Lectures notes on statistical and computational phase transitions in high-dimensional statistics
Winter 2026

Antoine Maillard
antoine.maillard@inria.fr

Last update: January 29, 2026

Overview

These notes form the core material for a 20-hours master's course in the program "Mathématiques de l'Aléatoire" (M2MDA) at Université Paris Saclay, from January to April 2026. Given the time constraints, it is likely that not all the present material will be/has been covered during the lectures: one purpose of these notes is for interested students to dive into the lecture topics at a deeper level.

Evaluation – For students that wish to validate, the course will be evaluated through presentations of research papers in the last session. A list of possible choices for papers will be given on the course's page on my website.

Acknowledgements – I am particularly grateful to B. Loureiro and F. Krzakala for helpful discussions during the writing of these notes.

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 January 29, 2026.*

Contents

1 Introduction

1.1 What are statistical and computational phase transitions?

This course is related to the fundamental question of computational complexity theory: which problems can be solved by computers? Precisely, one wishes to understand, for a given problem, what are the needed ressources (e.g. in memory or computation time) that are needed to solve it. Remarkably, some problems, while solvable in principle, seem to require prohibitively large resources to be solved as the size of the problem gets bigger: this phenomenon is known as *computational hardness*.

In this course, we introduce several tools to characterize the emergence of computational hardness in problems arising in a large class of problems in high-dimensional statistics. For concreteness, we will focus on two specific classes:

1. **Statistical estimation/inference**: Many problems in modern statistics and machine learning involve detecting or estimating structures from the indirect observation of a data. A typical modeling of this problem is the following: $\mathbf{x}_0 \in \mathbb{R}^d$, sometimes called the "signal", is only observed through an indirect observation $\mathbf{y} \in \mathbb{R}^n$, which can e.g. be corrupted by large amounts of noise. Given the observation of $\mathbf{y}$, and some possible "prior" knowledge about the structure of $\mathbf{x}_0$, one aims to recover it as well as possible. Importantly, we wish to solve such problems in a "modern statistics" framework, where both the number of observations n but also the number of parameters d to recover, are very large. A very non-exhaustive list of examples of such models include:

   (a) In *community detection*, one observes a large graph, and wishes to recover from it hidden *communities*, i.e. subgraphs where members of the same communities have a much higher chance of being connected than members of different communities. This kind of structure is very common in realistic networks, and understanding whether recovering communities is feasible has received a lot of attention: we refer to the course of L. Massoulié [MS23].

   (b) Interestingly, there is a toy model, dubbed *spiked matrix model*, that corresponds almost exactly to the community detection problem in a large random graph. There the observations take the form of a matrix, and the signal is assumed to have a *low-rank* structure:

   $$\mathbf{Y} = \sqrt{\lambda}\mathbf{x}_0\mathbf{x}_0^\top + \mathbf{W} \tag{1}$$

   Here $\mathbf{W}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ elements. One can also generalize this problem to recovering a rank-one tensor:

   $$\mathbf{Y} = \sqrt{\lambda}\mathbf{x}_0^{\otimes p} + \mathbf{W}, \tag{2}$$

   where $p \geq 2$ is the *order of the tensor*, and $W_{i_1,\cdots,i_p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Eq. (1) corresponds to the case $p = 2$. Here, the structure of the signal $\mathbf{x}_0$ can be modeled at will, e.g. by chosing a prior distribution $\mathbf{x}_0 \sim P_0$.

   (c) Imagine that $\mathbf{x}_0 \in \mathbb{R}^d$ corresponds to a signal written in a basis where it is k-sparse, i.e. all but k entries of $\mathbf{x}_0$ are zero, and $k \ll d$. This is for instance true of audio signals in the Fourier basis, or images in the wavelet basis. In *compressive sensing*, one aims at leveraging this structure to invert a large linear system

   $$\mathbb{R}^n \ni \mathbf{y} = \mathbf{A}\mathbf{x}_0, \tag{3}$$

where $n \ll d$, and $\mathbf{A}$ is a so-called "measurement" matrix, which is also known to the observer. The goal of compressive sensing is to exploit the sparsify of $\mathbf{x}_0$ to invert this under-determined linear system: it has particularly important applications in MRI imaging, and we refer to [BSS23, Chapter 10] for the basis of the theory of compressed sensing.

(d) Generalizing eq. (3), one may consider more general observations of the type

$$y_i = g\left(\mathbf{a}_i \cdot \mathbf{x}_0\right), \tag{4}$$

where the dataset is composed of $\mathcal{D} \coloneqq \{(y_i, \mathbf{a}_i)\}_{i=1}^d$. This is a so-called *single-index model*, and (along with natural generalizations known as *multi-index* models) can serve as a theoretical playground to understand the feasibility of learning some hidden structure in a large dataset, e.g. by neural networks.

2. **Optimization**: In these kind of problems, one is given a real function $R(\boldsymbol{\theta})$ on a high-dimensional set ($\boldsymbol{\theta} \in \mathcal{M}$), and the aim is to compute

$$\boldsymbol{\theta}^\star \coloneqq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} R(\boldsymbol{\theta}).$$

As we will discuss more in Section 6, the optimization of such high-dimensional *empirical risk/loss functions* is the workhorse of modern machine learning. There, a prototypical example of a function $R(\boldsymbol{\theta})$ may be given as

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

and depends on a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ of output/input pairs, from which we aim at learning the underlying input-to-output function. A somewhat simpler example is given by Maximum Likelihood Estimation (MLE) in the spiked tensor model above (eq. (2)). If we assume that $\|\mathbf{x}_0\|_2 = 1$, the MLE estimator is

$$\hat{\mathbf{x}} \coloneqq \arg\max_{\|\mathbf{x}\|=1} \langle \mathbf{x}^{\otimes p}, \mathbf{Y}\rangle = \arg\max_{\|\mathbf{x}\|=1} \left[\sum_{1 \le i_1, \cdots, i_p \le d} W_{i_1,\cdots,i_p} x_{i_1} \cdots x_{i_p} + \sqrt{\lambda}(\mathbf{x} \cdot \mathbf{x}_0)^p\right].$$

Statistical vs algorithmic performance – As we mentioned already, our goal is to answer, for *very high dimensions* ($d \gg 1$), the following questions:

1. When is estimation/detection/optimization possible at all (regardless of the computation time)?

2. If it is possible, can it be done with efficient algorithms, e.g. that run in polynomial time (in the parameters of the problem), or local optimization procedures?

The answer to these questions may change drastically as the parameters of the problem change, e.g. when the noise level gets smaller, or the size of the training dataset gets bigger: this can lead to sharp *phase transitions*, where the algorithmic feasibility of this problem can change very abruptly. Characterizing these phenomena is one of the main goals of this lecture.

Random high-dimensional measures – Tackling these questions has historically been a very inter-disciplianary endeavor, with a blend of tools from *probability theory, information theory, computer science, and statistical physics.* The later might be a bit surprising, but as we will see the main techniques we will see in this course have been developed in the broad study of *high-dimensional random probability measures*: probability distributions over $\mathbb{R}^d$ (with $d \gg 1$) which can usually be written as

$$\mu(\mathrm{d}\mathbf{x}) \propto e^{\beta H(\mathbf{x})} \mu_0(\mathrm{d}\mathbf{x}). \tag{5}$$

Here μ_0 is a deterministic reference measure (typically $\mu_0 = \mathrm{Unif}(\{\pm 1\}^d)$, or $\mu_0 = \mathrm{Unif}(\mathcal{S}^{d-1})$, the uniform distribution over the unit sphere). $\beta > 0$ is sometimes called the *inverse temperature*, $H : \mathbb{R}^d \to \mathbb{R}$, the *Hamiltonian* of the system[1], which is here a *random function.* While these distributions arose in the statistical physics of peculiar material called "spin glasses", it was soon realized that they are ubiquitous in other fields, among them high-dimensional statistics. To take the two examples we detailed above:

- In statistical inference/estimation, the *Bayesian posterior* $\mathbb{P}(\mathbf{x}_0|\mathbf{y}) \propto \mathbb{P}(\mathbf{y}|\mathbf{x}_0)\mathbb{P}(\mathbf{x}_0)$ is a random probability distribution over $\mathbb{R}^d$ (since $\mathbf{y}$ is random, e.g. through the noise). The *prior* distribution $\mathbb{P}(\mathbf{x}_0)$ plays the role of the reference measure in eq. (5), while the *log-likelihood* $\log \mathbb{P}(\mathbf{y}|\mathbf{x}_0)$ is akin to the Hamiltonian function.

- In optimization, a way to understand the feasibility of optimization is to study the geometry of the sub-level sets $S(\ell) \coloneqq \{\boldsymbol{\theta} : R(\boldsymbol{\theta}) \geq \ell\}$. The "Gibbs-Boltzmann" measure of eq. (5) can yield many information about the structure of these subelevel sets: for instance $\beta \to \infty$ corresponds to the uniform distribution over minimizers, and more generally we expect in many cases that μ_β is related to the uniform distribution over $S(\ell_\beta)$ for some $S(\ell_\beta)$. Notice that

1.2 Structure of the course

The lecture will be organized around different ways to investigate statistical and computational hardness in high-dimensional statistics. For the majority of the course, we will study the models of eqs. (1) and (2) as our driving examples, and mention extensions to other models along the way.

- We start in Section 2 by introducing a broad class of Gaussian additive models (which includes the spiked matrix and tensor models). We give some reminders of classical results in information theory, and introduce a statistical physics nomenclature. We also see a first example of a phase transition in a high-dimensional estimation problem (Gaussian mean location).

- Coming back to the spiked Wigner problem, we analyze in Section 3 a simple spectral method motivated by PCA, and derive sharp asymptotics for its performance using tools from random matrix theory.

- Section 4 is devoted to approaches from statistical physics. We will derive sharp information-theoretic results using this framework, as well as analyze *approximate message-passing*, a powerful class of algorithms: we will compare them to the performance of the PCA algorithm derived earlier. This will give us a sharp picture of statistical and computational phase transitions in the spiked Wigner model.

- In Section 5 we take a different point of view on computational hardness. We consider the *detection* problem: e.g. when can we distinguish a sample $\mathbf{Y}$ from eq. (1) from pure noise? We will introduce the important notion of contiguity

[1] Notice that in physics one usually considers the sign convention $e^{-\beta H}$ for the weight.

to argue about feasibility of detection problems, and introduce the so-called *low-degree likelihood ratio* method, based on the performance of algorithms which are low-degree polynomials of the data. This yields another way to probe statistical and computational hardness of many prolems in high-dimensional statistics, and we will compare its predictions to the statistical physics approach.

- Finally, in Section 6 we consider optimization problems in high-dimension, with the driving example of maximum likelihood estimation for the spiked tensor problem. We introduce the Kac-Rice formula of random differential geometry, and show how this allows to characterize the topology of high-dimensional non-convex landscapes, and probe when local optimization is feasible.

1.3 A disclaimer on mathematical rigor

This course is targeted at students in mathematics, with a good background in probability theory, and in particular some experience in high-dimensional probability (some reminders and classical results are given in Appendix A). While this course is mathematical, some of the arguments presented in Section 4 are inherently heuristic arguments of statistical physics, and some derivations and arguments there will not be rigorous. We will precise when this is the case, and also present how mathematicians have now been able to prove the large majority of these physics results.

1.4 References

Particular credit – These notes are heavily inspired by existing lectures and reviews, and I wish to give particular credit for many things that were borrowed from [El 21] (Ahmed El Alaoui. 2021. URL: https://courses.cit.cornell.edu/stsci6940/) in Sections 2, 4 and 5, in [Kun25] (Tim Kunisky. 2025. URL: http://www.kunisky.com/static/teaching/2025fall-rmt/rmt-notes-2025.pdf) in Section 3, and in [MS24] (Montanari and Sen (2024), *"A friendly tutorial on mean-field spin glass techniques for non-physicists"*) in Section 4.

Some other important references I used while making these notes include:

- Antoine Maillard. 2024. URL: https://anmaillard.github.io/assets/pdf/lecture_notes/MDS_Fall_2024.pdf: a set of lecture notes for a class I taught at ETH Zürich in 2024.
- [BN11] (Benaych-Georges and Nadakuditi (2011), *"The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices"*) for Section 3
- [KWB19] for Section 5.
- [Ben+19; Sel24] for Section 6.

More references are also given in the corresponding sections. Finally, here is a very non-exhaustive and personal list of some great books and reviews for readers interested in these topics.

- [AGZ10] : Anderson, Guionnet, and Zeitouni (2010), An introduction to random matrices
- [PB20] : Potters and Bouchaud (2020), A first course in random matrix theory: for physicists, engineers and data scientists
- [BSS23] : Bandeira, Singer, and Strohmer (2023), Mathematics of Data Science

- [Han14] : Ramon van Handel. *Probability in High Dimension*. 2014. URL: https://web.math.princeton.edu/~rvan/APC550.pdf

- [Ver18] : Vershynin (2018), High-dimensional probability: An introduction with applications in data science

- [Tal10] : Talagrand (2010), Mean field models for spin glasses: Volume I: Basic examples

- [ZK16] : Zdeborová and Krzakala (2016), *"Statistical physics of inference: Thresholds and algorithms"*

- [KZ24] : Krzakala and Zdeborová (2024), *"Statistical physics methods in optimization and machine learning"*

- [Bar19] : Barbier (2019), Mean-field theory of high-dimensional Bayesian inference

1.5 Notations

$x, \mathbf{x}, \mathbf{\Phi}$	Scalar, vector, matrix.
$\mathbf{x} \cdot \mathbf{y}$ or $\mathbf{x}^\intercal \mathbf{y}$	Dot product between $\mathbf{x}$ and $\mathbf{y}$.
$\mathcal{S}^{d-1}(r)$, $\mathcal{S}^{d-1}$	Euclidean sphere in $\mathbb{R}^d$ of radius r, unit Euclidean sphere in $\mathbb{R}^d$.
$\mathbb{S}_d$, $\mathbb{S}_d^+$	$d \times d$ symmetric matrices, $d \times d$ positive semidefinite matrices.
$v_{\max}(\mathbf{Y})$	Generic notation for the eigenvector of $\mathbf{Y} \in \mathbb{S}_d$ with the largest eigenvalue.
$\mathbb{R}_+$, $\mathbb{R}_+^\star$	Set of non-negative and strictly positive reals.
$\mathbb{C}_+$	Complex numbers with strictly positive imaginary part.
$x = \Theta(y)$	Two variables of the same order, i.e. $x = \mathcal{O}(y)$ and $y = \mathcal{O}(x)$.
I_n	The identity matrix of size n.
$\mathcal{P}(\mathbb{R})$	The set of real probability distributions.
$\mathbb{E}$	Expectation with respect to all involved random variables.
$\mathbb{E}_{X,Y}$	Expectation with respect to X, Y only.
$X \overset{\mathrm{d}}{=} Y$	X and Y have the same distribution.
$\|\mathbf{x}\|_0$	The number of non-zero elements of $\mathbf{x}$.

2 Gaussian additive models and reminders of Bayesian inference

2.1 Definition

We introduce here a particular class of statistical models for estimation and detection. They are conceptually very simple, which will allow us to develop a precise mathematical analysis of their high-dimensional limit while keeping the exposition relatively accessible. We will see specific examples of such models later on, here we introduce them in generality: informally, they correspond to recovering/detecting a signal "blurred" by additive Gaussian noise.

Definition 2.1 (*Gaussian additive model*)

Let $d \geq 1$, and $\mathbf{X}_0 \in \mathbb{R}^d$ be drawn from P_0 (called the "prior"), a probability distribution over $\mathbb{R}^d$ with a finite second moment. Let $\mathbf{W} \sim \mathcal{N}(0, \mathrm{I}_d)$ and $\lambda \geq 0$. We define $\mathbb{P}_\lambda$ as the law of

$$\mathbf{Y} = \mathbf{W} + \sqrt{\lambda}\mathbf{X}_0.$$

Remark – P_0, $\mathbb{P}_\lambda$ should be written as $P_0^{(d)}, \mathbb{P}_\lambda^{(d)}$, as they are sequences of probability distributions on $\mathbb{R}^d$. We however refrain from writing this d-dependency explicity, as it will always be clear in the arguments.

Signal-to-noise ratio – $\lambda \geq 0$ plays the role of a signal-to-noise ratio (SNR): equivalently one can write the observationshas $\widetilde{\mathbf{Y}} = \mathbf{X} + \sqrt{\Delta}\mathbf{W}$, with $\Delta := \lambda^{-1}$ the noise variance.

