Mean-field theory of high-dimensional Bayesian inference

Jean Barbier

International Center for Theoretical Physics, Trieste, Italy
jbarbier@ictp.it

These notes have been designed for a lecture at the school "Mathematical and Computational Aspects of Machine Learning" held at the Scuola Normale Superiore di Pisa from 7–11 october 2019. This course aims much more at pedagogy than completeness or presenting the very latest development in the field. It tries to be self-contained and accessible by a wide audience, so it assumes almost no prior knowledge. My goal is to recall basic (deep) concepts, as well as to provide *some* modern analytic and algorithmic tools used in high-dimensional inference. The language, techniques and concepts I will use are borrowed from various fields that share a lot and complement each other to form a young and dynamic research area at the intersection of statistical mechanics of disordered systems/spin glasses, information theory, high-dimensional probability and statistics, signal processing and machine learning. I will emphasize the connections between these fields and their specific terminology. I will also try to be rigorous whenever possible, and I will spend time on mathematical "details" that actually convey a lot of information. My motivation is that I've been trained as a physicist and discovered the beauty of mathematical proofs quite late, but I quickly realized that mastering all the ϵ's may yield a much deeper understanding of the physics hidden behind the apparent mess. Great references for those who want to go beyond on the connections between statistical physics and inference are [1–3].

A main question we will (very partially) try to answer is:

When does data contains enough information so that it can be used to infer something about the process that generated it?

This really is an *information-theoretic* question. Once the information-theoretic limits established we will then naturally ask ourselves:

Can we optimally extract/process the information from the data at low computational cost in order to perform efficient inference about the data-generating process?

This question is an *engineering/algorithmic* question. These two questions complement each other: without a clear answer to the information-theoretic question, people designing algorithms could lose time trying to improve algorithms with no hope of sucess, as they migh already be close to optimal. Also, the understanding of the barriers to information extraction provides guidance in algorithms design.

We will focus on a modern high-dimensional inference problem: the *spiked Wigner model*. This models the task of factorizing a large noisy data matrix in order to reduce its dimension, i.e., *principal component analysis* which is probably one of the most fundamental problem in machine learning. The choice of focusing on one problem rather than multiple ones is that it is simpler in terms of notations etc, yet its analysis requires all key ingredients necessary in the study of more complicated settings. Moreover applying numerous techniques to the same model helps in connecting them and see the global picture. Finally the spiked Wigner model is intimately linked to the most studied spin glass model: the Sherrington-Kirkpatrick model. Therefore it will be easier to make bridges with the rich literature in statistical physics that eventually lead to many of the ideas and techniques presented in this course.

A word about notations. Bold letters $(\boldsymbol{X}, \boldsymbol{y}, \dots)$ will be used for vectors, matrices etc, plain letters for scalars $(X_i, y_j, \dots)$. We will follow the information theory convention: random variables will be capital letters $(\boldsymbol{X}, Y_i, \dots)$. Their associated outcomes/realizations are in small letters $(\boldsymbol{x}, y_i, \dots)$.

Contents

1 Bayesian inference, information theory and statistical mechanics

1.1 Statistical inference, Bayes formula and decision theory

Statistical inference. For this part I highly recommend MacKay's book [4].

Before going into high dimensions lets go back to the basics: What is statistical inference? An important distinction is between *forward and inverse probabilities*. Consider a random generative process $\boldsymbol{x} \to \boldsymbol{y}$ where $\boldsymbol{x}$ are *parameters* of the model, also called *signal*, and $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ are *data* generated by the process $\to$. Forward probability involves computing the probability distribution (or various statistics) of functions of the data. The parameters are known and fixed, the data is not and is therefore modeled by a random variable. E.g., in "ball and boxes" exercises, something like computing the probability of drawing y_r red balls and y_b blue ones –$\boldsymbol{y}$ is the unknown data outcome–, the mean, or the variance of the random $\boldsymbol{Y}$ etc. In this case the known parameters could be the number of balls of different colors in the urn etc. Conceptually, in forward probability the random experiment has not yet taken place (or equivalently it took place but nothing is known about its outcome) and we try to *predict* what will happen, i.e., to predict the data it will generate which is therefore seen as a random variable:

