\mathbf{x}))$ is given by:

$$\begin{cases} f_p(\mathbf{x}) & \overset{\mathrm{d}}{=} Z, \\ \operatorname{grad} f_{n,p}(\mathbf{x}) & \overset{\mathrm{d}}{=} \sqrt{p}\mathbf{g}, \\ \operatorname{Hess} f_{n,p}(\mathbf{x}) & \overset{\mathrm{d}}{=} \sqrt{p(p-1)(d-1)}\mathbf{W} - pZ\,\mathrm{I}_{d-1}, \end{cases}$$

where $(Z, \mathbf{g}, \mathbf{W})$ are independent random variables, with $Z \sim \mathcal{N}(0,1)$, $\mathbf{g} \sim \mathcal{N}(0, \mathrm{I}_{d-1})$, and[14] $\mathbf{W} \sim \mathrm{GOE}(d-1)$.

Proof of Proposition 8.5 – Recall

$$f_p(\mathbf{x}) = \sum_{1 \le i_1, \cdots, i_p \le d} J_{i_1 \cdots i_p} x_{i_1} \cdots x_{i_p} \qquad (\mathbf{x} \in \mathcal{S}^{d-1}), \tag{66}$$

with $J_{i_1 \cdots i_p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. f_p has the following *rotation-invariance property*: for any orthogonal matrix $\mathbf{O} \in \mathcal{O}(d)$ and any $\mathbf{x}, \mathbf{y} \in \mathcal{S}^{d-1}$,

$$(f_p(\mathbf{x}), f_p(\mathbf{y})) \overset{\mathrm{d}}{=} (f_p(\mathbf{O}\mathbf{x}), f_p(\mathbf{O}\mathbf{y})).$$

This implies that, to prove Proposition 8.5, we can assume without loss of generality that $\mathbf{x} = (1, \cdots, 0)$.

Clearly $(f_p(\mathbf{x}), \nabla f_p(\mathbf{x}), \nabla^2 f_p(\mathbf{x}))$ are jointly Gaussian random variables (as sums of Gaussian random variables), and therefore so are $(f_p(\mathbf{x}), \operatorname{grad} f_p(\mathbf{x}), \operatorname{Hess} f_p(\mathbf{x}))$. It is

[14] Recall that we defined the $\mathrm{GOE}(d)$ distribution in eq. (21).

thus enough to compute their means and covariance to fully characterize their law. From eq. (66) we have for any $\mathbf{y}$:

$$\begin{cases} f_p(\mathbf{y}) & = \sum_{1\le i_1,\cdots,i_p\le d} J_{i_1\cdots i_p} \prod_{l=1}^{p} y_{i_l}, \\ \nabla_i f_p(\mathbf{y}) & = \sum_{k=1}^{p} \sum_{1\le i_1,\cdots,i_p\le d} \mathbb{1}[i_k = i]\, J_{i_1\cdots i_p} \prod_{l(\neq k)} y_{i_l}, \\ \nabla^2_{ij} f_p(\mathbf{y}) & = \sum_{\substack{k,l=1\\k\neq l}}^{p} \sum_{1\le i_1,\cdots,i_p\le d} \mathbb{1}[i_k = i]\,\mathbb{1}[i_l = j]\, J_{i_1\cdots i_p} \prod_{\substack{a=1\\a\notin\{k,l\}}}^{p} y_{i_a}. \end{cases}$$

Specializing to $\mathbf{y} = \mathbf{x} = (1, 0, \cdots, 0)$, we get, for any $i, j \ge 1$:

$$\begin{cases} f_p(\mathbf{x}) & = J_{1\cdots 1}, \\ \nabla_i f_p(\mathbf{x}) & = \sum_{k=1}^{p} \sum_{1\le i_1,\cdots,i_p\le d} \mathbb{1}[i_k = i] \prod_{a(\neq k)} \mathbb{1}[i_a = 1]\, J_{i_1\cdots i_p}, \\ & = J_{i1\cdots 1} + J_{1i1\cdots 1} + \cdots + J_{1\cdots 1i}, \\ \nabla^2_{ij} f_p(\mathbf{x}) & = \sum_{\substack{k,l=1\\k\neq l}}^{p} \sum_{1\le i_1,\cdots,i_p\le d} \mathbb{1}[i_k = i]\,\mathbb{1}[i_l = j] \prod_{a\notin\{k,l\}} \mathbb{1}[i_a = 1]\, J_{i_1\cdots i_p}, \\ & = J_{ij1\cdots 1} + J_{i1j\cdots 1} + \cdots + J_{1\cdots 1ji}, \end{cases} \tag{67}$$

For any $i, j \ge 2$ we find from eq. (67) (denoting $\nabla_i f$ for $\nabla_i f_p(\mathbf{x})$ to lighten notations):

$$\begin{cases} \mathbb{E}[f] = \mathbb{E}[\nabla_i f] = \mathbb{E}[\nabla^2_{ij} f] = 0, \\ \mathbb{E}[f^2] = 1, \\ \mathbb{E}[(\nabla_i f)(\nabla_j f)] = p\delta_{ij}, \\ \mathbb{E}[f(\nabla_i f)] = 0. \end{cases} \tag{68}$$

Similar computations give the covariance of the second derivative, for $2 \le i < j$ and $2 \le k < l$:

$$\begin{cases} \mathbb{E}[(\nabla^2_{ij} f)(\nabla^2_{kl} f)] = p(p-1)\delta_{ik}\delta_{jl}, \\ \mathbb{E}[(\nabla^2_{ii} f)(\nabla^2_{kl} f)] = 0, \\ \mathbb{E}[(\nabla^2_{ii} f)(\nabla^2_{kk} f)] = 2p(p-1)\delta_{ik}, \\ \mathbb{E}[f(\nabla^2_{ij} f)] = \mathbb{E}[f(\nabla^2_{ii} f)] = 0, \\ \mathbb{E}[(\nabla_k f)(\nabla^2_{ij} f)] = \mathbb{E}[(\nabla_k f)(\nabla^2_{ii} f)] = 0. \end{cases} \tag{69}$$

Since $\mathbf{x} = (1, 0, \cdots, 0)$, $P^{\perp}_{\mathbf{x}}$ is simply the projection to the last $d-1$ coordinates. So eq. (68) and eq. (69) imply that:

$$\begin{cases} f & \overset{\mathrm{d}}{=} Z, \\ P^{\perp}_{\mathbf{x}} \nabla f & \overset{\mathrm{d}}{=} \sqrt{p}\mathbf{g}, \\ P^{\perp}_{\mathbf{x}} \nabla^2 f P^{\perp}_{\mathbf{x}} & \overset{\mathrm{d}}{=} \sqrt{p(p-1)(d-1)}\mathbf{W}, \end{cases} \tag{70}$$

where $Z \sim \mathcal{N}(0,1)$, $\mathbf{g} \sim \mathcal{N}(0, \mathrm{I}_{d-1})$, $\mathbf{W} \sim \mathrm{GOE}(d-1)$, and all of them are independent. Further, $\langle \nabla f, \mathbf{x} \rangle = \nabla_1 f(\mathbf{x}) = p f_p(\mathbf{x})$: applying eqs. (64) and (65) end the proof. □

8.4.2 Plugging the distribution in the Kac-Rice formula

Let us make the following important remarks regarding what we proved in Proposition 8.5.