Estimation and detection – In a statistical setting, the statistician has access to a sample of $\mathbf{Y}$. Crucially, we will assume throughout this class that the statistican also *knows the value of* $\lambda > 0$ *and the prior distribution* P_0. The statistican wishes to answer the following questions:

- **Detection:** Can she distinguish a sample $\mathbf{Y} \sim \mathbb{P}_\lambda$ from a sample $\mathbf{W} \sim \mathbb{P}_0$?
- **Recovery/estimation:** Can she recover the value of $\mathbf{X}_0$ (exactly, or approximately) from $\mathbf{Y}$?

We will make these questions mathematically more precise later on. Crucially, we want to answer these questions *in the high-dimensional limit*, i.e. as $d \to \infty$.

2.2 Posterior measure, free energy, and mutual information

Let us now introduce some classical objects of Bayesian statistics applied to the Gaussian additive model. For more motivations on Bayesian statistics and inference, we refer the reader e.g. to the introduction of the course [Bar19].

Minimal MSE estimator – We focus for now on the recovery problem. For a given estimator $\hat{\mathbf{X}}(\mathbf{Y})$, a natural way to gauge its quality is via its *mean squared error* (MSE), which is defined as

$$\mathrm{MSE}(\hat{\mathbf{X}}) := \mathbb{E}_{\mathbf{Y}}\left[\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbf{X}_0\|_2^2\right]. \tag{6}$$

The best estimator in terms of MSE is simply the posterior average of $\mathbf{X}$.

Theorem 2.1 (*Bayes-optimal estimator*)

The estimator $\hat{\mathbf{X}} : \mathbb{R}^d \to \mathbb{R}^d$ that achieves the minimum MSE is given by the posterior mean

$$\hat{\mathbf{X}}_{\text{opt}}(\mathbf{Y}) := \mathbb{E}[\mathbf{X}|\mathbf{Y}].$$

We call its error the *minimal mean squared error* (MMSE)

$$\text{MMSE} := \arg\min_{\hat{\mathbf{X}}(\mathbf{Y})} \text{MSE}(\hat{\mathbf{X}}) = \mathbb{E}_{\mathbf{Y}}\left[\|\mathbb{E}[\mathbf{X}|\mathbf{Y}] - \mathbf{X}\|_2^2\right].$$

In probability terms, the conditional expectation $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ is the orthogonal projection of $\mathbf{X}$ onto the vector space of all square-integrables $\mathbf{Y}$-measurable random variables. **Proof of Theorem 2.1 –** For any estimator $\hat{\mathbf{X}}$, we have

$$\begin{aligned}\text{MSE}(\hat{\mathbf{X}}) &= \mathbb{E}[\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbf{X}_0\|^2], \\ &= \mathbb{E}[\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}] + \mathbb{E}[\mathbf{X}|\mathbf{Y}] - \mathbf{X}_0\|^2], \\ &= \text{MSE}(\mathbf{Y} \to \mathbb{E}[\mathbf{X}|\mathbf{Y}]) + \mathbb{E}[\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2] \\ &\quad + 2\mathbb{E}[(\hat{\mathbf{X}}(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}]) \cdot (\mathbb{E}[\mathbf{X}|\mathbf{Y}] - \mathbf{X}_0)].\end{aligned}$$

By the tower property of expectation:

$$\mathbb{E}[f(\mathbf{Y}) \cdot (\mathbb{E}[\mathbf{X}|\mathbf{Y}] - \mathbf{X}_0)] = \mathbb{E}_{\mathbf{Y}}[f(\mathbf{Y}) \cdot \mathbb{E}_{\mathbf{X}\sim\mathbb{P}(\cdot|\mathbf{Y})}(\mathbb{E}[\mathbf{X}|\mathbf{Y}] - \mathbf{X})] = 0.$$

Thus

$$\text{MSE}(\hat{\mathbf{X}}) = \text{MSE}(\mathbf{Y} \to \mathbb{E}[\mathbf{X}|\mathbf{Y}]) + \mathbb{E}[\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2],$$

which ends the proof. □

Posterior distribution – Theorem 2.1 motivates to consider the posterior distribution of $\mathbf{X}$ given $\mathbf{Y}$ (i.e. the probability that $\mathbf{Y}$ was generated by the value $\mathbf{X}_0 = \mathbf{X}$). It is given by Bayes' rule

$$\mathrm{d}\mathbb{P}(\mathbf{X}|\mathbf{Y}) = \frac{\varphi(\mathbf{Y}|\mathbf{X})}{\widetilde{\mathcal{Z}}(\mathbf{Y})} \cdot \mathrm{d}P_0(\mathbf{X}),$$

where $\varphi(\mathbf{Y}|\mathbf{X})$ is the density of $\mathbf{Y}$ given $\mathbf{X}_0 = \mathbf{X}$, and $\widetilde{\mathcal{Z}}(\mathbf{Y}) = \int \mathrm{d}P_0(\mathbf{X})\varphi(\mathbf{Y}|\mathbf{X})$ is a normalization[2]. In the Gaussian additive model of Definition 2.1, we get after simple manipulations:

$$\mathrm{d}\mathbb{P}(\mathbf{X}|\mathbf{Y}) = \frac{e^{-\frac{\lambda}{2}\|\mathbf{X}\|^2 + \sqrt{\lambda}\mathbf{Y}\cdot\mathbf{X}}}{\mathcal{Z}(\lambda; \mathbf{Y})} \mathrm{d}P_0(\mathbf{X}). \tag{7}$$

The statistical physics nomenclature – By analogy with the Gibbs-Boltzmann distribution in statistical physics (see the introduction), we introduce a series of definitions whoses names often come from statistical physics, but which are merely rebrandings of classical quantities in information theory. Still, we use the statistical physics terminology in the majority of this class: this will be particularly useful in Section 4, to connect to the existing literature connecting statistical physics and high-dimensional statistics.

Definition 2.2 (*Statistical physics nomenclature*)

We define several quantities for the problem of Definition 2.1.

[2] $\widetilde{\mathcal{Z}}(\mathbf{Y})$ is the density of the random variable $\mathbf{Y}$.

(1) The log-likelihood function, or *Hamiltonian*, is

$$H(\mathbf{X}) := -\frac{\lambda}{2}\|\mathbf{X}\|^2 + \sqrt{\lambda}\mathbf{Y}\cdot\mathbf{X}. \tag{8}$$

Notice that $H(\mathbf{X})$ also depends on $(\lambda, \mathbf{Y})$: it is a random function.

(2) The *partition function*, is

$$\mathcal{Z}(\lambda;\mathbf{Y}) := \int e^{-\frac{\lambda}{2}\|\mathbf{X}\|^2+\sqrt{\lambda}\mathbf{Y}\cdot\mathbf{X}}\mathrm{d}P_0(\mathbf{X}) = \int e^{H(\mathbf{X})}\mathrm{d}P_0(\mathbf{X}). \tag{9}$$

The corresponding *free entropy*[3] is

$$F(\lambda) := \mathbb{E}\log\mathcal{Z}(\lambda;\mathbf{Y}). \tag{10}$$

(3) The posterior distribution of eq. (7) is called the *Gibbs (or Gibbs-Boltzmann) measure*. Often, we will denote it

$$\langle g(\mathbf{X})\rangle := \mathbb{E}[g(\mathbf{X})|\mathbf{Y}], \tag{11}$$

omitting the dependency on $\mathbf{Y}$ of $\langle\cdot\rangle$ when it is not ambiguous. Keep in mind that this is a random probability measure!

Thermodynamic limit – Recall that we wish to consider these models in the high-dimensional limit, i.e. when $d \to \infty$. Sometimes, we will also use a physics language, and describe it as the *thermodynamic* limit.

The Nishimori identity – The following elementary property of posterior distributions will play a crucial role in our analysis later on.

Proposition 2.2 (*Nishimori identity*)

Recall that $\mathbf{Y} = \sqrt{\lambda}\mathbf{X}_0 + \mathbf{W}$. Let $\mathbf{X}_1, \mathbf{X}_2$ drawn independently from the posterior distribution of eq. (7). Then

$$(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) \overset{\mathrm{d}}{=} (\mathbf{X}_1, \mathbf{X}_0, \mathbf{Y})$$

Proposition 2.2 is called the "Nishimori identity" in statistical physics for historical reasons, however it is a quite trivial consequence of Bayes' formula.

Proof of Proposition 2.2 – It is equivalent to sample $(\mathbf{X}, \mathbf{Y})$ according to their joint law, or to sample first $\mathbf{Y}$ according to its marginal distribution and then sample $\mathbf{X}$ from the posterior distribution $\mathbb{P}(\cdot|\mathbf{Y})$. To make it more concrete, one can consider Ψ any test function, and write:

$$\begin{aligned}\mathbb{E}[\Psi(\mathbf{X}_1, \mathbf{X}_0, \mathbf{Y})] &= \mathbb{E}_{\mathbf{Y},\mathbf{X}_0}\mathbb{E}_{\mathbf{X}_1\sim\mathbb{P}(\cdot|\mathbf{Y})}[\Psi(\mathbf{X}_1, \mathbf{X}_0, \mathbf{Y})], \\ &= \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}(\cdot|\mathbf{Y})}\mathbb{E}_{\mathbf{X}_1\sim\mathbb{P}(\cdot|\mathbf{Y})}[\Psi(\mathbf{X}_1, \mathbf{X}_0, \mathbf{Y})], \\ &= \mathbb{E}[\Psi(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y})].\end{aligned}$$

□

A trivial corollary is the following, where we also introduce the notion of *overlap*, which will be very useful later.

[3] In physics, one often considers the *free energy*, which is equal to $-\mathbb{E}\log\mathcal{Z}(\lambda;\mathbf{Y})$. Sometimes there is also a global temperature factor.

Corollary 2.3 (*Equivalence of overlaps*)

Recall that $\mathbf{Y} = \sqrt{\lambda}\mathbf{X}_0 + \mathbf{W}$. Define the overlaps

$$\begin{cases} R_{01} & \coloneqq \mathbf{X}_0 \cdot \mathbf{X}_1, \\ R_{12} & \coloneqq \mathbf{X}_1 \cdot \mathbf{X}_2. \end{cases} \tag{12}$$

If $\mathbf{X}_0 \sim P_0$ and $\mathbf{X}_1, \mathbf{X}_2 \sim \mathbb{P}(\mathbf{X}|\mathbf{Y})$, then $R_{01} \overset{\mathrm{d}}{=} R_{12}$.

Mutual information – Recall that for two random variables (x, y), with joint distribution P_{xy}, and marginals (P_x, P_y), the mutual information is defined as[4]:

$$I(y; x) = I(x; y) \coloneqq D_{\mathrm{KL}}(P_{xy} || P_x \otimes P_y). \tag{13}$$

The following shows that the free entropy is essentially the mutual information, up to a sign and an additive constant.

Proposition 2.4 (*Free entropy and mutual information*)

For the model of Definition 2.1,

$$I(\mathbf{X}_0; \mathbf{Y}) = \frac{\lambda}{2}\mathbb{E}[\|\mathbf{X}_0\|^2] - F(\lambda).$$

Proof of Proposition 2.4 – To simplify, we denote here $\mathbb{P}_{\mathbf{X}} = P_0$, $\mathbb{P}_{\mathbf{Y}} = \mathbb{P}_\lambda$ the marginal laws of $\mathbf{X}_0$ and $\mathbf{Y}$, and $\mathbb{P}_{\mathbf{X},\mathbf{Y}}$ their joint law. Using the definition of the mutual information in eq. (13):

$$\begin{aligned}
I(\mathbf{X}_0; \mathbf{Y}) &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\mathbf{X},\mathbf{Y}}}{\mathrm{d}(\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}})}\right], \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\mathbf{Y}}(\mathbf{Y}) \cdot \mathrm{d}\mathbb{P}_{\mathbf{X}|\mathbf{Y}}(\mathbf{X})}{\mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{X}) \cdot \mathrm{d}\mathbb{P}_{Y}(\mathbf{Y})}\right], \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}|\mathbf{Y}}(\mathbf{X})}{\mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{X})}\right], \\
&\overset{\text{(a)}}{=} \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log \frac{e^{-\frac{\lambda}{2}\|\mathbf{X}\|^2 + \sqrt{\lambda}\mathbf{Y}\cdot\mathbf{X}}}{\mathcal{Z}(\lambda; \mathbf{Y})}\right], \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[-\frac{\lambda}{2}\|\mathbf{X}\|^2 + \sqrt{\lambda}\mathbf{Y} \cdot \mathbf{X}\right] - F(\lambda), \\
&\overset{\text{(b)}}{=} -\frac{\lambda}{2}\mathbb{E}[\|\mathbf{X}\|^2] + \sqrt{\lambda}\mathbb{E}[(\sqrt{\lambda}\mathbf{X} + \mathbf{W}) \cdot \mathbf{X}] - F(\lambda), \\
&= \frac{\lambda}{2}\mathbb{E}[\|\mathbf{X}\|^2] - F(\lambda),
\end{aligned}$$

using eq. (7) in (a), and Definition 2.1 in (b). □

Notice that $I(\mathbf{X}_0; \mathbf{Y}) \geq 0$: in particular, we showed that $F(\lambda) \leq (\lambda/2)\mathbb{E}[\|\mathbf{X}_0\|^2]$.

The MMSE is also related to the derivative of the free entropy (or of the mutual information) with respect to the SNR λ.

Proposition 2.5 (*I-MMSE formula*)

Consider the model of Definition 2.1, and denote its MMSE as $\mathrm{MMSE}(\lambda)$. For any

[4] Recall the KL divergence is $D_{\mathrm{KL}}(P||Q) \coloneqq \mathbb{E}_P \log \mathrm{d}P/\mathrm{d}Q$.

$\lambda \geq 0$ we have

$$F'(\lambda) = \frac{1}{2}\mathbb{E}[\|\mathbf{X}_0^2\|] - \frac{1}{2}\mathrm{MMSE}(\lambda) = \frac{1}{2}\mathbb{E}_{\mathbf{Y}}\left[\|\mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right].$$

This formula can be stated equivalently in the language of the mutual information by using Proposition 2.4:

$$\frac{\partial I(\mathbf{X};\mathbf{Y})}{\partial \lambda} = \frac{1}{2}\mathrm{MMSE}(\lambda). \tag{14}$$

Proof of Proposition 2.5 – First, the middle and right-hand side of the sought identity are equal, since by Proposition 2.2

$$\mathbb{E}[\|\mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2] = \mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}] \cdot \mathbf{X}_0]$$

For the rest of the proof, we leverage Gaussian integration by parts (see Lemma A.3). We have

$$F(\lambda) = \mathbb{E}\log\int e^{\sqrt{\lambda}\mathbf{Y}\cdot\mathbf{X} - \frac{\lambda}{2}\|\mathbf{X}\|^2}\,\mathrm{d}P_0(\mathbf{X}). \tag{15}$$

Recall that $Y = \sqrt{\lambda}\mathbf{X}_0 + \mathbf{W}$. We also recall the notations introduced in Definition 2.2. This yields:

$$\begin{aligned}
F'(\lambda) &= \frac{\partial}{\partial\lambda}\mathbb{E}_{\mathbf{W},\mathbf{X}_0}\log\int e^{\sqrt{\lambda}\mathbf{W}\cdot\mathbf{X} + \lambda\mathbf{X}_0\cdot\mathbf{X} - \frac{\lambda}{2}\|\mathbf{X}\|^2}\,\mathrm{d}P_0(\mathbf{X}), \\
&= \mathbb{E}_{\mathbf{W},\mathbf{X}_0}\left[\mathbf{X}_0\cdot\langle\mathbf{X}\rangle - \frac{1}{2}\langle\|\mathbf{X}\|^2\rangle + \frac{1}{2\sqrt{\lambda}}\mathbf{W}\cdot\langle\mathbf{X}\rangle\right], \\
&\overset{\text{(a)}}{=} \mathbb{E}_{\mathbf{W},\mathbf{X}_0}\left[\|\langle\mathbf{X}\rangle\|^2 - \frac{1}{2}\|\mathbf{X}_0\|^2 + \frac{1}{2\sqrt{\lambda}}\mathbf{W}\cdot\langle\mathbf{X}\rangle\right], \\
&\overset{\text{(b)}}{=} \mathbb{E}_{\mathbf{W},\mathbf{X}_0}\left[\|\langle\mathbf{X}\rangle\|^2 - \frac{1}{2}\|\mathbf{X}_0\|^2 + \frac{1}{2\sqrt{\lambda}}\sum_{i=1}^{d}\frac{\partial}{\partial W_i}\langle X_i\rangle\right], \\
&= \mathbb{E}_{\mathbf{W},\mathbf{X}_0}\left[\|\langle\mathbf{X}\rangle\|^2 - \frac{1}{2}\|\mathbf{X}_0\|^2 + \frac{1}{2}\sum_{i=1}^{d}(\langle X_i^2\rangle - \langle X_i\rangle^2)\right], \\
&\overset{\text{(c)}}{=} \frac{1}{2}\mathbb{E}[\|\langle\mathbf{X}\rangle\|^2].
\end{aligned}$$

In (a) and (c) we used the Nishimori identity (Proposition 2.2), and in (b) Gaussian integration by parts. □

Notably, a corollary of Proposition 2.5 is the following. Proving it involves heavy computations but follows exactly the same lines as the proof of Proposition 2.5, so we leave it as an exercise.