$$\boldsymbol{x} \to \boldsymbol{Y}.$$

In inverse probability, instead, the random experiment already took place and generated some *observed data* $\boldsymbol{y}$, which is therefore not random but fixed. The inference task is then to compute the conditional probability distribution of the parameters of the process *given* the observed data, in order to reconstruct the parameters which are now considered as the random variables:

$$\boldsymbol{y} \leftarrow \boldsymbol{X}.$$

This diagram emphasizes that we know the data, and want to inverse the process to reconstruct the unknown parameters/signal.

The Bayes formula. The random data generative process is modeled by the *likelihood function* (or simply likelihood) $\mathcal{L}(\boldsymbol{x}|\boldsymbol{y}) \equiv P(\boldsymbol{y}|\boldsymbol{x})$. The likelihood is known, exactly or approximately (there might be mismatch between the true likelihood that helped generate the data and the assumed one). Here is a subtle point: the likelihood *should not* be considered as a probability distribution: it is a function of the hidden parameters to infer. This is because the data is *fixed*. You should never say "the likelihood of the data given the parameters", this is wrong. Instead you should say "the likelihood of the parameters given the data". In the

likelihood $\boldsymbol{y}$ actually plays the role of parameters, and $\boldsymbol{x}$ its argument, thus the notation $\mathcal{L}(\boldsymbol{x}|\boldsymbol{y})$.

The parameters $\boldsymbol{x}$ are unknown, fine. But maybe you still know *something* about them, such as their domain etc. All the a-priori information, i.e., the information we have/assume about $\boldsymbol{x}$ *before getting the data* is encompassed by the *prior* $P(\boldsymbol{x})$. The prior should never depend on the data, and never been changed once the data acquired. Combining our a-priori knowledge and the one gained from the data is done using the *Bayes formula*: the posterior distribution, which summarizes our *belief*[1] about the parameters value given the data, reads

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{y}) = P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})}{\int dP(\boldsymbol{x}')P(\boldsymbol{y}|\boldsymbol{x}')} = \frac{P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})}{P(\boldsymbol{y})}. \tag{1}$$

The Bayes formula can also be written in the form of the *product rule*:

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x}).$$

The marginal of the data $P(\boldsymbol{y})$ is called *evidence*. In some cases extra parameters of the model $\boldsymbol{\theta}$ are known (called hypotheses). They can be included in the model by conditionning:

$$P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{P(\boldsymbol{y}|\boldsymbol{\theta})}. \tag{2}$$

High-dimensional inference. In classical statistics we are generally interested in settings where the number of parameters to infer n is small with respect to the number of data points m: $n \ll m$. Often in this setting performing inference from the posterior or from the likelihood alone give same results: there are so many data points that the likelihood completely dominates the prior, so the posterior converges to the likelihood.

Inference in high dimension means that both n, m are large, and the total *signal-to-noise ratio* (snr) is finite when $m, n \to +\infty$, so that the inference is not trivial. The snr measures the signal strenght compared to the noise. Often it

[1] In Bayesian inference the notion of probability is *subjective*: it quantifies our personal belief (or equivalently our level of ignorance) about the outcome of the experiment. In the Bayesian interpretation the outcome of an experiment is not fundamentally random, and therefore if we had access to more data about it and/or had a sufficient a-priori knowledge, the outcome would become fully predictable. In Bayesianism it makes perfect sense to speak about the probability of a one-shot experiment, as it simply measures how much we think we know about what will happen. Also, *Bayesian inference cannot be made without assumptions*: it requires assuming a prior, as well as model $P(\boldsymbol{y}|\boldsymbol{x})$ for the data generative process.