- The joint distribution of $(f_p(\mathbf{x}), \operatorname{grad} f_p(\mathbf{x}), \operatorname{Hess} f_p(\mathbf{x}))$ does not depend on the value of $\mathbf{x}$.
- The variables $(f_p(\mathbf{x}), \operatorname{Hess} f_p(\mathbf{x}))$ are independent from $\operatorname{grad} f_p(\mathbf{x})$.
- From the gradient distribution, one easily obtains its density evaluated in 0:

$$\varphi_{\mathbf{x}}(0) = e^{-\frac{d-1}{2}\log(2\pi p)}. \tag{71}$$

Using these properties, and the fact that the distribution of Proposition 8.5 is independent of the value of $\mathbf{x}$, we can simplify eq. (63) using $\mathbf{x} = (1, 0, \cdots, 0)$:

$$\mathbb{E}\operatorname{Crit}_{f_p}(B) = V(\mathcal{S}^{d-1})\varphi_{\mathbf{x}}(0)\,\mathbb{E}[|\det \operatorname{Hess} f_p(\mathbf{x})|\mathbb{1}\{f_p(\mathbf{x}) \in \sqrt{d}B\}], \tag{72}$$

where $V(\mathcal{S}^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$ is the volume of the unit sphere. In the end we get, with the same notations as in Proposition 8.5:

$$\mathbb{E}\operatorname{Crit}_{f_p}(B) = \frac{2\pi^{d/2}}{\Gamma(d/2)} e^{\frac{d-1}{2}\log\frac{(d-1)(p-1)}{2\pi}}\,\mathbb{E}\left[\left|\det\left(\mathbf{W} - \sqrt{\frac{p}{(d-1)(p-1)}}Z\mathrm{I}_{d-1}\right)\right|\,\mathbb{1}\{Z \in \sqrt{d}B\}\right]. \tag{73}$$

In this last equation, the expectation is over $Z \sim \mathcal{N}(0,1)$ and $\mathbf{W} \sim \mathrm{GOE}(d-1)$. It is now completely explicit that we reduced our original random differential geometry problem (counting the number of critical points of a random function) to a random matrix theory problem.

8.4.3 The large-d limit, and determinants of GOE matrices

Our goal is to understand the large-d behavior of eq. (73). To simplify, we take $B = [a, b]$ to be a real interval. We can then rewrite eq. (73) making explicit the expectation over Z, and changing variables $Z := z\sqrt{d}$:

$$\begin{aligned}&\mathbb{E}\operatorname{Crit}_{f_p}(B) \\ &= \frac{2\pi^{d/2}}{\Gamma(d/2)} e^{\frac{d-1}{2}\log\frac{(d-1)(p-1)}{2\pi}}\sqrt{\frac{d}{2\pi}}\int_a^b \mathrm{d}z\, e^{-\frac{dz^2}{2}}\,\mathbb{E}_{\mathbf{W}}\left[\left|\det\left(\mathbf{W} - \sqrt{\frac{pd}{(d-1)(p-1)}}z\mathrm{I}_{d-1}\right)\right|\right].\end{aligned} \tag{74}$$

From eq. (74) we can get a precise intuition of the large-d behavior of $\mathbb{E}\operatorname{Crit}_{f_p}(B)$. In what follows we sketch this informally: a rigorous proof is quite technical, and was done first in [ABČ13]. The first term of eq. (74) can be tackled easily using Stirling's formula for the Γ function.

Challenge 8.5. *Show that*

$$\lim_{d\to\infty}\frac{1}{d}\log\left[\frac{2\pi^{d/2}}{\Gamma(d/2)} e^{\frac{d-1}{2}\log\frac{(d-1)(p-1)}{2\pi}}\sqrt{\frac{d}{2\pi}}\right] = \frac{1+\log(p-1)}{2}. \tag{75}$$

We must now handle the integral in eq. (74). Its form is very reminiscent of the Laplace method: indeed, the first term is clearly exponentially large in d, and the main difficulty is in controlling the second term, which is of the type $\mathbb{E}[|\det(\mathbf{W} - t\mathrm{I}_{d-1})|]$, for some $t \in \mathbb{R}$.

What is a proper guess for the large-d behavior of such a term? Letting $r \coloneqq (d-1)$, if we denote $(\lambda_i)_{i=1}^r$ the eigenvalues of $\mathbf{W}$, and $\hat{\mu}_{\mathbf{W}} \coloneqq (1/r)\sum_i \delta_{\lambda_i}$ its empirical spectral distribution, we have

$$\mathbb{E}[|\det(\mathbf{W} - t\mathrm{I}_r)|] = \mathbb{E}\exp\left[\sum_{i=1}^{r} \log|\lambda_i - t|\right] = \mathbb{E}\exp\left[d\int \hat{\mu}_{\mathbf{W}}(\mathrm{d}x)\log|x-t|\right]. \tag{76}$$

We have seen the classical result (cf. Section 5, specifically eq. (22)) that $\hat{\mu}_{\mathbf{W}}$ converges (in the weak topology) a.s. to the semicircle distribution

$$\mu_{\text{s.c.}}(x) \coloneqq \frac{\sqrt{4-x^2}}{2\pi}\mathbb{1}\{|x| \le 2\}.$$

This observation can lead to the "wild" conjecture:

$$\mathbb{E}[|\det(\mathbf{W} - t\mathrm{I}_r)|] \overset{(?)}{=} \exp\left[d\int \mu_{\text{s.c.}}(\mathrm{d}x)\log|x-t| + o(d)\right]. \tag{77}$$

This is a very wild guess for two reasons:

(*i*) The function $x \mapsto \log|x-t|$ is not a bounded Lipschitz function, so weak convergence is *a priori* not enough to deduce convergence of the average of this function.

(*ii*) Much more importantly, there is an expectation outside the exponential in eq. (76)! And if X_d is a random variable, even if $X_d \to x$ as $d \to \infty$ (for x a real value), we might have

$$\lim_{d\to\infty} \frac{1}{d}\log \mathbb{E}[\exp(dX_d)] \neq x. \tag{78}$$

Challenge 8.6. *Find an example for eq. (78). Then, assuming $|X_d| \le M$ is bounded:*

(*a*) *Is there a clear bound between the left and right-hand sides of eq. (78)?*

(*b*) *Show that a sufficient condition for eq. (78) to be an equality is that for all $t > 0$:*

$$\lim_{d\to\infty} \frac{1}{d}\log \mathbb{P}[|X_d - x| \ge t] = -\infty. \tag{79}$$

Eq. (79) is called a *large deviations* upper bound: informally it is a very strong form of concentration, as events where X_d differ from x by a $\mathcal{O}(1)$ quantity have probability $\exp(-\omega(d))$.

Despite our wild guess, it turns out that eq. (77) is correct! The mathematical reason behind this result is the fact that the empirical spectral distribution $\hat{\mu}_{\mathbf{W}}$ actually concentrates to $\mu_{\text{s.c.}}$ even faster than what is required by eq. (79): it satisfies a large deviation principle in the scale d^2, meaning that typically

$$\lim_{d\to\infty} \frac{1}{d^2}\log \mathbb{P}\left[\left|\int \hat{\mu}_{\mathbf{W}}(\mathrm{d}x)\varphi(x) - \int \mu_{\text{s.c.}}(\mathrm{d}x)\varphi(x)\right| \ge t\right] = -I_\varphi(t), \tag{80}$$

for some finite $I_\varphi(t) > 0$. This was proven in [BG97], and I invite you to have a look at the ICM notes of A. Guionnet [Gui22] if you want to learn more about the (beautiful) theory of large deviations in random matrices. We finally note that the fact that the right-hand side of eq. (77) is an upper bound to the left-hand side can be proven using only the Hoffman-Wielandt inequality and Lipschitz concentration, see [Sel24a] (Lectures 6 and 7) if you want to see this proof!