Corollary 2.6 (*Properties of the free entropy*)

Consider the model of Definition 2.1. The free entropy $F : \lambda \geq 0 \mapsto F(\lambda)$ is a non-decreasing and non-negative function of λ, and further

$$F''(\lambda) = \frac{1}{2}\mathbb{E}\left[\|\mathrm{cov}(\mathbf{X}|\mathbf{Y})\|_F^2\right] = \frac{1}{2}\sum_{i,j}\mathbb{E}\left[(\langle X_iX_j\rangle - \langle X_i\rangle\langle X_j\rangle)^2\right]. \tag{16}$$

In particular, F is convex.

This last conclusion is intuitively very natural given Proposition 2.5: it is just saying that $\lambda \mapsto \mathrm{MMSE}(\lambda)$ is decreasing, i.e. that as the signal strength gets higher, the optimal mean-squared error decreases.

Other estimators – One can also consider other estimators $\hat{\mathbf{X}}(\mathbf{Y})$, which can optimize different objectives than the mean-squared error. Some examples include:

- When P_0 has a density, the *Maximum A Posteriori* estimator, which maximizes the posterior density:

$$\hat{\mathbf{X}}_{\text{MAP}}(\mathbf{Y}) \coloneqq \underset{\hat{\mathbf{X}}(\mathbf{Y})}{\arg\max} \log \mathbb{P}(\mathbf{X}|\mathbf{Y}) = \underset{\hat{\mathbf{X}}(\mathbf{Y})}{\arg\max}[\log \varphi(\mathbf{Y}|\hat{\mathbf{X}}) + \log P_0(\hat{\mathbf{X}})]. \tag{17}$$

- The *Maximum Likelihood* estimator, which maximizes only the likelihood term:

$$\hat{\mathbf{X}}_{\text{MLE}}(\mathbf{Y}) \coloneqq \underset{\hat{\mathbf{X}}(\mathbf{Y}) \in \text{supp}\, P_0}{\arg\max} \log \varphi(\mathbf{Y}|\hat{\mathbf{X}}). \tag{18}$$

  Notice that these two estimators coincide when P_0 is the uniform distribution on its support.

One can also define more general class of estimators. We will focus mainly on the MSE estimator for the moment, and will come back to the MLE/MAP estimators when discussing optimization procedures in Section 6 when discussing optimization. Indeed, notice that in a Gaussian additive model:

$$\hat{\mathbf{X}}_{\text{MLE}}(\mathbf{Y}) = \underset{\mathbf{X} \in \text{supp}\, P_0}{\arg\max} \left[\mathbf{Y} \cdot \mathbf{X} - \frac{\lambda}{2}\|\mathbf{X}\|^2\right]$$

and one can attack this problem e.g. by local optimization procedures.

2.3 The simplest example: scalar denoising

Let us start with the simplest instance of a Gaussian additive model: the scalar setting $d = 1$. The observations are generated as

$$y = \sqrt{\lambda} x_0 + z, \tag{19}$$

with $x_0 \sim P_0$ and $z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned} F(\lambda) &= \mathbb{E}_y \log \int e^{\sqrt{\lambda} xy - \frac{\lambda}{2} x^2} \mathrm{d}P_0(x), \\ &= \mathbb{E}_{z,x_0} \log \int e^{\sqrt{\lambda} xz + \lambda x x_0 - \frac{\lambda}{2} x^2} \mathrm{d}P_0(x). \end{aligned} \tag{20}$$

2.3.1 Gaussian prior

We start with the simplest example: $P_0 = \mathcal{N}(0,1)$. The integral is now explicit

$$\begin{aligned} F(\lambda) &= \mathbb{E} \log \int \frac{\mathrm{d}x}{\sqrt{2\pi}} e^{-\frac{1+\lambda}{2} x^2 + x[\lambda x_0 + \sqrt{\lambda} z]}, \\ &= \mathbb{E} \log \frac{e^{\frac{(\lambda x_0 + \sqrt{\lambda} z)^2}{2(1+\lambda)}}}{\sqrt{1+\lambda}}, \\ &= -\frac{1}{2} \log(1+\lambda) + \mathbb{E}\left[\frac{(\lambda x_0 + \sqrt{\lambda} z)^2}{2(1+\lambda)}\right], \\ &= -\frac{1}{2} \log(1+\lambda) + \frac{\lambda}{2}. \end{aligned} \tag{21}$$

From there we get the mutual information and MMSE as:

$$\begin{cases} I(x_0; y) &= \frac{1}{2} \log(1+\lambda), \\ \text{MMSE}(\lambda) &= \frac{1}{1+\lambda}. \end{cases} \tag{22}$$

The optimal estimator $\hat{x}_{\rm opt} = \mathbb{E}[x_0|y]$ of Theorem 2.1 is also easy to write here. Indeed, notice that (x_0, y) are jointly Gaussian random variables. We can thus use classical *Gaussian conditioning* result, which essentially states that the conditional expectation is linear in the case of jointly Gaussian random variables:

Theorem 2.7 (*Gaussian conditioning*)

Let $n, p \geq 1$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n \times \mathbb{R}^p$ be zero-mean and jointly Gaussian vectors. Then

$$\mathbb{E}[\mathbf{u}|\mathbf{v}] = \mathbf{A}^\star \mathbf{v}, \tag{23}$$

where $\mathbf{A}^\star \in \mathbb{R}^{n\times p}$ is the solution to the least-squares problem

$$\mathbf{A}^\star = \operatorname*{arg\,min}_{\mathbf{A}\in\mathbb{R}^{n\times p}} \mathbb{E}_{\mathbf{u},\mathbf{v}} \left[\|\mathbf{u} - \mathbf{A}^\star \mathbf{v}\|^2\right]. \tag{24}$$

We leave the proof of Theorem 2.7 as an exercise[5]. In our simple case, $n = p = 1$ and thus $\mathbb{E}[x_0|y]$ is the orthogonal projection of x_0 on y (with the L^2 norm), thus

$$\hat{x}_{\rm opt}(y) = \mathbb{E}[x_0|y] = \frac{\mathbb{E}[x_0 y]}{\mathbb{E}[y^2]} y = \frac{\sqrt{\lambda}}{1+\lambda} y. \tag{25}$$

2.3.2 Generic prior

We now consider a generic P_0 with mean zero and variance 1. Notice that the estimator of eq. (25) still reaches

$$\mathrm{MSE}\left(y \mapsto \frac{\sqrt{\lambda}}{1+\lambda} y\right) = \mathbb{E}\left[\left(x_0 - \frac{\sqrt{\lambda}}{1+\lambda} y\right)^2\right] \overset{\text{(a)}}{=} \frac{1}{1+\lambda}.$$

Indeed, notice that (a) holds for $P_0 = \mathcal{N}(0,1)$ (as we showed), and it clearly involves only the first two moments of P_0, which we assumed to be $(0,1)$. In particular, this implies that

$$\mathrm{MMSE}(P_0; \lambda) \leq \mathrm{MMSE}(\mathcal{N}(0,1); \lambda) = \frac{1}{1+\lambda}. \tag{26}$$

This formalizes that the Gaussian prior is thus the *"least-informative" one*, in the sense that the MMSE is the highest for this choice of prior. In information theory, this is known as the Shannon-Hartley theorem. By integrating out the I-MMSE formula, this can also be stated in terms of free entropy and mutual information:

$$\begin{cases} I_{P_0;\lambda}(x_0; y) & = \frac{1}{2}\int_0^\lambda \mathrm{MMSE}(P_0; t)\mathrm{d}t \leq \frac{1}{2}\log(1+\lambda) = I_{\mathcal{N}(0,1);\lambda}(x_0; y), \\ F_{P_0}(\lambda) & = \frac{1}{2}\left[1 - \int_0^\lambda \mathrm{MMSE}(P_0; t)\mathrm{d}t\right] \geq \frac{\lambda}{2} - \frac{1}{2}\log(1+\lambda) = F_{\mathcal{N}(0,1)}(\lambda), \end{cases} \tag{27}$$

where the inequalities holds for any P_0 with zero mean and unit variance.

[5] Recall that $\mathbb{E}[\mathbf{u}|\mathbf{v}]$ is the orthogonal projection of $\mathbf{u}$ onto the set of square-integrable $\mathbf{v}$-measurable random variables. It is thus enough to show that there exists $\mathbf{A}$ such that $\mathbf{u} - \mathbf{A}\mathbf{v}$ is independent from $\mathbf{v}$. Since these are Gaussian random variables, independence can be deduced simply from computing their correlation.

2.4 A warm-up: 1-sparse signal denoising

As a slightly harder warm-up, let us analyze a second, and not completely trivial, example of a Gaussian additive model. It will be useful to illustrate some of the phenomenology that will appear later in the class, as this is a high-dimensional model.

Definition 2.3 (1-*sparse signal denoising – Gaussian mean location*)

Let $d \geq 1$, and with $n := 2^d$, we denote $\mathbf{e}_1, \cdots, \mathbf{e}_n$ the canonical basis in $\mathbb{R}^n$. Let $\mathbf{z} \sim \mathcal{N}(0, \mathrm{I}_n)$, and $\sigma_0 \sim \mathrm{Unif}(\{1, \cdots, n\})$. We observe

$$\mathbf{y} := \sqrt{\lambda d} \cdot \mathbf{e}_{\sigma_0} + \mathbf{z}.$$

Definition 2.3 defines a Gaussian additive model in the sense of Definition 2.1, with $\mathbf{X}_0 := \sqrt{d}\,\mathbf{e}_{\sigma_0}$ a 1-sparse vector. Informally, we observe a very high-dimensional Gaussian vector, whose mean has been shifted slightly in one random direction of the canonical basis: our goal is to recover this direction.

2.4.1 Maximum likelihood estimation

Let us analyze a natural candidate for σ_0, when observing $\mathbf{y}$, which is the maximum-likelihood estimate of eq. (18): it is an estimate of σ_0 based on maximizing the log-likelihood $\log \varphi(\mathbf{y}|\sigma)$. Notice that for any $\sigma \in \{1, \cdots, n\}$, we have (we write equalities up to constants independent of σ):

$$\log \varphi(\mathbf{y}|\sigma) = -\frac{1}{2}\|\mathbf{y} - \sqrt{\lambda d}\mathbf{e}_\sigma\|^2 = \sqrt{\lambda d}\, y_\sigma + C(\mathbf{y})$$

The maximum likelihood estimator of eq. (18) is thus simply

$$\hat{\sigma}(\mathbf{y}) := \arg\max_{\sigma \in [n]} y_\sigma. \tag{28}$$

This is a very natural guess: we simply take the largest coordinate of $\mathbf{y}$. We have

$$y_\sigma = \sqrt{\lambda d}\mathbb{1}\{\sigma = \sigma_0\} + z_\sigma.$$

Recall $\log n = d \log 2$. By classical properties of the Gaussian distribution (Proposition A.4), for any $\varepsilon > 0$ we have with probability $1 - o(1)$ as $d \to \infty$:

$$\max_{\sigma \in [n] \setminus \{\sigma_0\}} y_\sigma \in \sqrt{2d \log 2} \cdot [1 - \varepsilon, 1 + \varepsilon].$$

On the other hand $y_{\sigma_0} = \sqrt{\lambda d} + z_{\sigma_0}$, where $z_{\sigma_0} \sim \mathcal{N}(0, 1)$.

Thus, if $\lambda > \lambda_{\mathrm{MLE}} := 2\log 2$, we have $y_{\sigma_0} > \max_{\sigma \in [n] \setminus \{\sigma_0\}} y_\sigma$ with probability $1 - o_d(1)$. On the other hand, for $\lambda < \lambda_{\mathrm{MLE}}$, then $y_{\sigma_0} < \max_{\sigma \in [n] \setminus \{\sigma_0\}} y_\sigma$ with probability $1 - o_d(1)$.

Stated differently, the MLE succeeds above the critical threshold $\lambda_{\mathrm{MLE}} = 2\log 2$, and fails below it: this is a first example of a sharp transition for recovery, here with the MLE estimator.

2.4.2 The free entropy / mutual information

Is the MLE threshold sharp, or can one still recover σ_0 for $\lambda < \lambda_{\mathrm{MLE}}$? We will investigate this question by computing the MMSE of the problem *for any* $\lambda > 0$. As motivated above, we achieve this by computing the free entropy (or mutual information) that we defined in Section 2.2.

$$
\begin{aligned}
F_d(\lambda) &\coloneqq \mathbb{E}_{\mathbf{y}} \log \mathcal{Z}_d(\lambda; \mathbf{y}), \\
&= \mathbb{E}_{\mathbf{y}} \log \left(\frac{1}{n} \sum_{\sigma=1}^{n} e^{-\frac{\lambda d}{2}\|\mathbf{e}_\sigma\|^2 + \sqrt{\lambda d}\,(\mathbf{y}\cdot\mathbf{e}_\sigma)} \right), \\
&= -\frac{\lambda d}{2} + \mathbb{E}_{\mathbf{y}} \log \left(\frac{1}{n} \sum_{\sigma=1}^{n} e^{\sqrt{\lambda d}\, y_\sigma} \right). \qquad (29)
\end{aligned}
$$

How to compute the RHS of eq. (29) ?