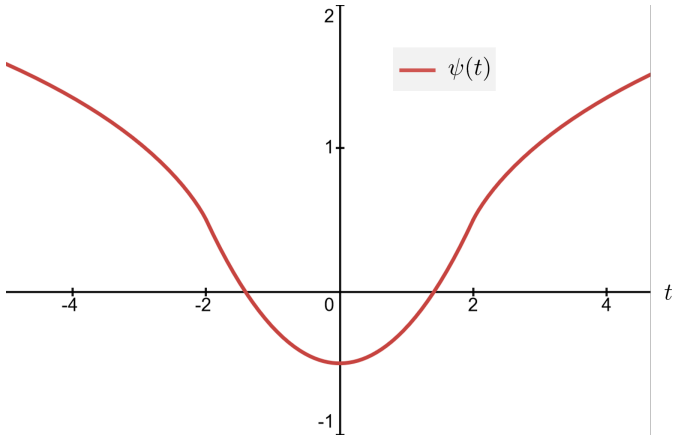
Coming back to eq. (74), using eq. (75) and the result of eq. (77), one can apply the Laplace method to obtain:

$$\lim_{d\to\infty}\frac{1}{d}\log\mathbb{E}\,\mathrm{Crit}_{f_p}([a,b]) = \max_{E\in(a,b)}\left[\frac{1+\log(p-1)}{2} - \frac{E^2}{2} + \int\mu_{\text{s.c.}}(\mathrm{d}x)\log\left|x - E\sqrt{\frac{p}{p-1}}\right|\right]. \tag{81}$$

It turns out that the integral over the semicircular density can be tackled with the residue theorem: we essentially computed its derivative already in Challenge 5.5!

Challenge 8.7. *Prove that for any $t\in\mathbb{R}$:*

$$\int\mu_{\text{s.c.}}(\mathrm{d}x)\log|x-t| = \frac{t^2}{4} - \frac{1}{2} + \mathbb{1}\{|t|\geq 2\}\left[-\frac{|t|\sqrt{t^2-4}}{4} + \log\frac{\sqrt{t^2-4}+|t|}{2}\right] =: \psi(t). \tag{82}$$



8.4.4 Discussion

Combining everything, we can very reasonably conjecture the following result, which is a theorem due to [ABČ13]:

Theorem 8.6 (*"Annealed" complexity of the pure spherical p-spin model [ABČ13]*)

For any $p\geq 2$, and any $-\infty\leq a<b\leq+\infty$:

$$\lim_{d\to\infty}\frac{1}{d}\log\mathbb{E}\,\mathrm{Crit}_{f_p}([a,b]) = \max_{E\in[a,b]} I_p(E),$$

where

$$I_p(E) := \frac{1+\log(p-1)}{2} - \frac{E^2}{2} + \psi\left(E\sqrt{\frac{p}{p-1}}\right). \tag{83}$$

Note that, from our calculation, the variable E actually represents the value $f_p(\mathbf{x})/\sqrt{d}$ (the "energy" value) at critical points of the function. We can already make some remarks concerning Theorem 8.6.

The case $p=2$ – For $p=2$, we have $f_2(\mathbf{x})/\sqrt{d} = \mathbf{x}^\top\mathbf{W}\mathbf{x}/\sqrt{2}$, where $\mathbf{W}\sim\mathrm{GOE}(d)$. We have seen (cf. Challenge 8.1) that the critical points of f_2 correspond to the eigenvectors of $\mathbf{W}$. On the other hand, we can show that $I_2(E)=0$ for all $|E|\leq E_0(2) := \sqrt{2}$, and $I_2(E)<0$ for all $|E|>E_0(2)$. Note that $E_0(2)$ corresponds well to the largest eigenvalue of the matrix $\mathbf{W}/\sqrt{2}$! Further, for all $|E|\leq\sqrt{2}$ there are eigenvalues of $\mathbf{W}/\sqrt{2}$ that approach E as $d\to\infty$: this explains why $I_2(E)=0$ for these values.

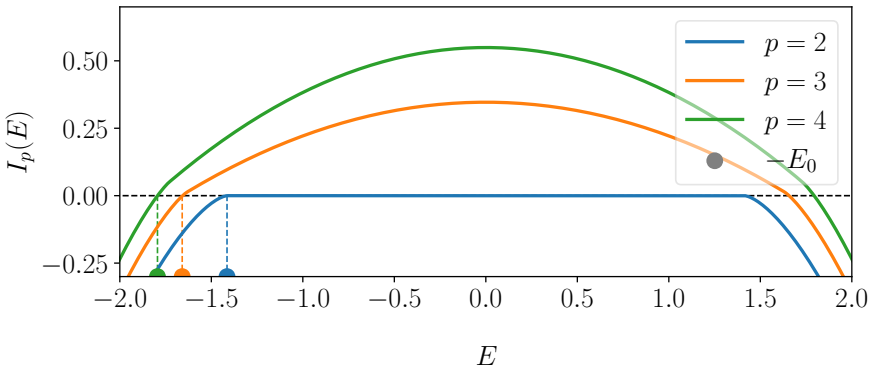


Figure 10: The function $I_p(E)$ from Theorem 8.6 for different values of p. We show the value $-E_0(p)$ such that $I_p(-E_0) = 0$.

The ground state energy – It is particularly interesting to consider the range of E for which $I_p(E) < 0$. Indeed, if we define $E_0(p) \geq 0$ by $I_p(-E_0) = 0$, then for any $\varepsilon > 0$:

$$\lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}\, \mathrm{Crit}_{f_p}((-\infty, -E_0 - \varepsilon]) < 0.$$

Since the expected number of critical points is exponentially small, by Markov's inequality:

$$\lim_{d\to\infty} \mathbb{P}[\mathrm{Crit}_{f_p}((-\infty, -E_0 - \varepsilon]) > 0] = 0.$$

Because f_p is a smooth function on the compact set $\mathcal{S}^{d-1}$, this implies that for any $\varepsilon > 0$, with high probability:

$$\frac{1}{\sqrt{d}} \min_{\mathbf{x}\in\mathcal{S}^{d-1}} f_p(\mathbf{x}) \geq -E_0(p) - \varepsilon.$$

8.5 Generalizations and open directions

8.5.1 Counting local minima

Theorem 8.6 gives a very detailed understanding of the average number of critical points of f_p. However, for local optimization procedures, the most important type of critical points are *local minima*, as the algorithms might get stuck in these points.

The Kac-Rice formula (Corollary 8.4) can handle counting only local minima: we reach then directly the equivalent to eq. (74):

$$\mathbb{E}\, \mathrm{Crit}^0_{f_p}(B) = \frac{2\pi^{d/2}}{\Gamma(d/2)} e^{\frac{d-1}{2}\log\frac{(d-1)(p-1)}{2\pi}} \sqrt{\frac{d}{2\pi}} \tag{84}$$
$$\times \int_a^b \mathrm{d}z\, e^{-\frac{dz^2}{2}} \mathbb{E}_{\mathbf{W}} \left[\left| \det\left(\mathbf{W} - \sqrt{\frac{pd}{(d-1)(p-1)}} z \mathrm{I}_{d-1} \right) \right| \mathbb{1}\left\{ \lambda_{\min}(\mathbf{W}) \geq \sqrt{\frac{pd}{(d-1)(p-1)}} z \right\} \right].$$

Therefore, in order to count local minima, we need to control (with $r := d-1$) quantities of the type

$$\Phi(t) := \lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}[|\det(\mathbf{W}) - t\mathrm{I}_r| \mathbb{1}\{\lambda_{\min}(\mathbf{W}) \geq t\}], \tag{85}$$

for $t \in \mathbb{R}$. We argued above that the determinant term concentrates extremely quickly (the large deviations are in the scale $\exp(-\Theta(d^2))$) on its limit, given by the semicircle

distribution. The same argument can be applied here, and a natural conjecture is thus

$$\Phi(t) = \int \mu_{\text{s.c.}}(\mathrm{d}x) \log|x-t| + \lim_{d\to\infty} \frac{1}{d} \log \mathbb{P}\{\lambda_{\min}(\mathbf{W}) \geq t\}. \tag{86}$$

We see appearing in eq. (86) the *large deviations* of the smallest eigenvalue of a GOE matrix, i.e. regimes in which it can be very atypical: we illustrate it in Fig. 11. The fact that these large deviations appear in our computation is not such a surprise: we wrote the Kac-Rice formula for the number of *local minima* of the function f_p, i.e. critical points with positive Hessian matrix, and in order to condition on this positivity we have to understand the law of the smallest eigenvalue of the Hessian.

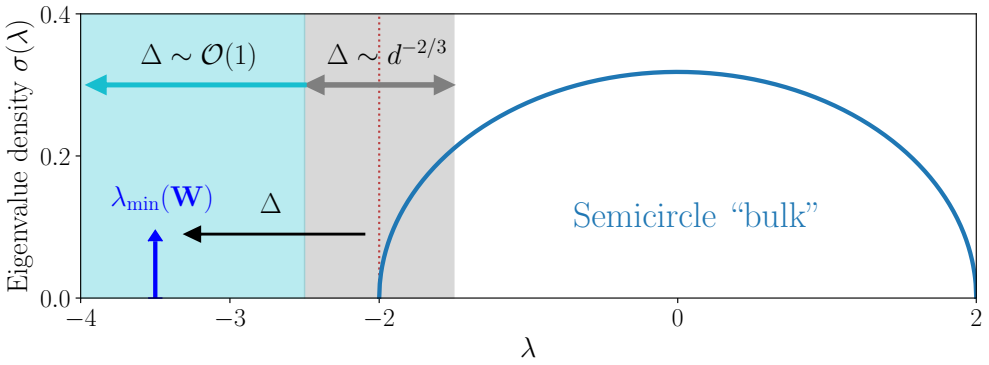


Figure 11: Illustration of a *large deviations* event in which the smallest eigenvalue of a random matrix $\mathbf{W} \sim \text{GOE}(d)$ is macroscopically far from its expected value -2. Δ is the shift of $\lambda_{\max}$ from its expected value, at the left edge of the bulk. Note that the smallest eigenvalue typically fluctuates in the scale $d^{-2/3}$, as shown in the grey area (these fluctuations are connected to the Tracy-Widom law [TW94]), but large deviations instead correspond to *macroscopic* fluctuations, which are exponentially rare in d (cyan area).

The large deviations of $\lambda_{\min}(\mathbf{W})$ have been studied and are well understood [BDG01].

Lemma 8.7 (*Large deviations of* $\lambda_{\min}(\mathbf{W})$)

Let $\mathbf{W} \sim \text{GOE}(d)$. Then, for any $a < b$:

$$\lim_{d\to\infty} \frac{1}{d} \log \mathbb{P}\{\lambda_{\min}(\mathbf{W}) \in [a,b]\} = -\inf_{t\in[a,b]} \Xi(t),$$

where the so-called *rate function* $\Xi(t)$ is given by:

$$\Xi(t) := \begin{cases} \frac{1}{2}\int_2^{-t} \mathrm{d}z \sqrt{z^2-4} = \frac{-t\sqrt{t^2-4}}{4} + \log\frac{-t-\sqrt{t^2-4}}{2} & \text{if } t \leq -2, \\ +\infty & \text{if } t > -2. \end{cases}. \tag{87}$$

In particular, it is clear that

$$\lim_{d\to\infty} -\frac{1}{d} \log \mathbb{P}\{\lambda_{\min}(\mathbf{W}) \geq t\} = J(t) := \begin{cases} 0 & \text{if } t \leq -2, \\ +\infty & \text{if } t > -2. \end{cases}. \tag{88}$$

Using eq. (88), eq. (86), and the Laplace method in eq. (84), we reach that

$$\lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}\,\text{Crit}^0_{f_p}([a,b]) = \max_{E\in[a,b]} \left[\frac{1+\log(p-1)}{2} - \frac{E^2}{2} + \psi\left(E\sqrt{\frac{p}{p-1}}\right) - J\left(E\sqrt{\frac{p}{p-1}}\right)\right].$$

We reach the following theorem, again proven in [ABČ13].

Theorem 8.8 (*"Annealed" complexity of local minima [ABČ13]*)

For any $p \geq 2$, and any $-\infty \leq a < b \leq +\infty$:

$$\lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}\, \mathrm{Crit}^0_{f_p}([a,b]) = \max_{E\in[a,b]} I^0_p(E),$$

where (with ψ defined in eq. (82)):

$$I^0_p(E) := \begin{cases} -\infty & \text{if } E > -2\sqrt{\dfrac{p-1}{p}}, \\ \dfrac{1+\log(p-1)}{2} - \dfrac{E^2}{2} + \psi\left(E\sqrt{\dfrac{p}{p-1}}\right) & \text{if } E \leq -2\sqrt{\dfrac{p-1}{p}}. \end{cases} \tag{89}$$

The calculation we sketched for local minima can be generalized to critical points of any fixed index $k \in \mathbb{N}$, and if we define:

$$\mathrm{Crit}^k_{f_p}(B) := \sum_{\mathrm{grad} f_p(\mathbf{x})=0} \mathbb{1}\{f_p(\mathbf{x}) \in \sqrt{d}B \text{ and } i(\mathrm{Hess} f_p(\mathbf{x})) = k\},$$

and

$$\begin{cases} \Theta_k(u) & := \lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}\, \mathrm{Crit}^k_{f_p}((-\infty,u]), \\ \Theta(u) & := \lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}\, \mathrm{Crit}_{f_p}((-\infty,u]), \end{cases}$$

there exists analytic expressions for the $\Theta_k(u)$ functions, given in [ABČ13]. We plot these functions for $p = 3, 4$ in Fig. 12. In particular, all the functions $\Theta_k(u)$ agree for

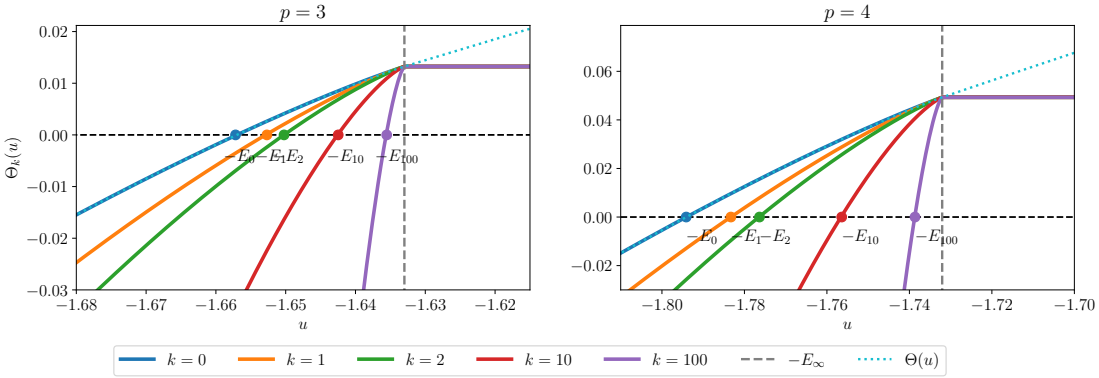


Figure 12: The functions Θ_k for $k \in 0, 1, 2, 10, 100$, for two values of p.