A first bound: Jensen's inequality – A first upper bound on $F_d(\lambda)$ is obtained by using Jensen's inequality, since $\mathbb{E} \log[\cdots] \leq \log \mathbb{E}[\cdots]$. In the physics jargon, this is called an *annealed* upper bound on the free entropy. Here, this yields:

$$
F_d(\lambda) \leq -\frac{\lambda d}{2} + \log \left(\frac{1}{n} \sum_{\sigma=1}^{n} \mathbb{E}_{\mathbf{y}} \left[e^{\sqrt{\lambda d}\, y_\sigma} \right] \right). \qquad (30)
$$

For any $\sigma \in [n]$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{y}} \left[e^{\sqrt{\lambda d}\, y_\sigma} \right] &= \left(\mathbb{E}_{\sigma_0} e^{\lambda d \mathbb{1}\{\sigma = \sigma_0\}} \right) \cdot \left(\mathbb{E}_{z \sim \mathcal{N}(0,1)} e^{\sqrt{\lambda d} z} \right), \\
&= \left(\frac{1}{n} [(n-1) + e^{\lambda d}] \right) \cdot e^{\frac{\lambda d}{2}}.
\end{aligned}
$$

Plugging it back in eq. (30) we get (recall $n = 2^d$):

$$
F_d(\lambda) \leq \log \left(1 - 2^{-d} + e^{(\lambda - \log 2) d} \right).
$$

Taking $d \to \infty$, we reach:

$$
\limsup_{d \to \infty} \frac{1}{d} F_d(\lambda) \leq \max(0, \lambda - \log 2). \qquad (31)
$$

In particular, by eq. (31), if $\lambda < \lambda_{\text{ann.}} \coloneqq \log 2$, $(1/d) F_d(\lambda) \to 0$ as $d \to \infty$. By the I-MMSE theorem (Proposition 2.5), this implies that

$$
Q_d(\lambda) \coloneqq \mathbb{E}[\|\mathbb{E}[\mathbf{X}|\mathbf{y}]\|^2] = d\mathbb{E}[\|\mathbb{E}[\mathbf{e}_\sigma|\mathbf{Y}]\|^2]
$$

satisfies, for any $\lambda \in [0, \log 2)$:

$$
\frac{1}{d} \int_0^\lambda Q_d(\tau) \mathrm{d}\tau = \frac{2}{d} F_d(\lambda) \to 0.
$$

Thus $Q_d(\lambda)/d \to 0$ as $d \to \infty$, for almost every $\lambda < \log 2$. Since $\lambda \to Q_d(\lambda)$ is non-decreasing by Corollary 2.6, we reach that $Q_d(\lambda)/d \to 0$ as $d \to \infty$ for all $\lambda < \log 2$. Formally, for $\lambda < \lambda_{\text{ann.}} = \log 2$, it is impossible to estimate σ_0 with a mean-squared error that is asymptotically better than the trivial estimator:

$$
\frac{1}{d} \text{MMSE}(\lambda) = \frac{1}{d} \mathbb{E}_{P_0}[\|\mathbf{X}\|^2] - \frac{1}{d} Q_d(\lambda) = 1 - o(1).
$$

Finer control: conditional Jensen's inequality – Still, this is not completely satisfactory: combining this with the results of Section 2.4.1 leaves an open region for $\lambda_{\text{ann.}} = \log 2 < \lambda < \lambda_{\text{MLE}} = 2 \log 2$. Moreover, we know from the relation between free entropy and mutual information (Proposition 2.4) that $F_d(\lambda) \leq (\lambda/2)$, so eq. (31) can not be tight. A finer control can be achieved by conditioning explicitly on y_{σ_0} in the use of Jensen's inequality. We come back to eq. (29):

$$\begin{aligned}
F_d(\lambda) &= -\frac{\lambda d}{2} + \mathbb{E}_{\mathbf{y}} \log\left(\frac{1}{n}\sum_{\sigma=1}^{n} e^{\sqrt{\lambda d}\, y_\sigma}\right), \\
&\leq -\frac{\lambda d}{2} + \mathbb{E}_{\sigma_0, y_{\sigma_0}} \log\left(\frac{1}{n}\mathbb{E}\left[\sum_{\sigma=1}^{n} e^{\sqrt{\lambda d}\, y_\sigma}\middle| y_{\sigma_0}\right]\right), \\
&= -\frac{\lambda d}{2} + \mathbb{E}_{\sigma_0, y_{\sigma_0}} \log\left(\frac{1}{n}e^{\sqrt{\lambda d} y_{\sigma_0}} + \frac{n-1}{n}e^{\frac{\lambda d}{2}}\right), \\
&= \mathbb{E}_{\sigma_0, y_{\sigma_0}} \log\left(e^{\sqrt{\lambda d} y_{\sigma_0} - \frac{\lambda d}{2} - d\log 2} + 1 - 2^{-d}\right), \\
&\overset{\text{(a)}}{=} \mathbb{E}_{z_{\sigma_0}} \log\left(e^{\left(\frac{\lambda}{2} - \log 2\right)d + \sqrt{\lambda d} z_{\sigma_0}} + 1 - 2^{-d}\right), \\
&\leq \mathbb{E}_{z\sim\mathcal{N}(0,1)} \log\left(1 + e^{\left(\frac{\lambda}{2} - \log 2\right)d + \sqrt{\lambda d} z}\right).
\end{aligned}$$

In (a) we used $y_{\sigma_0} = \sqrt{\lambda d} + z_{\sigma_0}$. Thus we have (since $1 + e^x \leq 2e^{\max(0,x)}$):

$$\frac{1}{d}F_d(\lambda) \leq \frac{\log 2}{d} + \mathbb{E}_{z\sim\mathcal{N}(0,1)} \max\left\{0, \underbrace{\left(\frac{\lambda}{2} - \log 2\right) + \sqrt{\frac{\lambda}{d}}z}_{=:w_d}\right\}.$$

Since $w_d \to (\lambda/2 - \log 2)$ in probability as $d \to \infty$, and $\mathbb{E}[\max(0, w_d)^2] \leq \mathbb{E}[w_d^2] = \mathcal{O}_d(1)$, we get:

$$\limsup_{d\to\infty} \frac{1}{d}F_d(\lambda) \leq \max\left(0, \frac{\lambda}{2} - \log 2\right).$$

One can easily obtain a corresponding lower bound:

$$\begin{aligned}
\frac{1}{d}F_d(\lambda) &= -\frac{\lambda}{2} + \frac{1}{d}\mathbb{E}_{\mathbf{y}} \log\left(\frac{1}{n}\sum_{\sigma=1}^{n} e^{\sqrt{\lambda d}\, y_\sigma}\right), \\
&\geq -\frac{\lambda}{2} + \frac{1}{d}\mathbb{E}_{\mathbf{y}} \log\left(\frac{1}{n}e^{\sqrt{\lambda d}\, y_{\sigma_0}}\right), \\
&= \frac{\lambda}{2} - \log 2 + \sqrt{\frac{\lambda}{d}}\,\mathbb{E}_{z\sim\mathcal{N}(0,1)}[z], \\
&= \frac{\lambda}{2} - \log 2.
\end{aligned}$$

Recalling that $F_d(\lambda) \geq 0$, we have finally proven the following

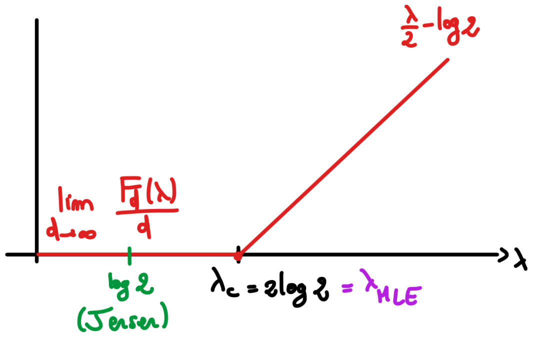
Lemma 2.8

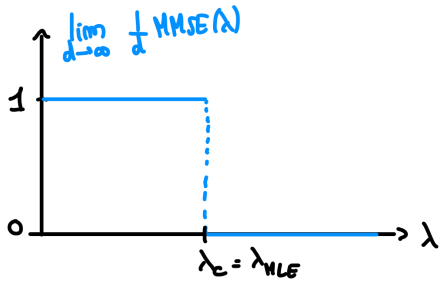
For any $\lambda \geq 0$,

$$\lim_{d\to\infty} \frac{1}{d}F_d(\lambda) = \max\left(0, \frac{\lambda}{2} - \log 2\right).$$

From there we can deduce the behavior of the MMSE. Recall that

$$\frac{1}{d}\text{MMSE}(\lambda) = \mathbb{E}\|\mathbf{e}_{\sigma_0} - \langle \mathbf{e}_\sigma \rangle\|_2^2 = 1 - \mathbb{E}[\|\langle \mathbf{e}_\sigma \rangle\|^2].$$





Corollary 2.9

The asymptotic overlap and asymptototic MMSE satisfy, for all $\lambda \neq \lambda_c$:

$$\lim_{d\to\infty} \frac{1}{d}\mathrm{MMSE}(\lambda) = 1 - \lim_{d\to\infty} \frac{1}{d} Q_d(\lambda) = \mathbb{1}\{\lambda < \lambda_c\}.$$

Proof of Corollary 2.9 – It is a simple consequence of Lemma 2.8 combined with the I-MMSE theorem (Proposition 2.5), and the following classical result of convex analysis

Lemma 2.10

If $f_d : \mathbb{R} \to \mathbb{R}$ is a sequence of convex and differentiable functions, which converge pointwise to a limit f. Then *(i)* f is convex, and *(ii)* for all $t \in \mathbb{R}$ at which f is differentiable[6], we have $f_d'(t) \to f_d(t)$ as $d \to \infty$.

As a remark, recall that any convex function is differentiable everywhere but in a *countable* set of points. □

A first-order phase transition – The results above draw the picture of a sharp transition for recovery of the hidden direction σ_0:

- For $\lambda < \lambda_c := 2\log 2$, one cannot estimate the direction σ_0 better than a random guess, and the asymptotic MMSE is simply the norm of the prior distribution

$$\mathrm{MMSE}(\lambda) = \frac{1}{d}\mathbb{E}[\|\mathbf{X}_0\|^2] - o_d(1) = 1 - o_d(1).$$

- For $\lambda > \lambda_c$, recovery of σ_0 is possible with a probability $1 - o_d(1)$, and an explicit procedure is given by the MLE estimator of eq. (28).

Notice that the asymptotic free entropy has a discontinuous derivative at $\lambda = \lambda_c$: in the physics jargon, this is called a *first-order phase transition*: it corresponds to a discontinuity in the MMSE, and a sharp transition from impossible non-trivial recovery to perfect recovery. On the other hand, a *second-order phase transition* would correspond to a discontinuous second derivative of $F(\lambda)$: in this kind of transitions, the MMSE is continuous at the critical λ_c: we will see examples of both transitions in the following.

Why did naïve Jensen failed ? – The failure of the naïve use of Jensen's inequality is symptomatic of a phenomenon where a random variable X_d[7] can have a seemingly simple behavior, e.g. $X_d \to x$ as $d \to \infty$ in L^2 (for $x \in \mathbb{R}$ a real value), however

$$\lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}[\exp(dX_d)] > \lim_{d\to\infty} \mathbb{E}[X_d] = x. \tag{32}$$

[7] Here $X_d = (1/d)\log \mathcal{Z}_d(\lambda; \mathbf{y})$.

Notice that the LHS of eq. (32) is always greater than the RHS by Jensen's inequality. The strict inequality in eq. (32) can arise if $\mathbb{E}[e^{dX_d}]$ is dominated by rare events, where X_d is much greater than its typical value x. In the Gaussian mean location problem, exemples of such events are

$$\mathcal{E}_\tau \coloneqq \{z_{\sigma_0} \geq \sqrt{\tau d}\}. \tag{33}$$

Clearly, under $\mathcal{N}(0,1)$, $\mathcal{E}_\tau$ has probability $\mathbb{P}(\mathcal{E}_\tau)$ such that $\log \mathbb{P}(\mathcal{E}_\tau) \sim -\frac{\tau d}{2}$ for any fixed $\tau > 0$. While this probability is exponentially small, notice that

$$\begin{aligned}\log \mathbb{E}\mathcal{Z}_d(\lambda;\mathbf{y}) &\geq \log \mathbb{E}\left[\mathcal{Z}_d(\lambda;\mathbf{y})|\mathcal{E}_\tau\right] + \log \mathbb{P}[\mathcal{E}_\tau], \\ &\geq -\frac{\tau d}{2} - d\log 2 + \frac{\lambda d}{2} + \sqrt{\lambda\tau}d + o(d).\end{aligned}$$

Taking $\tau = \lambda$ to maximize this lower bound, we reach that

$$\log \mathbb{E}\mathcal{Z}_d(\lambda;\mathbf{y}) \geq (\lambda - \log 2)d + o(d).$$

What we just showed is that the "annealed" average $\mathbb{E}\mathcal{Z}_d(\lambda;\mathbf{y})$ is actually dominated by the events $\mathcal{E}_\lambda$ of eq. (33), although these events have exponentially small probability. As we later conditioned on z_{σ_0} before applying Jensen's inequality, such spurious events could no longer impact the annealed average.

The following is a sufficient condition for Jensen's inequality to be asymptotically sharp.

Challenge 2.1. *Assume X_d is a real r.v. such that $X_d \to x$ (in probability) as $d \to \infty$, and $|X_d| \leq M$ (a.s.) for some $M > 0$. Show that a sufficient condition for eq. (32) to be an equality is that for all $t > 0$:*

$$\lim_{d\to\infty} \frac{1}{d} \log \mathbb{P}[|X_d - x| \geq t] = -\infty. \tag{34}$$

Eq. (34) 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 $\exp(-\omega(d))$.

2.5 Spiked matrix and spiked tensor models

For much of this class (in the majority of Sections 3,4,5,6), we will consider a specific instance of Gaussian additive models as our toy setting for questions of detection, estimation and optimization. In these models, the observations $\mathbf{Y}$ are given in the form of a matrix or a tensor, and the signal $\mathbf{X}_0$ has a *low-rank structure.*

2.5.1 The spiked Wigner/spiked matrix model

We first introduce the matrix setting of this problem, for which we need to define a Gaussian distribution over symmetric matrices.

Definition 2.4 (*Gaussian Orthogonal Ensemble*)

Let $d \geq 1$. We say that $\mathbf{W} \in \mathbb{S}_d$ is drawn from the *Gaussian orthogonal ensemble* (or GOE(d)) if its elements are drawn independently (up to the symmetry $W_{ij} = W_{ji}$), with

$$\begin{cases} W_{ij} \sim \mathcal{N}(0, 1/d) & \text{for } i < j, \\ W_{ii} \sim \mathcal{N}(0, 2/d). \end{cases} \tag{35}$$

The normalization convention for diagonal and off-diagonal elements in Definition 2.4 implies the nice fact that the probability density of $\mathbf{W}$ can be written (up to a constant) in the compact form:

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

We can now introduce the spiked Wigner (or spiked matrix) model, which is an instance of a Gaussian additive model where the signal is a rank-one matrix.

Definition 2.5 (*Spiked Wigner/Spiked matrix model*)

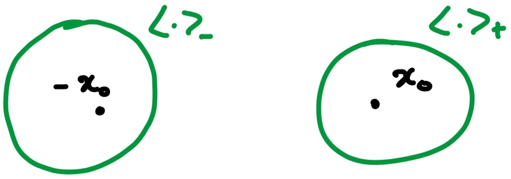
Let $d \geq 1$, $\lambda \geq 0$, and $\mathbf{x}_0 \in \mathbb{R}^d$ be drawn from a prior distribution P_0 over $\mathbb{R}^d$ such that $\mathbb{E}[\|\mathbf{x}\|^2] = d$. We observe $\mathbf{Y} \in \mathbb{S}_d$, the symmetric matrix built as

$$\mathbf{Y} = \frac{\sqrt{\lambda}}{d}\mathbf{x}_0\mathbf{x}_0^\top + \mathbf{W}, \tag{36}$$

where $\mathbf{W} \sim \mathrm{GOE}(d)$.

Remark – The normalization $\mathbb{E}[\|\mathbf{x}_0\|^2] = d$ ensures that the two matrices in eq. (36) have comparable spectral norms as we will discuss in Section 3. Note that this just amounts to a rescaling of λ.

A remark on symmetry – Notice that if P_0 is symmetric around the origin, then the Bayes-optimal estimator of Theorem 2.1 is identically zero by symmetry, as the Gibbs (posterior) measure $\langle\cdot\rangle$ is invariant under reflections $A \to -A$. In particular $\mathbb{E}[\mathbf{x}|\mathbf{Y}] = 0$. Still, the posterior measure might have information about $\mathbf{x}_0$, it just has a global symmetry and can be decomposed into two components.



$$\langle\cdot\rangle = \frac{1}{2}\left[\langle\cdot\rangle_- + \langle\cdot\rangle_+\right]$$

$$\langle g(x)\rangle_+ = \langle g(-x)\rangle_-$$

In the following, we will mostly ignore this problem, and notice that it is usually solved in several ways:

1. Slightly break the symmetry of P_0, e.g. by setting $\mathbb{E}[x_i] = \varepsilon \ll 1$. One takes then the limit $\varepsilon \downarrow 0$ *after* $d \to \infty$.

2. Another similar fix consists in adding a small side information to the model, e.g.

$$\mathbf{y}' = \sqrt{\varepsilon}\mathbf{x}_0 + \mathbf{z},$$

with $\mathbf{z} \sim \mathcal{N}(0, \mathrm{I}_d)$ Gaussian noise, and again $\varepsilon \to 0$ after $d \to \infty$. In both these cases, the assumption is that when taking $\varepsilon \downarrow 0$ after $d \to \infty$, the various expectations we will compute become expectations under $\langle\cdot\rangle_+$.

3. The arguably cleanest approach is simply to consider the estimation of the rank-one matrix $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\top$, e.g. computing the MMSE for $\mathbf{X}_0$ instead of the one

of $\mathbf{x}_0$. Notice that from $\hat{\mathbf{X}}_{\mathrm{opt}}(\mathbf{Y}) = \mathbb{E}[\mathbf{X}|\mathbf{Y}] = \langle \mathbf{X} \rangle$, denoting $(\lambda_{\max}, \mathbf{v}_{\max})$ its top eigenvalue-eigenvector pair, one can build easily an estimator for $\mathbf{x}_0$ (up to a global sign) as

$$\hat{x}(\mathbf{Y}) \coloneqq \sqrt{\lambda_{\max}[\hat{\mathbf{X}}(\mathbf{Y})]}\, \mathbf{v}_{\max}[\hat{\mathbf{X}}(\mathbf{Y})].$$

We refer to [MS24, Section 1.1.2] for more details on this point. The PhD thesis [Mio19] is also a great reference on spiked models.

Further motivations – Let us mention a few motivations behind the spiked Wigner model:

1. **Group synchronization** – In the group synchronization problem, one is given a finite graph $G = (V, E)$ (with $V = [n]$) and a group $\mathcal{G}$. We assign to each edge a group element $g_i \in \mathcal{G}$, and for each edge $(i, j) \in E$ we observe

$$Y_{ij} = g_i g_j^{-1} + \text{noise}.$$

The goal is to recover $\{g_i\}_{i\in[n]}$ from these noisy observations. This has applications in imaging for instance: consider the problem of reconstructing a 3D image from various 2D pictures taken by cameras in different positions. Determining the relative positions of the cameras is then a group synchronization problem with $\mathcal{G} = \mathrm{SO}(3)$. We refer to [Abb+18] for more details. The arguably simplest setting of this problem is $\mathbb{Z}_2$*-synchronization*, where $\mathcal{G} = \mathbb{Z}_2$, $G = K_n$ is the complete graph, and the noise is Gaussian. This corresponds exactly to the spiked Wigner model of eq. (36), with $\mathbf{x}_0 \in \{\pm 1\}^d$!

2. **Sparse PCA** – A model for sparse PCA (i.e. computing a sparse large-variance direction in the data) is the following. Let $\mathbf{x}_0 \in \mathbb{R}^d$ be k-sparse, i.e. $\|\mathbf{x}_0\|_0 = k$. We observe n samples from a Gaussian with a preferred sparse direction:

$$\mathbf{y}_1, \cdots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \mathrm{I}_d + \frac{\sqrt{\lambda}}{k}\mathbf{x}_0\mathbf{x}_0^\top\right).$$

The question is then to recover $\mathbf{x}_0$ from the *empirical covariance matrix*

$$\mathbf{Y} \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top \overset{\mathrm{d}}{=} \left(\mathrm{I}_d + \frac{\sqrt{\lambda}}{k}\mathbf{x}_0\mathbf{x}_0^\top\right)^{1/2} \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^\top \left(\mathrm{I}_d + \frac{\sqrt{\lambda}}{k}\mathbf{x}_0\mathbf{x}_0^\top\right)^{1/2},$$

with $\mathbf{z}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d)$. This is sometimes known as a *spiked Wishart* model. The spiked Wigner model corresponds to a simplification where the low-rank perturbation is additive, and the noise matrix is Wigner instead of Wishart. All the tools we will develop in this class for the spiked Wigner model can be generalized to spiked Wishart models.

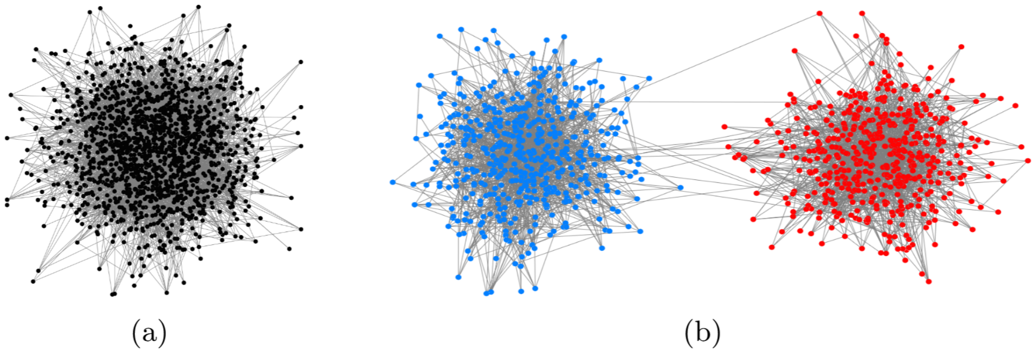
3. **Community detection** – This topic is discussed in detail in [MS23]. Consider a *stochastic block model* (SBM) with two communities: for some $\sigma \in \{\pm 1\}^n$ representing the two communities, onw draws the adjacency matrix $A_{ij} \in \{0, 1\}$ for $i < j$ with independent elements, and

$$\mathbb{P}(A_{ij} = 1|\sigma_i, \sigma_j) = \begin{cases} p_{\mathrm{in}} & \text{if } \sigma_i = \sigma_j, \\ p_{\mathrm{out}} & \text{if } \sigma_i \neq \sigma_j. \end{cases}$$

It is then easy to check that, up to global rank-one change

$$\bar{\mathbf{A}} \coloneqq \mathbf{A} - \frac{p_{\mathrm{in}} + p_{\mathrm{out}}}{2}\mathbf{1}\mathbf{1}^\top = \Delta\boldsymbol{\sigma}\boldsymbol{\sigma}^\top + \mathbf{W},$$

with $\Delta := (p_{\mathrm{in}} - p_{\mathrm{out}})/2$, and $\mathbf{W} := \mathbf{A} - \mathbb{E}[\mathbf{A}]$ is a noise matrix, with independent elements. Replacing the distribution of these elements by i.i.d. centered Gaussians yields again a spiked Wigner model. Beyond [MS23], we refer to [DAM16] for a rigorous connection, and to [BSS23, Section 7.2] for a short introduction to the SBM.



A graph generated from a SBM (a), and the same graph with the communities colored (b). From [BSS23].

2.5.2 Tensor PCA and the spiked tensor model

The spiked Wigner model can be generalized to tensors, i.e. multi-dimensional arrays. It was introduced in [MR14], and we refer to this work for other motivations and its connection to so-called *tensor PCA*. To define the model formally, we we first generalize Definition 2.4 to a notion of symmetric Gaussian tensors.

Definition 2.6 (*Symmetric Gaussian tensor*)

Let $d \geq 1$ and $k \geq 2$. Let $\mathbf{G} \in (\mathbb{R}^d)^{\otimes k}$ with $G_{i_1,\cdots,i_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. For a permutation $\pi \in \mathfrak{S}_k$, $\mathbf{G}^\pi$ is the tensor with indices $G^\pi_{i_1,\cdots,i_k} := G_{i_{\pi(1)},\cdots,i_{\pi(k)}}$. We say that $\mathbf{W} \in (\mathbb{R}^d)^{\otimes k}$ is drawn as a symmetric Gaussian tensor (denoted $\mathbf{W} \sim \mathrm{ST}(k;d)$) if it is distributed as

$$\mathbf{W} = \frac{1}{\sqrt{k!d}} \sum_{\pi \in \mathfrak{S}_k} \mathbf{G}^\pi .$$

Remarks –

(*i*) For $k = 2$ we recover the $\mathrm{GOE}(d)$ distribution: $\mathrm{ST}(2;d) = \mathrm{GOE}(d)$.

(*ii*) For all $i_1 < \cdots < i_k$, we have $W_{i_1 \cdots i_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$.

(*iii*) $\mathbf{W} \sim \mathrm{ST}(k;d)$ is a symmetric tensor: for all $\pi \in \mathfrak{S}_k$, $\mathbf{W}^\pi = \mathbf{W}$.

(*iv*) The distribution $\mathrm{ST}(k;d)$ enjoys a rotation-invariance property. For $\mathbf{O} \in \mathcal{O}(d)$ and $\mathbf{T} \in (\mathbb{R}^d)^{\otimes k}$, we define $(\mathbf{T}\#\mathbf{O})_{i_1,\cdots,i_k} := \sum_{j_1,\cdots,j_k} T_{j_1,\cdots,j_k} O_{i_1 j_1} \cdots O_{i_k j_k}$ the rotation of $\mathbf{T}$ by $\mathbf{O}$. If $\mathbf{W} \sim \mathrm{ST}(k;d)$, then for any $\mathbf{O} \in \mathcal{O}(d)$, $\mathbf{W}\#\mathbf{O} \sim \mathrm{ST}(k;d)$. In the case $k = 2$, for any $\mathbf{W} \sim \mathrm{GOE}(d)$ we have $\mathbf{O}\mathbf{W}\mathbf{O}^\top \sim \mathrm{GOE}(d)$: in particular, the eigenvectors of $\mathbf{W}$ form an orthogonal matrix drawn from the Haar measure on the orthogonal group $\mathcal{O}(d)$, and they are independent of the eigenvalues of $\mathbf{W}$.

We can now introduce the spiked tensor model, the counterpart to Definition 2.7 in the tensor world. Note that we use slightly different normalizations.

Definition 2.7 (*Spiked tensor model*)

Let $d \geq 1$, and $\mathbf{x}_0 \in \mathbb{R}^d$ be drawn from a prior distribution P_0 over $\mathbb{R}^d$. Let $k \geq 1$ and $\mathbf{W} \sim \mathrm{ST}(k; d)$. We observe $\mathbf{Y} \in (\mathbb{R}^d)^{\otimes k}$, the symmetric tensor built as

$$\mathbf{Y} = \mathbf{W} + \sqrt{\lambda}\, \mathbf{x}_0^{\otimes k}.$$

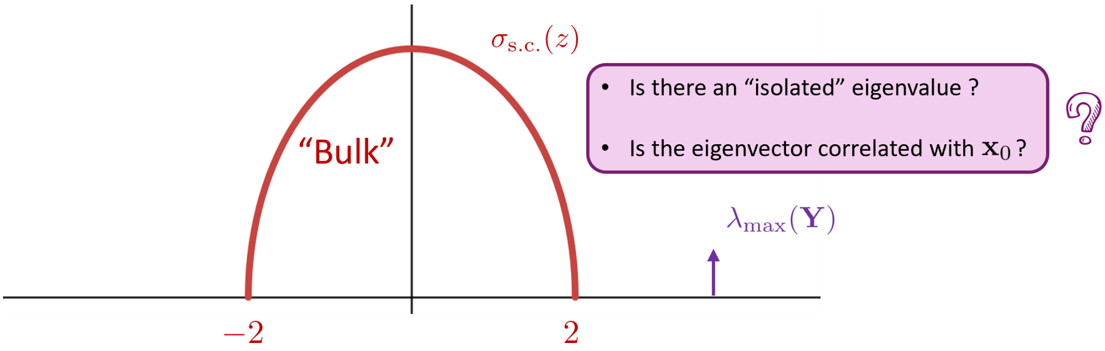

$\sigma_{\text{s.c.}}(z)$
"Bulk"
Is there an "isolated" eigenvalue ?
Is the eigenvector correlated with $\mathbf{x}_0$?
$\lambda_{\max}(\mathbf{Y})$
-2
2

Figure 1: Schematic view of the question we want to answer regarding the model of eq. (37).

3 Spectral algorithms in the spiked matrix model

We consider the spiked Wigner model of Definition 2.5. The statistician is given an observation under the form of a symmetric matrix $\mathbf{Y}$, built as:

$$\mathbf{Y} = \mathbf{W} + \frac{\sqrt{\lambda}}{d}\mathbf{x}_0\mathbf{x}_0^\top \in \mathbb{S}_d$$

In this section, we will assume that $\mathbf{x} = \mathbf{x}_0$ is fixed, and on the Euclidean sphere of radius $\sqrt{d}$. Notice that by rescaling it as $\mathbf{x} \to \mathbf{x}/\sqrt{d}$, it is equivalent to consider

$$\mathbf{Y} = \mathbf{W} + \sqrt{\lambda}\mathbf{x}\mathbf{x}^\top \in \mathbb{S}_d, \tag{37}$$

with $\|\mathbf{x}\| = 1$, i.e. $\mathbf{x} \in \mathcal{S}^{d-1}$. The normalization will be more convenient for this section.

The main goal in Section 3 is to answer the following question:

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

Since $v_{\max}(\mathbf{Y})$ is efficient to compute, this estimator (the *PCA estimator*) already gives us a baseline for efficient recovery of $\mathbf{x}$ in a general spiked Wigner model. Notice that what we will discuss can be generalized for $\mathbf{W}$ beyond Gaussian matrices to other i.i.d. matrices, as well as a large class of matrix distributions that enjoy a rotation-invariance property: see [Mai24, Section 5] for more on this point.

3.1 The asymptotic spectrum of Wigner matrices: reminders

The seminal work of Wigner [Wig55], that can be seen as the start of random matrix theory, proves that the GOE(d) ensemble satisfies the following:

Theorem 3.1 (*Asymptotic spectrum of Wigner matrices*)

Let $\mathbf{W} \sim \mathrm{GOE}(d)$, with eigenvalues $w_1 \geq \cdots \geq w_d$. Then:

(i) The empirical spectral distribution of $\mathbf{W}$ converges[8]:

$$\hat{\mu}_{\mathbf{W}} \coloneqq \frac{1}{d}\sum_{i=1}^{d}\delta_{w_i} \xrightarrow[d\to\infty]{\text{weakly}} \sigma_{\text{s.c.}} \quad \text{(a.s.)},$$

where $\sigma_{\text{s.c.}}$ is called Wigner's *semicircle* law

$$\sigma_{\text{s.c.}}(\mathrm{d}x) \coloneqq \frac{\sqrt{4-x^2}}{2\pi}\mathbb{1}\{|x| \leq 2\}\mathrm{d}x. \tag{38}$$

(ii) The top eigenvalue of $\mathbf{W}$ converges to the right edge of the support of $\sigma_{\text{s.c.}}$:

$$w_1 = \max_{i\in[d]} z_i \xrightarrow[d\to\infty]{} 2 \quad \text{(a.s.)}$$

3.1.1 The bulk of Wigner matrices: sketch of proof

We sketch here a proof of Theorem 3.1-(i) using the *Stieltjes/Cauchy transform*, or *resolvent*, method. As this result is very classical, we only aim to present the main ideas, and we refer to [AGZ10; Kun25] for mathematical proofs. The resolvent method is very powerful and will play a crucial role in the spectral analysis of the spiked model.

Definition 3.1 (*Resolvent and Cauchy transform*)

For a matrix $\mathbf{M} \in \mathbb{S}_d$, we define its resolvent $\mathbf{R}_{\mathbf{M}}(z)$ and Cauchy transform $G_{\mathbf{M}}(z)$ as follows:

$$\begin{cases} \mathbf{R}_{\mathbf{M}}(z) & := (z\mathrm{I}_d - \mathbf{M})^{-1}, \\ G_{\mathbf{M}}(z) & := (1/d)\mathrm{Tr}[\mathbf{R}(z)], \end{cases}$$

for any $z \in \mathbb{C}\backslash\mathrm{Sp}(\mathbf{M})$. $z \mapsto -G_{\mathbf{M}}(z)$ is usually called the *Stieltjes* transform.

More generally, one can define the Cauchy transform of any real probability measure as

Definition 3.2 (*Cauchy transform*)

For any $\mu \in \mathcal{P}(\mathbb{R})$ and $z \in \mathbb{C}\backslash\,\mathrm{supp}(\mu)$, we define the Cauchy transform as:

$$G_\mu(z) := \mathbb{E}_{X\sim\mu}[(z - X)^{-1}].$$

The Cauchy transform enjoys remarkable properties: in particular it fully characterizes the associated probability measure as this next theorem shows. We refer to [AGZ10, Section 2.4] for more properties, and their associated proofs.

Proposition 3.2 (*Properties of the Cauchy transform*)

If $(\mu_n)_{n\geq 1}$ and μ are real probability measures, then

$$\mu_n \xrightarrow[n\to\infty]{\text{(w.)}} \mu \Leftrightarrow \lim_{n\to\infty} G_{\mu_n}(z) = G_\mu(z)\ \forall z \in \mathbb{C}\backslash\mathbb{R}.$$

We will sketch here a proof that $G_{\mathbf{W}}(z) \to G_{\text{s.c.}}(z)$, for any $z \in \mathbb{C}\backslash\mathbb{R}$ and as $d \to \infty$, which will thus imply Theorem 3.1-(i). A complete proof is available in [AGZ10, Section 2.4]. We start with this simple property.

Lemma 3.3

The Stieltjes transform $G_{\text{s.c.}}$ of the semicircle law of eq. (38) satisfies, for all $t > 2$:

$$G_{\text{s.c.}}(t) = \frac{t - \sqrt{t^2 - 4}}{2}. \tag{39}$$

[8] 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.