$u \geq -E_\infty$, which is often referred to as the *threshold* energy

$$E_\infty(p) := 2\sqrt{\frac{p-1}{p}}. \tag{90}$$

We also denote $-E_k$ the energy value at which the function $\Theta_k(u)$ becomes positive: $\Theta_k(-E_k) = 0$, cf Fig. 12. Note that the local minima always dominate the average

complexity of critical points for all energies below $-E_\infty$. For $u > -E_\infty$, we instead have $\Theta(u) > \Theta_k(u)$: one can show that there the complexity is dominated instead by critical points of index diverging with d, and that there is no longer any local minimum with these energy values (in fact this last point follows from Theorem 8.8 and Markov's inequality).

Connection to local optimization – Coming back to local optimization algorithms (think of a variant of a gradient descent algorithm: physicists often study so-called *Langevin dynamics* where one adds some amount of noise to the iterates), this picture has led physicists to conjecture for a long time that these algorithms are able to find $\mathbf{x}$ with $f_p(\mathbf{x}) \simeq -\sqrt{d}E_\infty(p)$, but that any $f_p(\mathbf{x}) \simeq \sqrt{d}u$ for $u < -E_\infty(p)$ is not reachable in polynomial time [CK93]. Indeed, while for $u > -E_\infty(p)$ all critical points have a number $\Theta(d)$ of "descent" directions, for $u < -E_\infty(p)$ such critical points have essentially disappeared, and on the other hand there is an exponential number of local minima which can trap the local optimization dynamics! This conjecture about the threshold energy $-E_\infty(p)$ being a barrier for the Langevin dynamics algorithm was recently proven in [Sel24b].

8.5.2 Concentration and the second moment method

Let us come back to the ground state energy. We have seen that the averaged complexity becoming negative allows (via Markov's inequality) to give the high-probability bound, for any $\varepsilon > 0$:

$$\frac{1}{\sqrt{d}} \min_{\mathbf{x}\in\mathcal{S}^{d-1}} f_p(\mathbf{x}) \geq -E_0(p) - \varepsilon. \tag{91}$$

Instead, for $u > -E_0(p)$, there is on average an exponentially large number of critical points with $f_d(\mathbf{x}) \simeq \sqrt{d}u$: this seems to suggest that the bound on eq. (91) might actually be an equality as d gets large.

How to get this matching lower bound? A natural way is to prove *concentration* of the number of critical points, by using the second moment method we introduced in Section 7. As we mentioned as a remark, the Kac-Rice formula can be generalized to moments, and in particular we have a formula of the type

$$\mathbb{E}[\mathrm{Crit}_{f_p}(B)^2] = \cdots, \tag{92}$$

where the right-hand side of eq. (92) involves now an integral over $\mathbf{x}$ and $\mathbf{y}$, and – for instance – the averaged value of $|\det \mathrm{Hess}\, f_p(\mathbf{x})||\det \mathrm{Hess}\, f_p(\mathbf{y})|$, conditioned on the gradients at $\mathbf{x}$ and $\mathbf{y}$ both being 0. The same ideas we sketched for the average can be generalized to the second moment, and being very careful with constants, the following result was shown in [Sub17].

Theorem 8.9 (*[Sub17]*)

For any $p \geq 3$, and any $u \in (-E_0(p), -E_\infty(p))$:

$$\lim_{d\to\infty} \frac{\mathbb{E}[\mathrm{Crit}_{f_p}((-\infty,u])^2]}{\mathbb{E}[\mathrm{Crit}_{f_p}((-\infty,u])]^2} = 1.$$

A direct corollary of Theorem 8.9 and Theorem 8.6 is that indeed $-E_0(p)$ is the minimum (or ground state) of $f_p(\mathbf{x})$.

Corollary 8.10 (*Ground state of the spherical p-spin model*)

For any $p \geq 3$,

$$\frac{1}{\sqrt{d}} \min_{\mathbf{x} \in \mathcal{S}^{d-1}} f_p(\mathbf{x}) \xrightarrow[d\to\infty]{\mathrm{p}} -E_0(p).$$

Recall that we characterized $-E_0(p)$ as the unique zero of the averaged complexity function, and the latter has a completely analytical form, see Theorem 8.6 and Fig. 10.

Proof of Corollary 8.10 – Let $X \coloneqq (1/\sqrt{d}) \min_{\mathbf{x} \in \mathcal{S}^{d-1}} f_p(\mathbf{x})$. We know that $\mathbb{P}[X \geq -E_0 - \varepsilon] = 1 - o_d(1)$, so it only remains to prove the corresponding upper bound. Notice that $X \leq u$ is equivalent to $\mathrm{Crit}((-\infty, u]) > 0$, since the global minimum must be a critical point. On the other hand, by the Paley-Zygmund inequality (Lemma 7.3), we have

$$\mathbb{P}[\mathrm{Crit}((-\infty, u]) > 0] \geq \frac{\mathbb{E}[\mathrm{Crit}((-\infty, u])^2]}{\mathbb{E}[\mathrm{Crit}((-\infty, u])]^2}.$$

The result then follows directly from Theorem 8.9. □

As a remark, an even weaker statement than Theorem 8.10, just showing that

$$\frac{\mathbb{E}[\mathrm{Crit}_{f_p}((-\infty, u])^2]}{\mathbb{E}[\mathrm{Crit}_{f_p}((-\infty, u])]^2} = e^{o(d)},$$

is already enough to deduce Corollary 8.10 if one uses classical concentration inequalities, see Appendix D of [Sub17] or the notes [Sel24a].

8.5.3 Towards more complex and realistic models

To conclude our discussion, let us mention a few generalizations and extensions of the Kac-Rice calculation that we described.

- There are simple models (for instance $f(\mathbf{x}) \coloneqq f_3(\mathbf{x}) + f_4(\mathbf{x})$), where it is known that the second moment method fails, and Theorem 8.9 does not hold. In this case, one can use the Kac-Rice formulas to compute $\mathbb{E}[(\mathrm{Crit}_f)^l]$ for all $l \geq 1$, and try to characterize the typical value of the number of critical points

  $$\Sigma_{\mathrm{qu.}} \coloneqq \lim_{d\to\infty} \frac{1}{d} \mathbb{E} \log \mathrm{Crit}_f.$$

  Indeed, if Crit_f in the scale $\exp\{\Theta(d)\}$, $(1/d) \log \mathrm{Crit}_f$ will often concentrate on its average, so $\Sigma_{\mathrm{qu.}}$ – which is called the *quenched complexity* in the physics language – really describes the typical number of critical points of the function. On the other hand, here we computed the so-called *annealed* complexity

  $$\Sigma_{\mathrm{an.}} \coloneqq \lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}\, \mathrm{Crit}_f,$$

  which might be different from the quenched one if Crit_f does not concentrate around its average. As we saw, in the pure p-spin model $f_p(\mathbf{x})$, annealed and quenched complexities are identical thanks to Theorem 8.9.

  While computing the quenched complexity from all the moments $\mathbb{E}[(\mathrm{Crit}_f)^l]$ is technically possible (since the moments characterize fully the distribution), it has not yet been done in a rigorous manner, even in Gaussian models. On the other hand, theoretical physicists have used a non-rigorous method (called the "replica method") that allows to easily obtain the quenched complexity from the integer moments: this was performed in [Ros+19] for generalizations of the p-spin model.

- Going beyond Gaussian models, there have been applications of the Kac-Rice method to empirical risk models that we introduced in Section 8.1. In [MAB20] the annealed and quenched complexity are computed for an empirical risk of the form:

$$\mathcal{R}(\mathbf{x}) \coloneqq \frac{1}{2}\sum_{i=1}^{n}[\phi(\mathbf{z}_\mu \cdot \mathbf{x}) - \phi(\mathbf{z}_\mu \cdot \mathbf{x}^\star)]^2,$$

  where $\mathbf{z}_\mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d)$. $\mathcal{R}$ can be seen as the empirical risk of a *generalized linear model*, where the statistician is given $\{y_i = \varphi(\mathbf{z}_\mu \cdot \mathbf{x}^\star)\}_{i=1}^n$, and must infer $\mathbf{x}^\star$ from these observations. The essentials of how to perform the Kac-Rice method in this setting are described in the introduction of [MAB20], and a summary can also be found in Lecture 13 of [Kun23].

Bibliographic notes

The main mathematical references for the Kac-Rice calculation in the spherical p-spin model are [AB13; ABČ13], but these calculations appeared before in the theoretical physics literature [Fyo04]. If you are interested in knowing more after this chapter, you can take a look at the recent review [RF23] (disclaimer: it is written in a theoretical physics language), or at the lecture notes of [Kun23] and [Sel24a] related to the Kac-Rice formula.

Solution to (some) challenges

Challenge 8.2

The set $g^{-1}(\{0\})$ is closed (because g is continuous), and made of isolated points since $g(x) = 0 \Rightarrow g'(x) \neq 0$, so there exists $\varepsilon_x > 0$ such that $g(x+h) \neq 0$ for any $|h| \leq \varepsilon_x$. Closed sets of isolated points in a compact set must be finite (prove it, see also Challenge 8.3), and so $g^{-1}(\{0\})$ is finite.

We order the zeros of g as $a < x_1 < \cdots < x_n < b$. Because g is $\mathcal{C}^1$, the set $C \coloneqq (g')^{-1}(\{0\})$ of critical points of g is compact, and disjoint from $g^{-1}(\{0\})$. We let

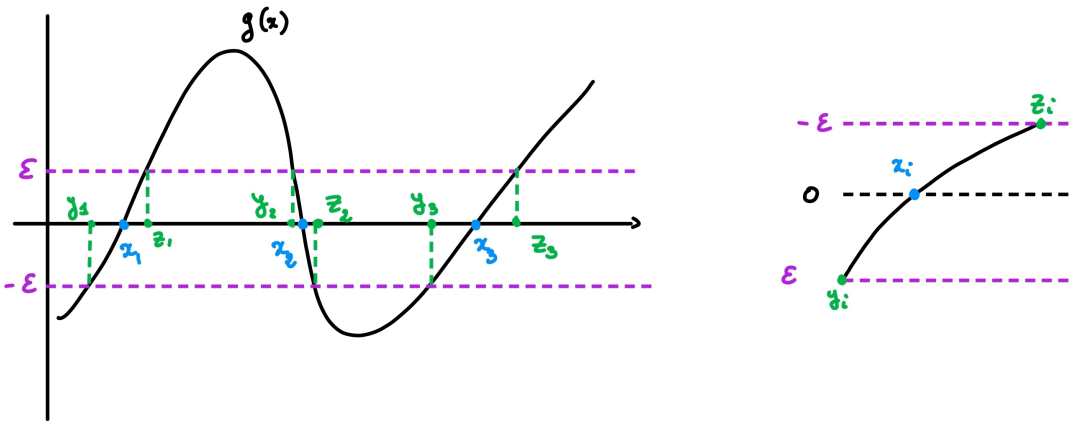
$$\varepsilon_0 \coloneqq \min_{x \in C} |g(x)| > 0.$$

Let $\varepsilon \in (\varepsilon, \varepsilon_0)$. Because $\Gamma_\varepsilon \coloneqq \{|g(x)| < \varepsilon\}$ is open, it can be expressed as of countable union of disjoint intervals $\Gamma_\varepsilon = \bigcup_{i \geq 1}(y_i, z_i)$. Moreover, we have:

(*i*) $\Gamma_\varepsilon \cap C = \emptyset$, since $\varepsilon < \varepsilon_0$. Since g' is continuous, it therefore has constant sign on any (y_i, z_i).

(*ii*) $|g(y_i)| = |g(z_i)| = \varepsilon$, by continuity of g.

(*iii*) $g(y_i)g(z_i) < 0$. Indeed, otherwise by (*ii*) we have $g(y_i) = g(z_i) \neq 0$, and Rolle's theorem would imply that there is a zero of g' in (y_i, z_i), which contradicts (*i*).



By points (i) and (iii), g has exactly one zero in each interval (y_i, z_i). So in the end we have

$$\Gamma_\varepsilon = \bigcup_{i=1}^{n} (y_i, z_i), \text{ with } y_i < x_i < z_i.$$

Finally, we have:

$$\begin{aligned}
\frac{1}{2\varepsilon}\int_a^b \mathbb{1}\{|g(x)| < \varepsilon\}|g'(x)|\mathrm{d}x &= \frac{1}{2\varepsilon}\int_{\Gamma_\varepsilon} |g'(x)|\mathrm{d}x, \\
&= \frac{1}{2\varepsilon}\sum_{i=1}^{n}\int_{y_i}^{z_i} |g'(x)|\mathrm{d}x, \\
&\overset{\text{(a)}}{=} \frac{1}{2\varepsilon}\sum_{i=1}^{n}\left|\int_{y_i}^{z_i} g'(x)\right|\mathrm{d}x, \\
&\overset{\text{(b)}}{=} \frac{1}{2\varepsilon}\sum_{i=1}^{n}(2\varepsilon), \\
&= n,
\end{aligned}$$

where we used in (a) that g' has constant sign on (y_i, z_i), and we used (ii) in (b). □

Challenge 8.6

Clearly an example would need to be a random variable X_d that possesses low-probability events where it is far from its average. An example is the following:

$$X_d = \begin{cases} 0 & \text{with probability } 1 - e^{-d}, \\ 2 & \text{with probability } e^{-d}. \end{cases}$$