Challenge 3.1. *Prove Lemma 3.3. (Hint: try to write it as an integral over the complex unit circle, and use the residue theorem)*

The crux of the proof is a leave-one-out argument (also called “cavity method” in statistical physics, we will revisit this later on!). Notice that

$$G_{\mathbf{W}}(z) = \frac{1}{d}\sum_{i=1}^{d}[z\mathrm{I}_d - \mathbf{W}]^{-1}_{ii}.$$

The matrix element of this inverse can be expressed using the Schur complement formula:

$$\begin{pmatrix} a & \mathbf{b}^\top \\ \mathbf{b} & \mathbf{C} \end{pmatrix}^{-1}_{11} = \frac{1}{a - \mathbf{b}^\top \mathbf{C}^{-1}\mathbf{b}}, \tag{40}$$

for any $a, \mathbf{b}, \mathbf{C}$ (symmetric) such that these quantities are well-defined. Using eq. (40):

$$G_{\mathbf{W}}(z) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{(z - W_{ii}) - \widetilde{\mathbf{w}}_i \cdot (z\mathrm{I}_{d-1} - \mathbf{W}_{-i})^{-1}\widetilde{\mathbf{w}}_i}. \tag{41}$$

Up to now, our derivation was exact. We now give the sketch of the rest of the proof at a very heuristic level: the rigorous derivation follows exactly the same lines, using precise concentration inequalities in several steps. Notice that $W_{ii} = \Theta(1/\sqrt{d})$, so we simplify eq. (41) to leading order as $d \to \infty$ as:

$$G_{\mathbf{W}}(z) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{z - \widetilde{\mathbf{w}}_i \cdot (z\mathrm{I}_{d-1} - \mathbf{W}_{-i})^{-1}\widetilde{\mathbf{w}}_i}. \tag{42}$$

Here $\mathbf{W}_{-i}$ is the $(d-1)\times(d-1)$ matrix with i-th row and column removed, and $\widetilde{\mathbf{w}}_i \in \mathbb{R}^{d-1}$ is the i-th row of $\mathbf{W}$ with the i-th element removed. The crucial remark is that $\widetilde{\mathbf{w}}_i$ *is independent of* $\mathbf{W}_{-i}$! Therefore by concentration of measure (see Appendix A.4 e.g.,) we have

$$\widetilde{\mathbf{w}}_i \cdot (z\mathrm{I}_{d-1} - \mathbf{W}_{-i})^{-1}\widetilde{\mathbf{w}}_i \simeq \mathbb{E}\widetilde{\mathbf{w}}_i \cdot (z\mathrm{I}_{d-1} - \mathbf{W}_{-i})^{-1}\widetilde{\mathbf{w}}_i = \frac{1}{d}\mathrm{Tr}[(z\mathrm{I}_{d-1} - \mathbf{W}_{-i})^{-1}].$$

Plugging it back in eq. (42), since all elements of the sum have the same law, and using again concentration of measure, we expect:

$$\begin{aligned} G_{\mathbf{W}}(z) &\simeq \mathbb{E}_{\mathbf{W}} G_{\mathbf{W}}(z), \\ &\simeq \frac{1}{z - \frac{1}{d}\mathbb{E}\mathrm{Tr}[(z\mathrm{I}_{d-1} - \mathbf{W}_{-1})^{-1}]}, \\ &\simeq \frac{1}{z - \mathbb{E}G_{\mathbf{W}^{(d-1)}}(\mathbf{z})}. \end{aligned}$$

This heuristic derivations suggests that $G_{\mathbf{W}}(z) \to G(z)$ for $G(z)$ a solution to

$$G(z) = \frac{1}{z - G(z)}. \tag{43}$$

One checks then easily from Lemma 3.3 that $G_{\mathrm{s.c.}}(z)$ is the only solution to eq. (43) such that $G(z) \to 0$ as $|z| \to \infty$. □

3.1.2 The top eigenvalue of Wigner matrices

We give here a proof of point (ii) of Theorem 3.1. First notice that

$$\{w_1 < 2 - \varepsilon\} \Rightarrow \hat{\mu}_{\mathbf{W}}([2-\varepsilon, 2]) = 0.$$

Using point (i) of Theorem 3.1, we reach that, almost surely:

$$\liminf_{d\to\infty} w_1 \geq 2. \tag{44}$$

The upper bound can be obtained in several steps. The first is to use Sudakov-Fernique's inequality, see Lemma A.5, to control $\mathbb{E}[w_1]$. Indeed, notice that

$$w_1 = \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathbf{W}\mathbf{x}.$$

Let $X(\mathbf{x}) \coloneqq (\sqrt{d/2})\, \mathbf{x}^\top \mathbf{W}\mathbf{x}$. Then X is a Gaussian process (indexed by the unit sphere $\mathcal{S}^{d-1}$). Define $Y(\mathbf{x}) \coloneqq \sqrt{2}\,(\mathbf{x}\cdot\mathbf{g})$. for $\mathbf{g} \sim \mathcal{N}(0, \mathrm{I}_d)$. Then we have, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{S}^{d-1}$:

$$\begin{cases} \mathbb{E}[X(\mathbf{x})] & = \mathbb{E}[Y(\mathbf{x})] = 0, \\ \mathbb{E}[(X(\mathbf{x}) - X(\mathbf{x}'))^2] & = 2[1 - (\mathbf{x}\cdot\mathbf{x}')^2], \\ \mathbb{E}[(Y(\mathbf{x}) - Y(\mathbf{x}'))^2] & = 4[1 - (\mathbf{x}\cdot\mathbf{x}')]. \end{cases}$$

Since $1 - q^2 \leq 2(1-q)$ for all $q \in [-1, 1]$, applying Lemma A.5, we get:

$$\begin{aligned} \sqrt{\frac{d}{2}}\,\mathbb{E}[w_1] &= \mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{S}^{d-1}} X(\mathbf{x})\right], \\ &\leq \mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{S}^{d-1}} Y(\mathbf{x})\right], \\ &= \sqrt{2}\mathbb{E}[\|\mathbf{g}\|], \\ &\leq \sqrt{2\mathbb{E}[\|\mathbf{g}\|^2]}, \\ &= \sqrt{2d}. \end{aligned}$$

We showed $\mathbb{E}[w_1] \leq 2$.

Next, denote Z_{ij} for $i \leq j$ be the i.i.d. $\mathcal{N}(0,1)$ random variables such that $W_{ij} = W_{ji} = Z_{ij}/\sqrt{d}$, and $W_{ii} = (\sqrt{2/d})\, Z_{ii}$. For any $\mathbf{W}, \mathbf{W}'$ (and associated $\mathbf{Z}, \mathbf{Z}'$), we have

$$\|\lambda_{\max}(\mathbf{W} - \mathbf{W}')\|^2 \leq \|\mathbf{W} - \mathbf{W}'\|_F^2 = \frac{2}{d}\sum_{i\leq j}(Z_{ij} - Z'_{ij})^2.$$

Stated differently, $\mathbf{Z} \mapsto \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathbf{W}\mathbf{x}$ is $(\sqrt{2/d})$-Lipschitz. We can thus leverage Gaussian concentration (Theorem A.8), which gives for any $t \geq 0$:

$$\mathbb{P}\left(|w_1 - \mathbb{E}[w_1]| \geq t\right) \leq 2\exp\left\{-\frac{dt^2}{4}\right\}.$$

Combining it with the bound $\mathbb{E}[w_1] \leq 2$, we reach

$$\mathbb{P}\left(w_1 \geq 2 + t\right) \leq 2\exp\left\{-\frac{dt^2}{4}\right\}.$$

By the Borel-Cantelli lemma (Lemma A.1), almost surely:

$$\limsup_{d\to\infty} w_1 \leq 2. \tag{45}$$

Combining eqs. (44) and (45) ends the proof. □

3.2 Emergence of a single outlier

The following proposition shows that the eigenvalues of $\mathbf{W}$ and $\mathbf{Y}$ are interlaced.

Proposition 3.4 (*Interlacing*)

Let $\lambda \geq 0$, and $\mathbf{W} \in \mathbb{S}_d$, $\mathbf{x} \in \mathcal{S}^{d-1}$. Let $\mathbf{Y} = \mathbf{W} + \sqrt{\lambda}\mathbf{x}\mathbf{x}^\top$. Denote $y_1 \geq \cdots y_d$ and $w_1 \geq \cdots w_d$ the eigenvalues of $\mathbf{Y}$ and $\mathbf{W}$. Then

(i) $w_1 \leq y_1$.

(ii) $w_i \leq y_i \leq w_{i-1}$ for all $i \in \{2, \cdots, d\}$.

Proof of Proposition 3.4 – The lower bound on y_i (for $i \in [d]$) is trivial, since $\sqrt{\lambda}\mathbf{x}\mathbf{x}^\top \succeq 0$. Let us denote $\mathbf{u}_1, \cdots, \mathbf{u}_d$ the eigenvectors of $\mathbf{W}$. The lower bound on y_i for $i \geq 2$ follows from the Courant-Fischer characterization of eigenvalues:

$$y_i = \max_{\dim(V)=i} \min_{\substack{\mathbf{v}\in V \\ \|\mathbf{v}\|=1}} \mathbf{v}^\top \mathbf{Y}\mathbf{v}.$$

Since $i \geq 2$, any subspace $V \subseteq \mathbb{R}^d$ with dimension i must contain a non-zero vector $\mathbf{v}$ orthogonal to $\mathrm{Span}(\mathbf{x}, \{\mathbf{u}_j\}_{j\leq i-2})$. Thus

$$\mathbf{v}^\top \mathbf{Y}\mathbf{v} = \mathbf{v}^\top \mathbf{W}\mathbf{v} \overset{(a)}{\leq} w_{i-1},$$

where (a) comes from $\mathbf{v}$ being orthogonal to $\{\mathbf{u}_j\}_{j\leq i-2}$. □

Proposition 3.4 is a special case of Weyl's interlacing inequality. It implies that the "bulk" of eigenvalues of $\mathbf{Y}$ and $\mathbf{W}$ behave similarly as $d \to \infty$.

Corollary 3.5

For $\mathbf{Y}$ as in Proposition 3.4, the empirical distribution of $\mathbf{Y}$ converges:

$$\hat{\mu}_{\mathbf{Y}} \coloneqq \frac{1}{d}\sum_{i=1}^d \delta_{y_i} \xrightarrow[d\to\infty]{\text{weakly}} \sigma_{\text{s.c.}} \qquad \text{(a.s.)},$$

More specifically, all $y_2 \geq \cdots y_d$ will (with high probability) lie in the interval $[-2 - o(1), 2 + o(1)]$. Thus it is *only* $y_1 = \lambda_{\max}(\mathbf{Y})$ that might be an outlier, see Fig. 1.

A simple bound – Notice that $y_1 = \max_{\|\mathbf{v}\|=1} \mathbf{v}^\top \mathbf{Y}\mathbf{v} \geq \mathbf{x}^\top \mathbf{Y}\mathbf{x} = \mathbf{x}^\top \mathbf{W}\mathbf{x} + \sqrt{\lambda}$. Furthermore, it is easy to see that, for any $\mathbf{x} \in \mathcal{S}^{d-1}$, $z \coloneqq \mathbf{x}^\top \mathbf{W}\mathbf{x} \sim \mathcal{N}(0, 2/d)$. In particular $\mathbb{P}(z \geq t) \leq \exp\{-dt^2/4\}$ for any $t \geq 0$. This can be easily shown to imply (via the Borel-Cantelli lemma) that

$$\liminf_{d\to\infty} y_1 \geq \sqrt{\lambda}. \qquad \text{(a.s.)} \tag{46}$$

In particular, if $\lambda > 4$, then $\liminf y_1 > 2$ (a.s.): we see an outlier in the spectrum of $\mathbf{Y}$! Further, if $\mathbf{v}_1$ is the top eigenvector of $\mathbf{Y}$, we have

$$y_1 = \mathbf{v}_1^\top \mathbf{Y}\mathbf{v}_1 \leq w_1 + \sqrt{\lambda}(\mathbf{v}_1 \cdot \mathbf{x})^2.$$

Using Theorem 3.1 and eq. (46), we have for $\lambda > 2$:

$$\liminf_{d\to\infty} (\mathbf{v}_1 \cdot \mathbf{x})^2 \geq 1 - \frac{2}{\sqrt{\lambda}}. \qquad \text{(a.s.)} \tag{47}$$

So the outlier is associated to an eigenvector which correlates positively with $\mathbf{x}$! Further, this correlation goes to 1 as $\lambda \to \infty$. But are eqs. (46),(47) sharp? We saw that $\lambda > 4$ is sufficient for an outlier to appear in the spectrum, with an eigenvector positively correlated with $\mathbf{x}$: is this also *necessary?*

3.3 The BBP transition

The following theorem is the main result of this section. It is usually refered to as the *Baik-Ben Arous-Péché* (BBP) transition, from the authors of [BBP05], and it provides a sharp answer to the question above. While the authors of [BBP05] analyzed a spiked version of covariance matrices (see the discussion on spiked Wishart models in Section 2.5), the statement for the spiked Wigner model can be found in [FP07][9], and a much generalized version in [BN11].

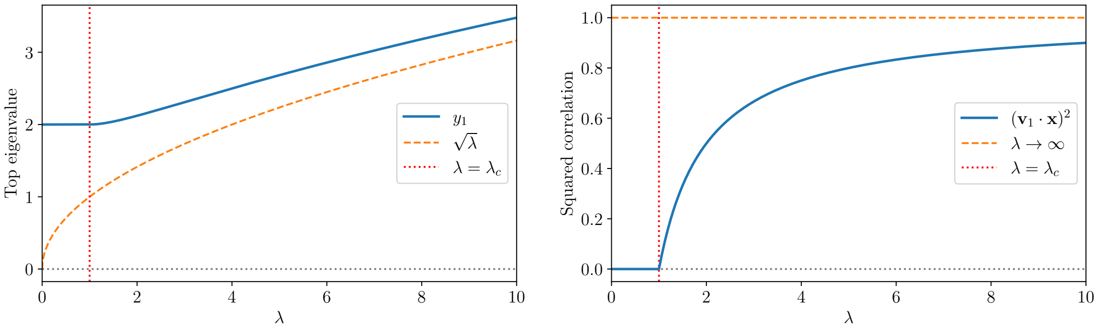
Theorem 3.6 (*The BBP transition in the spiked Wigner model*)

Let $d \geq 1$ and $\lambda > 0$. Let $\mathbf{x} \in \mathcal{S}^{d-1}$ an arbitrary unit-norm vector. We draw $\mathbf{Y} = \mathbf{W} + \sqrt{\lambda}\mathbf{x}\mathbf{x}^\top$ with $\mathbf{W} \sim \mathrm{GOE}(d)$. Denote $y_1 \geq \cdots \geq y_d$ the eigenvalues of $\mathbf{Y}$, and $\mathbf{v}_1, \cdots, \mathbf{v}_d$ a set of corresponding eigenvectors (unit-normed). Then:

(i) If $\lambda \leq 1$, then $y_1 \xrightarrow[d\to\infty]{\text{(a.s.)}} 2$, and $(\mathbf{v}_1 \cdot \mathbf{x})^2 \xrightarrow[d\to\infty]{\text{(a.s.)}} 0$.

(ii) If $\lambda > 1$, then $y_1 \xrightarrow[d\to\infty]{\text{(a.s.)}} \lambda^{1/2} + \lambda^{-1/2}$, and $(\mathbf{v}_1 \cdot \mathbf{x})^2 \xrightarrow[d\to\infty]{\text{(a.s.)}} 1 - \lambda^{-1}$.

Moreover, for any $\lambda > 0$, $y_2 \xrightarrow[d\to\infty]{\text{(a.s.)}} 2$.



3.3.1 Discussion

Theorem 3.6 shows several key features:

1. There is sharp *phase transition* at $\lambda_c = 1$: the model behaves very differently for $\lambda < \lambda_c$ and $\lambda > \lambda_c$!

2. The sufficient condition $\lambda > 4$ to have an outlier that we derived in Section 3.2 is not sharp. What happens is that for $\lambda = 1 + \varepsilon$, the top eigenvector is very slightly correlated with $\mathbf{x}$, but not enough to make $\mathbf{x}^\top\mathbf{Y}\mathbf{x}$ be dominated by the rank-one perturbation.