Then $X_d \to 0$ as $d \to \infty$ (in probability). However

$$\mathbb{E}[e^{dX_d}] = (1 - e^{-d}) + e^d,$$

and so

$$\lim_{d\to\infty} \frac{1}{d}\log \mathbb{E}[e^{dX_d}] = 1 \neq 0.$$

Point (a) – Jensen's inequality implies that for any random variable $Y > 0$:

$$\log \mathbb{E}Y \geq \mathbb{E}\log Y.$$

In particular

$$\frac{1}{d}\log \mathbb{E}[e^{dX_d}] \geq \frac{1}{d}\mathbb{E}\log[e^{dX_d}] = \mathbb{E}[X_d].$$

So we have the general bound ($\mathbb{E}[X_d] \to x$ since $|X_d| \leq M$ and $X_d \to x$ in probability):

$$\liminf_{d\to\infty} \frac{1}{d}\log \mathbb{E}[e^{dX_d}] \geq x. \tag{93}$$

Point (b) – We assume here that eq. (79) holds. Let $\varepsilon > 0$. Notice that $|X_d - x| \leq 2M$. Thus[15]:

$$\begin{aligned}
\mathbb{E}[e^{d(X_d-x)}] &\leq e^{2Md-\omega(d)} + \mathbb{E}\left[e^{d(X_d-x)}\mathbb{1}\{|X_d - x| < \varepsilon\}\right], \\
&\leq e^{2Md-\omega(d)} + e^{d\varepsilon}\mathbb{P}[|X_d - x| < \varepsilon] \\
&\leq e^{2Md-\omega(d)} + e^{d\varepsilon}.
\end{aligned}$$

Thus we have

$$\limsup_{d\to\infty} \frac{1}{d}\log \mathbb{E}[e^{d(X_d-x)}] \leq \varepsilon.$$

Since this is true for any $\varepsilon \geq 0$, we have

$$\limsup_{d\to\infty} \frac{1}{d}\log \mathbb{E}[e^{d(X_d-x)}] \leq 0.$$

Combining this with eq. (93) we get

$$\lim_{d\to\infty} \frac{1}{d}\log \mathbb{E}[e^{d(X_d-x)}] = 0. \qquad \square$$

[15] The notation $a_d = \omega(b_d)$ means that $a_d/b_d \to \infty$ as $d \to \infty$.

References

[AB13] Antonio Auffinger and Gérard Ben Arous. "Complexity of random smooth functions on the high-dimensional sphere". In: *The Annals of Probability* (2013), pp. 4214–4247.

[ABČ13] Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. "Random matrices and complexity of spin glasses". In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 165–201.

[Ada+11] Radoslaw Adamczak et al. "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling". In: *Constructive Approximation* 34 (2011), pp. 61–88.

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. 118. Cambridge university press, 2010.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. "Finding a large hidden clique in a random graph". In: *Random Structures & Algorithms* 13.3-4 (1998), pp. 457–466.

[Alq+24] Abdulmajeed Alqasem et al. "On a conjecture of Feige for discrete log-concave distributions". In: *SIAM Journal on Discrete Mathematics* 38.1 (2024), pp. 93–102.

[AS16] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[AT09] R.J. Adler and J.E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer New York, 2009.

[AW09] Jean-Marc Azais and Mario Wschebor. *Level sets and extrema of random processes and fields*. John Wiley & Sons, 2009.

[Ban+24] Afonso S Bandeira et al. "On the concentration of Gaussian Cayley matrices". In: *Applied and Computational Harmonic Analysis* 73 (2024), p. 101694.

[Ban10] Nikhil Bansal. "Constructive algorithms for discrepancy minimization". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 3–10.

[Ban98] Wojciech Banaszczyk. "Balancing vectors and Gaussian measures of n-dimensional convex bodies". In: *Random Structures & Algorithms* 12.4 (1998), pp. 351–360.

[Bar+19] Boaz Barak et al. "A nearly tight sum-of-squares lower bound for the planted clique problem". In: *SIAM Journal on Computing* 48.2 (2019), pp. 687–735.

[BBH23] Afonso S Bandeira, March T Boedihardjo, and Ramon van Handel. "Matrix concentration inequalities and free probability". In: *Inventiones mathematicae* 234.1 (2023), pp. 419–487.

[BBP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *Annals of Probability* 33.5 (2005), pp. 1643–1697.

[BDG01] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. "Aging of spherical spin glasses". In: *Probability theory and related fields* 120 (2001), pp. 1–67.

[BE76] Béla Bollobás and Paul Erdös. "Cliques in random graphs". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 80. 3. Cambridge University Press. 1976, pp. 419–427.

[BG97] Gérard Ben Arous and Alice Guionnet. "Large deviations for Wigner's law and Voiculescu's non-commutative entropy". In: *Probability theory and related fields* 108 (1997), pp. 517–542.

[BJM23] Nikhil Bansal, Haotian Jiang, and Raghu Meka. "Resolving matrix spencer conjecture up to poly-logarithmic rank". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023, pp. 1814–1819.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BMZ22] Afonso S Bandeira, Antoine Maillard, and Nikita Zhivotovskiy. "A remark on Kashin's discrepancy argument and partial coloring in the Komlós conjecture". In: *Portugaliae Mathematica* 79.3 (2022), pp. 311–316.

[BN11] Florent Benaych-Georges and Raj Rao Nadakuditi. "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1 (2011), pp. 494–521.

[BPW18] Afonso S Bandeira, Amelia Perry, and Alexander S Wein. "Notes on computational-to-statistical gaps: predictions using statistical physics". In: *Portugaliae mathematica* 75.2 (2018), pp. 159–186.

[BR13] Quentin Berthet and Philippe Rigollet. "Complexity theoretic lower bounds for sparse principal component detection". In: *Conference on learning theory*. PMLR. 2013, pp. 1046–1066.

[BSS23] A. S. Bandeira, A. Singer, and T. Strohmer. *Mathematics of Data Science*. Book draft available here. 2023.

[CK93] Leticia F Cugliandolo and Jorge Kurchan. "Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model". In: *Physical Review Letters* 71.1 (1993), p. 173.

[CS92] Andrea Crisanti and H-J Sommers. "The spherical p-spin interaction spin glass model: the statics". In: *Zeitschrift fur Physik B Condensed Matter* 87.3 (1992), pp. 341–354.

[Dia+16] Mohamad Dia et al. "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". In: *Advances in Neural Information Processing Systems* 29 (2016).

[DM15] Yash Deshpande and Andrea Montanari. "Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time". In: *Foundations of Computational Mathematics* 15 (2015), pp. 1069–1128.

[Fei06] Uriel Feige. "On sums of independent random variables with unbounded variance and estimating the average degree in a graph". In: *SIAM Journal on Computing* 35.4 (2006), pp. 964–984.

[Fel+17] Vitaly Feldman et al. "Statistical algorithms and a lower bound for detecting planted cliques". In: *Journal of the ACM (JACM)* 64.2 (2017), pp. 1–37.

[Fyo04] Yan V Fyodorov. "Complexity of Random Energy Landscapes, Glass Transition, and Absolute Value of the Spectral Determinant of Random Matrices". In: *Physical review letters* 92.24 (2004), p. 240601.

[Gar20] Brian Garnett. "Small deviations of sums of independent random variables". In: *Journal of Combinatorial Theory, Series A* 169 (2020), p. 105119.

[GB23] Cédric Gerbelot and Raphaël Berthier. “Graph-based approximate message passing iterations”. In: *Information and Inference: A Journal of the IMA* 12.4 (2023), pp. 2562–2628.

[Glu89] Efim Davydovich Gluskin. “Extremal properties of orthogonal parallelepipeds and their applications to the geometry of Banach spaces”. In: *Mathematics of the USSR-Sbornik* 64.1 (1989), p. 85.

[GM75] Geoffrey R Grimmett and Colin JH McDiarmid. “On colouring random graphs”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 77. Cambridge University Press. 1975, pp. 313–324.

[GMZ22] David Gamarnik, Cristopher Moore, and Lenka Zdeborová. “Disordered systems insights on computational hardness”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2022.11 (2022), p. 114015.

[Gui22] Alice Guionnet. “Rare events in random matrix theory”. In: *Plenary LectureS* 2 (2022), pp. 1008–1052.

[Guo+20] Jiayi Guo et al. “Bounding probability of small deviation on sum of independent random variables: Combination of moment approach and Berry-Esseen theorem”. In: *arXiv preprint arXiv:2003.03197* (2020).

[GZ19] David Gamarnik and Ilias Zadik. “The landscape of the planted clique problem: Dense subgraphs and the overlap gap property”. In: *arXiv preprint arXiv:1904.07174* (2019).

[HZZ10] Simai He, Jiawei Zhang, and Shuzhong Zhang. “Bounding probability of small deviation: A fourth moment approach”. In: *Mathematics of Operations Research* 35.1 (2010), pp. 208–232.

[Jer92] Mark Jerrum. “Large cliques elude the Metropolis process”. In: *Random Structures & Algorithms* 3.4 (1992), pp. 347–359.

[Kar76] Richard M Karp. “The probabilistic analysis of some combinatorial search algorithms.” In: (1976).

[Kuč95] Luděk Kučera. “Expected complexity of graph partitioning problems”. In: *Discrete Applied Mathematics* 57.2-3 (1995), pp. 193–212.

[Kun23] Tim Kunisky. 2023. URL: http://www.kunisky.com/teaching/2023spring-modernprob/.

[KWB19] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. “Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio”. In: *ISAAC Congress (International Society for Analysis, its Applications and Computation)*. Springer. 2019, pp. 1–50.

[KXZ16] Florent Krzakala, Jiaming Xu, and Lenka Zdeborová. “Mutual information in rank-one matrix estimation”. In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE. 2016, pp. 71–75.

[KZ23] Dmitriy Kunisky and Peiyuan Zhang. “Online algorithms and lower bounds for average-case matrix discrepancy”. In: *arXiv preprint arXiv:2307.10055* (2023).

[Li+18] Hao Li et al. “Visualizing the loss landscape of neural nets”. In: *Advances in neural information processing systems* 31 (2018).

[LKZ15a] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. “MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel”. In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2015, pp. 680–687.

[LKZ15b] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. "Phase transitions in sparse PCA". In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2015, pp. 1635–1639.

[LM17a] Rafał Latała and Dariusz Matlak. "Royen's proof of the Gaussian correlation inequality". In: *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*. Springer. 2017, pp. 265–275.

[LM17b] Marc Lelarge and Léo Miolane. "Fundamental limits of symmetric low-rank matrix estimation". In: *Conference on Learning Theory*. PMLR. 2017, pp. 1297–1301.

[MAB20] Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. "Landscape complexity for the empirical risk of generalized linear models". In: *Mathematical and Scientific Machine Learning*. PMLR. 2020, pp. 287–327.

[Mai24] Antoine Maillard. "Average-case matrix discrepancy: satisfiability bounds". In: *arXiv preprint arXiv:2410.17887* (2024).

[Mat70] D Matual. "On the complete subgraph of a random graph". In: *Combinatory Mathematics and its Appl.* (1970).

[MBM18] Song Mei, Yu Bai, and Andrea Montanari. "The landscape of empirical risk for nonconvex losses". In: *The Annals of Statistics* 46.6A (2018), pp. 2747–2774.

[Mek14a] Raghu Meka. *Discrepancy and beating the union bound*. https://windowsontheory.org/2014/02/07/discrepancy-and-beating-the-union-bound. 2014.

[Mek14b] Raghu Meka. *Discrepancy and Rounding Linear Programs*. https://windowsontheory.org/2014/02/27/discrepancy-and-rounding-linear-programs/. 2014.

[Mek14c] Raghu Meka. *Discrepancy Bounds from Convex Geometry*. https://windowsontheory.org/2014/02/17/discrepancy-bounds-from-convex-geometry/. 2014.

[Mil63] John Willard Milnor. *Morse theory*. 51. Princeton university press, 1963.

[Mon15] Andrea Montanari. "Finding one community in a sparse graph". In: *Journal of Statistical Physics* 161 (2015), pp. 273–299.

[MP67] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Matematicheskii Sbornik* 114.4 (1967), pp. 507–536.

[NP33] Jerzy Neyman and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[Per+18] Amelia Perry et al. "Optimality and sub-optimality of PCA I: Spiked random matrix models". In: *The Annals of Statistics* 46.5 (2018), pp. 2416–2451.

[RF23] Valentina Ros and Yan V Fyodorov. "The high-dimensional landscape paradigm: Spin-glasses, and beyond". In: *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking After 40 Years*. World Scientific, 2023, pp. 95–114.

[Ros+19] Valentina Ros et al. "Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions". In: *Physical Review X* 9.1 (2019), p. 011003.

[Roy14] Thomas Royen. "A simple proof of the Gaussian correlation conjecture extended to multivariate gamma distributions". In: *arXiv preprint arXiv:1408.1028* (2014).

[Sch42] IJ Schoenberg. "Positive definite functions on spheres". In: *Duke Mathematical Journal* 9.1 (1942).

[Sel24a] Mark Sellke. 2024. URL: https://msellke.com/courses/STAT_291/course_page_website.html.

[Sel24b] Mark Sellke. "The threshold energy of low temperature Langevin dynamics for pure spherical spin glasses". In: *Communications on Pure and Applied Mathematics* 77.11 (2024), pp. 4065–4099.

[Šid67] Zbyněk Šidák. "Rectangular confidence regions for the means of multivariate normal distributions". In: *Journal of the American statistical association* 62.318 (1967), pp. 626–633.

[Spe85] Joel Spencer. "Six standard deviations suffice". In: *Transactions of the American mathematical society* 289.2 (1985), pp. 679–706.

[Spe94] Joel Spencer. *Ten lectures on the probabilistic method*. SIAM, 1994.

[Sub17] Eliran Subag. "The complexity of spherical p-spin models—a second moment approach". In: *The Annals of Probability* (2017), pp. 3385–3450.

[TW94] Craig A Tracy and Harold Widom. "Level-spacing distributions and the Airy kernel". In: *Communications in Mathematical Physics* 159 (1994), pp. 151–174.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Wig55] Eugene P Wigner. "Characteristic Vectors of Bordered Matrices With Infinite Dimensions". In: *Annals of Mathematics* 62.3 (1955), pp. 548–564.

[Wis28] John Wishart. "The generalised product moment distribution in samples from a normal multivariate population". In: *Biometrika* (1928), pp. 32–52.

[WX23] Yihong Wu and Jiaming Xu. "Statistical inference on graphs: Selected Topics". In: *Lecture notes* (2023).

[Zou12] Anastasios Zouzias. "A matrix hyperbolic cosine algorithm and applications". In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2012, pp. 846–858.