3. Notice that $(y_1, \mathbf{v}_1)$ are *inconsistent* estimators of $(\sqrt{\lambda}, \pm\mathbf{x})$ even if $\lambda > 1$. Indeed, for any such λ, $y_1 \to \sqrt{\lambda} + 1/\sqrt{\lambda} > \sqrt{\lambda}$, and $\max_{\varepsilon\in\{\pm1\}} \|\mathbf{v}_1 - \varepsilon\mathbf{x}\|_2 \not\to 0$. Instead, the distance $\max_{\varepsilon\in\{\pm1\}} \|\mathbf{v}_1 - \varepsilon\mathbf{x}\|_2 \to 2[1 - \sqrt{1 - \lambda^{-1}}] \sim \lambda^{-1}$ is finite, and goes to 0 only as $\lambda \to \infty$.

[9] The authors of [BBP05; FP07] analyzed much more detailed properties of the fluctuations of the top eigenvalue, not just the value of its limit as stated here.

3.3.2 Proof of Theorem 3.6: eigenvalue transition

We prove here the statements of Theorem 3.6 related to the eigenvalues y_i, by simplifying the proof of [BN11]. Notice first that the statement on y_2 is a direct consequence of our analysis in Section 3.2, so we focus on the statement concerning y_1.

Like the proof of Theorem 3.1 that we discussed in Section 3.1.1, a possible proof is based on the analysis of the Cauchy, or Stieltjes, transform of probability measures, see Definition 3.1.

Denote $w_1 \geq \cdots w_d$ the eigenvalues of $\mathbf{W}_d$, with a set of corresponding eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_d$. By the remark below Definition 2.6, $(\mathbf{u}_1, \cdots, \mathbf{u}_d)$ is an orthogonal matrix uniformly drawn from the Haar measure on $\mathcal{O}(d)$, and is independent of $(w_1, \cdots, w_d)$.

Recall that y_1 is the largest eigenvalue of $\mathbf{Y}$. In the following we sometimes denote it $y_1^{(d)}$ to clarify its dependency on the dimension. We now use the fact that eigenvalues are roots of the characteristic polynomial, so $y_1^{(d)}$ is a solution to:

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

Furthermore, $y_1 \geq \mathbf{u}_1^\top \mathbf{W} \mathbf{u}_1 = w_1 + \sqrt{\lambda}(\mathbf{u}_1 \cdot \mathbf{x})^2$, so $y_1 > w_1$ with probability 1 since $\mathbf{u}_1 \sim \mathrm{Unif}(\mathcal{S}^{d-1})$. In particular, $(y_1\mathrm{I}_d - \mathbf{W})$ is almost surely invertible, and we reach:

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

i.e. 1 is an eigenvalue of $\sqrt{\lambda}\mathbf{x}\mathbf{x}^\top(y\mathrm{I}_d - \mathbf{W})^{-1}$. This is a rank-one matrix, so it has a single non-zero eigenvalue, which is also equal to its trace. Combining this fact with Proposition 3.4, this yields that $y = y_1^{(d)}$ is the only solution in (w_1, ∞) to the equation

$$\frac{1}{\sqrt{\lambda}} = \mathbf{x}^\top (y\mathrm{I}_d - \mathbf{W})^{-1}\mathbf{x}. \tag{48}$$

We can decompose $\mathbf{x} = \sum_{i=1}^d \alpha_i \mathbf{u}_i$ along the eigenbasis of $\mathbf{x}$. Because $\|\mathbf{x}\|_2 = 1$ and $\mathbf{x}$ is independent of $\mathbf{W}$, $\boldsymbol{\alpha} := (\alpha_1, \cdots, \alpha_d)$ is uniformly sampled from the unit sphere $\mathcal{S}^{d-1}$ and . independent of $(w_1, \cdots, w_d)$. Eq. (48) reads:

$$\frac{1}{\sqrt{\lambda}} = \sum_{i=1}^d \frac{\alpha_i^2}{y - w_i}. \tag{49}$$

Let us denote

$$\nu_d := \sum_{i=1}^d \alpha_i^2 \delta_{w_i} \in \mathcal{P}(\mathbb{R}). \tag{50}$$

Eq. (49) can be reframed by saying that $y_1^{(d)}$ is the unique zero in (w_1, ∞) of the function

$$M_d(y) := 1 - \sqrt{\lambda} G_{\nu_d}(y), \tag{51}$$

with G_{ν_d} the Cauchy transform of ν_d.

Notice that $\mathbb{E}[\nu_d] = \mathbb{E}[\mu_{\mathbf{W}}] \to \sigma_{\text{s.c.}}$ as $d \to \infty$ by Theorem 3.1. The next lemma crucially shows that, by concentration of measure, one can essentially replace ν_d by $\sigma_{\text{s.c.}}$ as $d \to \infty$.

Lemma 3.7 (*Convergence of ν_d*)

(i) $\nu_d \xrightarrow[d\to\infty]{\text{(a.s.)}} \sigma_{\text{s.c.}}$, for the weak convergence of probability measures.

(ii) For all $\eta > 0$, $G_{\nu_d}(z) \xrightarrow[d\to\infty]{\text{(a.s.)}} G_{\text{s.c.}}(z)$, uniformly on $K_\eta := \{z \in \mathbb{C} : \mathrm{d}(z, [-2, 2]) \geq \eta\}$.

The following properties of $G_{\text{s.c.}}(z)$ are elementary consequences of Lemma 3.3 and left to show as an exercise:

Proposition 3.8

Let $\lambda \geq 0$, and $M_\lambda(z) := 1 - \sqrt{\lambda} G_{\text{s.c.}}(z)$ for $z \in \mathbb{C} \backslash [-2, 2]$. Then, $y \mapsto M_\lambda(y)$ is strictly increasing on $(2, \infty)$, with $M_\lambda(\infty) = 1$, and $M_\lambda(2^+) = 1 - \sqrt{\lambda}$. For $\lambda > 1$, we denote $y_\star(\lambda)$ the unique zero of $M_\lambda(y)$ on $(2, \infty)$. Then:

(i) $y_\star(\lambda) = \lambda^{1/2} + \lambda^{-1/2}$.

(ii) $y_\star(\lambda)$ is a simple root of $M_\lambda(z)$.

Finally, the next lemma (borrowed from [BN11] and tailored for our setting), follows from considerations in complex analysis.

Lemma 3.9

Let $(a_d, b_d)_{d \geq 1}$ such that $\lim_{d\to\infty} a_d = -2$, $\lim_{d\to\infty} b_d = 2$, and $N_d(z)$ be an analytic function of z defined on $\mathbb{C} \backslash [a_d, b_d]$, and such that:

(i) For all $d \geq 1$ and $z \in \mathbb{C} \backslash \mathbb{R}$, $N_d(z) \neq 0$.

(ii) For all $\eta > 0$, $N_d(z) \to M_\lambda(z)$, uniformly on $K_\eta := \{z \in \mathbb{C} : \mathrm{d}(z, [-2, 2]) \geq \eta\}$.

Then, if $\lambda > 1$, there exists a real sequence $(\gamma_d)_{d \geq 1}$ such that $\gamma_d > b_d$, and:

(a) $\gamma_d \to y_\star(\lambda)$ as $d \to \infty$.

(b) γ_d is a simple root of N_d.

(c) For all $\varepsilon > 0$ small enough and $d \geq 1$ large enough,

$$\forall y \in (2 + \varepsilon, \infty), \quad N_d(y) = 0 \Leftrightarrow y = \gamma_d.$$

Further, if $\lambda \leq 1$, then any $(\gamma_d)_{d \geq 1}$ such that $\gamma_d > b_d$ and $N_d(\gamma_d) = 0$ must satisfy $\gamma_d \to 2$ as $d \to \infty$.

We defer the proofs of Lemma 3.7 and 3.9 to Section 3.3.4.

We know that $y_1^{(d)}$ is the unique zero of $M_d(y)$ on (w_1, ∞) and that $w_1 \xrightarrow[d\to\infty]{\text{(a.s.)}} 2$. Lemma 3.7-$(ii)$ shows then that one can apply Lemma 3.9 to $N_d = M_d$ given by eq. (51), and we reach that

- If $\lambda \leq 1$, $y_1^{(d)} \xrightarrow[d\to\infty]{\text{(a.s.)}} 2$.
- If $\lambda > 1$, $y_1^{(d)} \xrightarrow[d\to\infty]{\text{(a.s.)}} y_\star(\lambda) > 2$. □

3.3.3 Proof of Theorem 3.6: eigenvector correlation

We now consider the correlation of the top eigenvector $\mathbf{v}_1$ (associated with the eigenvalue y_1) with the signal $\mathbf{x}$. By definition:

$$(y_1 \mathrm{I}_d - \mathbf{W})\mathbf{v}_1 = \sqrt{\lambda}(\mathbf{v}_1 \cdot \mathbf{x})\mathbf{x}.$$

As we argued above, $(y_1 \mathrm{I}_d - \mathbf{W})$ is almost surely invertible, which yields:

$$\mathbf{v}_1 = \sqrt{\lambda}(\mathbf{v}_1 \cdot \mathbf{x})(y_1 \mathrm{I}_d - \mathbf{W})^{-1}\mathbf{x}.$$

While this equation still involves $\mathbf{v}_1$ on both sides, since $\|\mathbf{v}_1\| = 1$ we have:

$$\mathbf{v}_1 = \pm \frac{(y_1 \mathrm{I}_d - \mathbf{W})^{-1}\mathbf{x}}{\sqrt{\mathbf{x}^\top (y_1 \mathrm{I}_d - \mathbf{W})^{-2}\mathbf{x}}}.$$

And in particular:

$$(\mathbf{v}_1 \cdot \mathbf{x})^2 = \frac{\left(\mathbf{x}^\top (y_1 \mathrm{I}_d - \mathbf{W})^{-1}\mathbf{x}\right)^2}{\mathbf{x}^\top (y_1 \mathrm{I}_d - \mathbf{W})^{-2}\mathbf{x}}.$$

By eq. (48), we can further simplify it into:

$$(\mathbf{v}_1 \cdot \mathbf{x})^2 = \left(\lambda \mathbf{x}^\top (y_1 \mathrm{I}_d - \mathbf{W})^{-2}\mathbf{x}\right)^{-1}. \tag{52}$$

We now analyze the limit as $d \to \infty$ of eq. (52) in a very similar way to what we did to analyze the limit of $\mathbf{x}^\top (y_1 \mathrm{I}_d - \mathbf{W})^{-1}\mathbf{x}$ above. We separate the cases $\lambda \le 1$ and $\lambda > 1$.

$\boldsymbol{\lambda} > \mathbf{1}$ – Using the same notations as in eqs. (49) and (50):

$$\mathbf{x}^\top (y_1 \mathrm{I}_d - \mathbf{W})^{-2}\mathbf{x} = \sum_{i=1}^d \frac{\alpha_i^2}{(y_1 - w_i)^2} = \int \frac{\nu_d(\mathrm{d}w)}{(y_1 - w)^2}.$$

Since $y_1 \xrightarrow[d\to\infty]{\text{(a.s.)}} y_\star(\lambda) > 2$, and $\nu_d \xrightarrow[d\to\infty]{\text{(a.s.)}} \sigma_{\text{s.c.}}$ by Lemma 3.7-(i), we immediately obtain

$$(\mathbf{v}_1 \cdot \mathbf{x})^{-2} = \lambda \int \frac{\nu_d(\mathrm{d}w)}{(y_1 - w)^2} \xrightarrow[d\to\infty]{\text{(a.s.)}} \lambda \int \frac{\rho_{\text{s.c.}}(\mathrm{d}w)}{(y_\star(\lambda) - w)^2} = -\lambda G'_{\text{s.c.}}[y_\star(\lambda)].$$

Since $y_\star(\lambda) = \lambda^{1/2} + \lambda^{-1/2}$, and by Lemma 3.3, we get $\lambda G'_{\text{s.c.}}[y_\star(\lambda)] = -(1 - \lambda^{-1})^{-1}$.

$\boldsymbol{\lambda} \le \mathbf{1}$ – Notice that $G'_{\text{s.c.}}(2^+) = -\infty$, i.e.

$$\int \frac{\rho_{\text{s.c.}}(\mathrm{d}w)}{(2 - w)^2} = +\infty.$$

Since $y_1 \xrightarrow[d\to\infty]{\text{(a.s.)}} 2$ and $\nu_d \xrightarrow[d\to\infty]{\text{(a.s.)}} \sigma_{\text{s.c.}}$ by Lemma 3.7-(i):

$$\widetilde{\nu}_d \coloneqq \sum_{i=1}^d \alpha_i^2 \delta_{w_i + 2 - y_1} \xrightarrow[d\to\infty]{\text{(a.s.)}} \sigma_{\text{s.c.}}.$$

Again, convergence is meant in the sense of weak convergence of probability measures. Thus, almost surely:

$$\liminf_{d\to\infty} (\mathbf{v}_1 \cdot \mathbf{x})^{-2} = \liminf_{d\to\infty} \lambda \int \frac{\widetilde{\nu}_d(\mathrm{d}w)}{(2 - w)^2} \overset{\text{(a)}}{\ge} \lambda \int \frac{\rho_{\text{s.c.}}(\mathrm{d}w)}{(2 - w)^2} = +\infty,$$

using Fatou's lemma in (a). □

3.3.4 Proof of Theorem 3.6: auxiliary results

We prove here the technical Lemmas 3.7 and 3.9.

Proof of Lemma 3.7 – We start with (i). Let f be a continuous bounded function on $\mathbb{R}$. By concentration of measure (Theorem A.9), for any $\mathbf{x} \in \mathbb{R}^d$ and any $t > 0$:

$$\mathbb{P}_{\boldsymbol{\alpha}}\left[\left|\sum_{i=1}^d \alpha_i^2 x_i - \frac{1}{d}\sum_{i=1}^d x_i\right| \geq t\right] \leq 2\exp\left\{-\frac{cdt^2}{\|\mathbf{x}\|_\infty^2}\right\},$$

for some universal constant $c > 0$. Indeed, if $g(\boldsymbol{\alpha}) \coloneqq \sum_i \alpha_i^2 x_i$, then $\|\nabla g(\boldsymbol{\alpha})\|_2 \leq 2\|\mathbf{x}\|_\infty$ for any $\boldsymbol{\alpha} \in \mathcal{S}^{d-1}$. Using the Borel-Cantelli lemma, and combining it with Theorem 3.1 which implies $(1/d)\sum_{i=1}^d f(w_i) \to \int \sigma_{\rm s.c.}(\mathrm{d}w)\, f(w)$ almost surely as $d \to \infty$, we obtain

$$\int f(w)\nu_d(\mathrm{d}w) = \sum_{i=1}^d \alpha_i^2\, f(w_i) \xrightarrow[d\to\infty]{\text{(a.s.)}} \int \rho_{\rm s.c.}(\mathrm{d}w)\, f(w),$$

which proves point (i).

We turn to point (ii). Let $\eta > 0$. Since $w_1 \overset{\text{a.s.}}{\to} 2$ and $w_d \overset{\text{a.s.}}{\to} -2$ as $d \to \infty$,

$$G_{\nu_d}(z) = \sum_{i=1}^d \frac{\alpha_i^2}{z - w_i}$$

are a.s. uniformly bounded and Lipschitz on K_η. By the Arzelà-Ascoli theorem, any subsequence of (G_{ν_d}) must admit a subsequence that is uniformly convergent on K_η. Moreover, for any $z \in K_\eta$, $G_{\nu_d}(z) \to G(z)$ (a.s.) by point (i). This implies the almost-sure convergence of $G_{\nu_d}(z)$ to $G(z)$ holds uniformly over $z \in K_\eta$. □

Proof of Lemma 3.9 – Notice that $G_{\rm s.c.}(z) \to 0$ as $|z| \to \infty$. By (ii), this implies that for some $R > 0$, and $d \geq 1$ large enough, $N_d(z) = 0 \Rightarrow |z| \leq R$. By (i), we even have that for all $z \in \mathbb{C}$, $N_d(z) = 0 \Rightarrow z \in [-R, R]$. We will show

(H) Let[10] $(a, b) \in (2, \infty)\backslash\{y_\star(\lambda)\}$ such that $a < b$. Let $\Gamma_d(a, b)$ be the number of zeroes of $N_d(z)$ located inside the real interval (a, b), counted with multiplicity. Then

$$\Gamma_d(a,b) \xrightarrow[d\to\infty]{} \Gamma(a,b) \coloneqq \mathbb{1}\{y_\star(\lambda) \in (a,b)\}.$$

Indeed, assume (H) holds, and $\lambda > 1$. Then for all $\varepsilon > 0$ small enough, and $d \geq 1$ large enough, there is exactly one zero[11] $\gamma_d \in (2+\varepsilon, R]$ of $\gamma \mapsto N_d(\gamma)$, and it is a simple root. By the point above, it is the only zero in $(2+\varepsilon, \infty)$. Further, since $\Gamma_d(y_\star - \eta, y_\star + \eta) \to 1$ for any $\eta > 0$, we get $\gamma_d \to y_\star(\lambda)$ as $d \to \infty$. Similarly, if $\lambda \leq 1$, then for any $\varepsilon > 0$ there is no zero of $\gamma \mapsto N_d(\gamma)$ in $(2+\varepsilon, \infty)$, so any $\gamma_d > b_d$ with $N_d(\gamma_d) = 0$ must satisfy $\gamma_d \to 2$ as $d \to \infty$.

It remains to prove (H). Let $\mathcal{C}$ be the circle in the complex plane with diameter $[a, b]$. Since $a, b \neq y_\star(\lambda)$, by (ii), $N_d(z)$ does not vanish on $\mathcal{C}$. Thus, by the argument principle and the remark above on the zeroes of N_d:

$$\Gamma_d(a,b) = \frac{1}{2i\pi}\oint_{\mathcal{C}} \frac{N_d'(z)}{N_d(z)}\mathrm{d}z.$$

[10] If $\lambda \leq 1$, set $y_\star(\lambda) = 2$ by convention.

[11] Notice that γ_d does not depend on the choice of ε.

By the Cauchy integral formula, if $N_d \to M_\lambda$ uniformly on K_η, then $N'_d \to M'_\lambda$ uniformly on K_η. Therefore, we get

$$\lim_{d\to\infty} \Gamma_d(a,b) = \frac{1}{2i\pi}\oint_{\mathcal{C}} \frac{M'_\lambda(z)}{M_\lambda(z)}\mathrm{d}z = \mathbb{1}\{y_\star(\lambda) \in (a,b)\},$$

which ends the proof. □

3.4 (Some) generalizations

The careful reader will have noticed that the proof of Theorem 3.6 is very generic, and one can generalize it in several ways. Let us mention a few of them.

- **Beyond rotational invariance** – We used critically that the eigenvectors of $\mathbf{W}$ are *completely delocalized*, since the distribution of $\mathbf{W}$ is rotationally invariant. This can be relaxed to approximate delocalization, allowing in particular matrices with i.i.d. non-Gaussian elements. Moreover, one can even completely drop randomness assumptions on the eigenvectors of $\mathbf{W}$, by assuming instead that the signal $\mathbf{x}$ is randomly sampled (independently of $\mathbf{W}$).

- **Beyond the semicircular law** – Our proof can be straightforwardly applied to any noise matrix $\mathbf{W}$ that satisfies the delocalization property just mentioned, and a convergence of its spectrum and extreme eigenvalues to some density ν, similar to Theorem 3.1. In this case the BBP threshold λ_c and the asymptotic values of y_1 and $(\mathbf{v}_1 \cdot \mathbf{x})^2$ depend on the Cauchy transform of ν: see [Mai24, Chapter 5] for more details.

- **Multiple spikes** – The argument can also be generalized to the case of "multi-spike" models, i.e. we consider instead

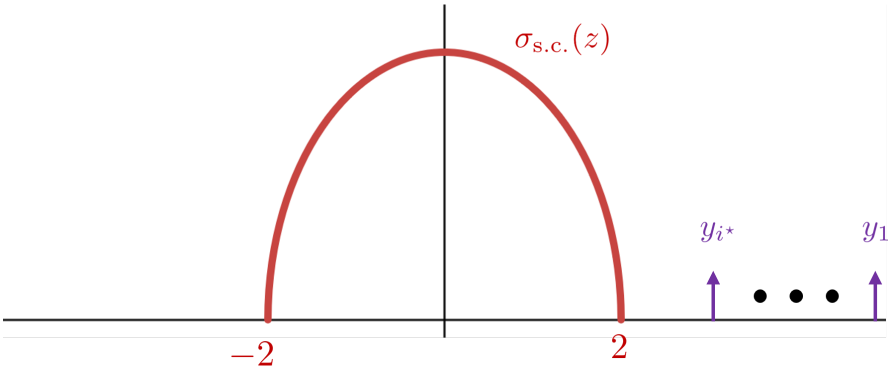
$$\mathbf{Y} = \mathbf{W} + \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{x}_i\mathbf{x}_i^\top, \tag{53}$$

  for some $r \geq 1$ (fixed as $d \to \infty$). We can assume without loss of generality that the $\mathbf{x}_i$'s are orthonormal vectors, and we assume that $\lambda_1 > \cdots > \lambda_r$. We get the following generalization of Theorem 3.6.

Theorem 3.10 (*"Multi-spike" BBP transition*)

Let $\mathbf{Y}$ be generated from eq. (53), denote its eigenvalues $y_1 \geq \cdots \geq y_d$, and corresponding eigenvectors $(\mathbf{v}_1, \cdots, \mathbf{v}_d)$. Let $i^\star \in \{0, \cdots, r\}$ such that $\lambda_{i^\star} > 1 \geq \lambda_{i^\star+1}$. Then:

- For all $i \in \{1, \cdots, i^\star\}$, $y_i \xrightarrow[d\to\infty]{\text{(a.s.)}} \lambda_i^{1/2} + \lambda_i^{-1/2}$, and $(\mathbf{v}_i \cdot \mathbf{x})^2 \xrightarrow[d\to\infty]{\text{(a.s.)}} 1 - \lambda_i^{-1}$.
- For all $i \in \{i^\star, \cdots r\}$, $y_i \xrightarrow[d\to\infty]{\text{(a.s.)}} 2$, and $(\mathbf{v}_i \cdot \mathbf{x})^2 \xrightarrow[d\to\infty]{\text{(a.s.)}} 0$.



Everything happens as if the different spikes in eq. (53) each had its own independent BBP transition! The case where there is degeneracy in the spiked matrix, i.e. if $\lambda_i = \lambda_j$ is slightly more subtle, and we refer to [BN11] for more on this setting.

4 Optimal estimation: approaches from statistical physics

5 Detection: contiguity, likelihood ratio, and the low-degree method

6 Optimization: Local minima in high-dimensional landscapes

References

[Abb+18] Emmanuel Abbe et al. “Group synchronization on grids”. In: *Mathematical Statistics and Learning* 1.3 (2018), pp. 227–256.

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. 118. Cambridge university press, 2010.

[Bar19] Jean Barbier. *Mean-field theory of high-dimensional Bayesian inference*. 2019. URL: https://jeanbarbier.github.io/jeanbarbier/docs/main.pdf.

[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.

[Ben+19] Gerard Ben Arous et al. “The landscape of the spiked tensor model”. In: *Communications on Pure and Applied Mathematics* 72.11 (2019), pp. 2282–2330.

[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.

[BSS23] A. S. Bandeira, A. Singer, and T. Strohmer. *Mathematics of Data Science*. Book draft available here. 2023.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. “Asymptotic mutual information for the binary stochastic block model”. In: *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2016, pp. 185–189.

[El 21] Ahmed El Alaoui. 2021. URL: https://courses.cit.cornell.edu/stsci6940/.

[FP07] Delphine Féral and Sandrine Péché. “The largest eigenvalue of rank one deformation of large Wigner matrices”. In: *Communications in mathematical physics* 272.1 (2007), pp. 185–228.

[Han14] Ramon van Handel. *Probability in High Dimension*. 2014. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[Kun25] Tim Kunisky. 2025. URL: http://www.kunisky.com/static/teaching/2025fall-rmt/rmt-notes-2025.pdf.

[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.

[KZ24] Florent Krzakala and Lenka Zdeborová. “Statistical physics methods in optimization and machine learning”. In: *Lecture Notes* (2024).

[Mai24] Antoine Maillard. 2024. URL: https://anmaillard.github.io/assets/pdf/lecture_notes/MDS_Fall_2024.pdf.

[Mio19] Léo Miolane. “Fundamental limits of inference: A statistical physics approach.” PhD thesis. Ecole normale supérieure-ENS PARIS; Inria Paris, 2019.

[MR14] Andrea Montanari and Emile Richard. “A statistical model for tensor PCA”. In: *Advances in neural information processing systems* 27 (2014).

[MS23] Laurent Massoulié and Ludovic Stéphan. "Inference in large random graphs". 2023.

[MS24] Andrea Montanari and Subhabrata Sen. "A friendly tutorial on mean-field spin glass techniques for non-physicists". In: *Foundations and Trends® in Machine Learning* 17.1 (2024), pp. 1–173.

[PB20] Marc Potters and Jean-Philippe Bouchaud. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.

[Sel24] Mark Sellke. 2024. URL: https://msellke.com/courses/STAT_291/course_page_website.html.

[Tal10] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*. Vol. 54. Springer Science & Business Media, 2010.

[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.

[ZK16] Lenka Zdeborová and Florent Krzakala. "Statistical physics of inference: Thresholds and algorithms". In: *Advances in Physics* 65.5 (2016), pp. 453–552.

A Some reminders in probability theory

A.1 General reminders in probability

We assume here that we have fixed a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all the following events and random variables are defined.

Lemma A.1 (*Borel-Cantelli*)

Let $(E_n)_{n\geq 1}$ be a sequence of events. Then

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \Rightarrow \mathbb{P}(\limsup_{n\to\infty} E_n) = 0.$$

Recall that $\limsup E_n \coloneqq \cap_{n\geq 1} \cup_{k\geq n} E_k$.

Let $(X_n)_{n\geq 1}$ be a sequence of real-valued random variables. Recall that $X_n \overset{\text{a.s.}}{\to}_{n\to\infty} X$ if $\mathbb{P}(\lim X_n = X) = 1$.

Proposition A.2 (*Reminder on almost sure convergence*)

Let $(X_n)_{n\geq 1}$ be a sequence of real-valued random variables. Then

(*i*) $X_n \overset{\text{a.s.}}{\to} X$ as $n \to \infty$ if and only if, for all $\varepsilon > 0$:

$$\mathbb{P}\left(\limsup_{n\to\infty}\{|X_n - X| \geq \varepsilon\}\right) = 0.$$

(*ii*) If for all $\varepsilon > 0$, $\sum_{n\geq 1} \mathbb{P}(|X_n - X| \geq \varepsilon) < \infty$, then $X_n \overset{\text{a.s.}}{\to} X$ as $n \to \infty$.

A.2 Gaussian random variables

The following elementary result is sometimes called Stein's lemma.

Lemma A.3 (*Gaussian integration by parts*)

Let $X \sim \mathcal{N}(0, \sigma^2)$ and $g : \mathbb{R} \to \mathbb{R}$ a differentiable function such that $\mathbb{E}[|Xg(X)|] < \infty$, $\mathbb{E}[|g'(X)|] < \infty$. Then

$$\mathbb{E}[Xg(X)] = \sigma^2 \mathbb{E}[g'(X)].$$

The following is a classical property of maxima of Gaussian random variables [Ver18]:

Proposition A.4 (*Maximum of independent Gaussians*)

Let $z_1, \cdots, z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Then

$$\lim_{n\to\infty} \frac{\mathbb{E}[\max_{i\in[n]} z_i]}{\sqrt{2\log n}} = 1 \text{ and } \operatorname*{p-lim}_{n\to\infty} \frac{\max_{i\in[n]} z_i}{\sqrt{2\log n}} = 1.$$

Finally, we cite a useful result allowing to compare the maxima Gaussian processes through their covariances. Recall that a random process $(X_t)_{t\in T}$ is called a *Gaussian process* if for every finite subset $T_0 \subseteq T$ the vector $(X_t)_{t\in T_0}$ has normal distribution.

Lemma A.5 (*Slepian/Sudakov-Fernique inequality*)

Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean-zero Gaussian processes. Assume that for all

$(s,t) \in T$ we have

$$\mathbb{E}[(X_s - X_t)^2] \le \mathbb{E}[(Y_s - Y_t)^2].$$

Then

$$\mathbb{E}[\max_{t\in T} X_t] \le \mathbb{E}[\max_{t \in T} Y_t].$$

A.3 Sub-Gaussian random variables

Definition A.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 $t > 0$, $\mathbb{P}[|X| \ge t] \le 2\exp\{-t^2/(2K_1^2)\}$, for some $K_1 > 0$.

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

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

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

Challenge A.1. *Check that the conditions (i), (ii), (iii) in Definition A.1 are equivalent, and that (K_1, K_2, K_3) differ by at most an absolute multiplicative constant.*

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

Challenge A.2. *Show that if $|X| \le a$, then X is $\mathrm{SG}(Ka)$, for $K > 0$ an absolute constant.*

We have a similar upper bound to Proposition A.4 when considering sub-Gaussian random variables.

Proposition A.6 (*Maximum of sub-Gaussian random variables*)

Let $n \ge 2$, and $X_1, \cdots, X_n$ be sub-Gaussian random variables, not necessarily independent. Then $Z \coloneqq \max_{i\in[n]} X_i$ is also sub-Gaussian, and we have ($\lesssim$ means "up to a global constant")

$$\begin{cases} \|Z\|_{\psi_2} & \lesssim \left(\max\limits_{i\in[n]} \|X_i\|_{\psi_2}\right) \cdot \sqrt{\log n}, \\ \mathbb{E}[Z] & \lesssim \left(\max\limits_{i\in[n]} \|X_i\|_{\psi_2}\right) \cdot \sqrt{\log n}. \end{cases}$$

A.4 Concentration inequalities

The following concentration inequality is very useful.

Theorem A.7 (*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 $t > 0$:

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

Beyond sums of independent random variables, one can show that Lipschitz functions of Gaussian random variables also enjoy fast concentration properties.

Theorem A.8 (*Gaussian concentration*)

Let $d \geq 1$ and $\mathbf{X} \sim \mathcal{N}(0, \mathrm{I}_d)$. Let $F : \mathbb{R}^d \to \mathbb{R}$ a L-Lipschitz function, i.e. such that $|F(\mathbf{x}) - F(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then, for any $t > 0$

$$\mathbb{P}(|F(\mathbf{X}) - \mathbb{E}F(\mathbf{X})| \geq t) \leq 2\exp\left\{-\frac{t^2}{2L^2}\right\}.$$

In particular, for any F as in Theorem A.8 and any $\gamma \in \mathbb{R}$ we have

$$\mathbb{E}e^{\gamma[F(\mathbf{X}) - \mathbb{E}F(\mathbf{X})]} \leq e^{\frac{c\gamma^2 L^2}{2}}, \tag{54}$$

for some $c > 0$ a universal constant.

A similar result holds for the uniform distribution on the unit sphere.

Theorem A.9 (*Lipschitz concentration on the sphere*)

There is $c > 0$ such that the following holds. Let $d \geq 1$ and $\mathbf{u} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$. Let $F : \mathbb{R}^d \to \mathbb{R}$ a L-Lipschitz function for the Euclidean distance. Then for any $t > 0$

$$\mathbb{P}(|F(\mathbf{u}) - \mathbb{E}F(\mathbf{u})| \geq t) \leq 2\exp\left\{-\frac{cdt^2}{L^2}\right\}.$$

B Solutions to problems

B.1 Section 3

Solution of Challenge 3.1 – Let $t > 2$. Changing variables to $x = 2\cos\theta$ we get:

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

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

$$\begin{aligned} G_{\text{s.c.}}(t) &= \frac{1}{\pi}\oint_{|\zeta|=1} \left(\frac{\zeta - \zeta^{-1}}{2i}\right)^2 \frac{1}{t - (\zeta + \zeta^{-1})}\frac{\mathrm{d}\zeta}{i\zeta}, \\ &= \frac{1}{4i\pi}\oint_{|\zeta|=1} \frac{(\zeta^2 - 1)^2}{\zeta^2(\zeta^2 - t\zeta + 1)}\mathrm{d}\zeta. \end{aligned} \tag{55}$$

The integrand in eq. (55) 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. (55), we finally find

$$G_{\text{s.c.}}(t) = \frac{t - \sqrt{t^2-4}}{2},$$

which ends the proof. □