In contrary in frequentism, probability is an *objective* notion, i.e., a fundamental feature of a system and the best way we have to characterize its physical behavior. In this view the probability of a one-shot event makes no sense, as it is defined as the event's frequency when the random experiment is repeated. The outcome of a single experiment is fundamentally random; this randomness is not due to our lack of knowledge. Frequentist statistics require no assumptions, only repeating experiments and looking at frequencies of events.

means that $m = \Theta(n) \gg 1$ but we will see that not always. This regime is related to "big-data" and machine learning applications, where the number of data points is huge, but the number of parameters in the model (like the weights of a neural network) is also large[2].

Bayesian decision theory. In a inference task one wants to reconstruct parameters from data. To precisely define what we mean by "reconstructing", we need to provide an error metric that quantifies the inference quality of our estimate of the signal. We will see that whatever reasonable metric we use, the optimal estimator for this particular metric is always derived from the posterior.

Denote $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y}(\boldsymbol{x}))$ the output of our algorithm, or *estimator*:

$$\hat{\boldsymbol{x}} = \text{algo}(\boldsymbol{y}(\boldsymbol{x})).$$

Let's say that your goal is to perfectly reconstruct the parameters, i.e., to minimize

$$E(\hat{\boldsymbol{x}}, \boldsymbol{x}) \equiv 1 - \mathbf{1}(\hat{\boldsymbol{x}} = \boldsymbol{x}). \tag{3}$$

$E(\hat{\boldsymbol{x}}, \boldsymbol{x})$ is called the *objective function* that we want to optimize. When the objective is to be minimized, we also call it the *loss*, *cost*, *regret* or *energy* (not surprisingly)[3]. But we do *not* have access to the loss, as it depends on the unknown parameters. So what is the best way to approximate the loss? By averaging it over the posterior $P(\boldsymbol{x}|\boldsymbol{y})$ of course, as the posterior properly combines all information we have about the signal. This leads to the definition of the *risk* $R(\hat{\boldsymbol{x}}, \boldsymbol{y})$ of the estimator $\hat{\boldsymbol{x}}$ associated to this loss[4]. In the case of the loss (3), the risk is the so-called *block-error-rate* (a terminology coming from communication):

$$\begin{aligned} R(\hat{\boldsymbol{x}}, \boldsymbol{y}) = p_B(\hat{\boldsymbol{x}}, \boldsymbol{y}) &\equiv \sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x}|\boldsymbol{y}) E(\hat{\boldsymbol{x}}, \boldsymbol{x}) \\ &= 1 - \sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x}|\boldsymbol{y}) \mathbf{1}(\hat{\boldsymbol{x}} = \boldsymbol{x}) = 1 - P(\hat{\boldsymbol{x}}|\boldsymbol{y}). \end{aligned}$$

$p_B(\hat{\boldsymbol{x}}, \boldsymbol{y})$ is the *a-posteriori probability of the estimator $\hat{\boldsymbol{x}}$ to be wrong, given $\boldsymbol{y}$*. This quantity, as opposed to the loss, can be in principle computed, as the $\boldsymbol{x}$ is now averaged over the posterior (in the expresison above $\boldsymbol{x}$ is a dummy variable, not the true value of the signal/parameters). Maybe computing $p_B(\hat{\boldsymbol{x}}, \boldsymbol{y})$ is computationally expensive, but that is another issue.

So now the question becomes: which estimator/algorithm, that assumes only knowledge of $\boldsymbol{y}$ (and maybe some prior knowledge about $\boldsymbol{x}$), minimizes this loss?

[2] In modern neural networks it is often the case that $n \gg m$, which should a-priori lead to *overfitting*, as there are more free parameters than data points, so in theory the noise can be fitted. The reason why, empirically, largely overparametrized neural networks do not overfit is a question at the forefront of the research in the field; see, e.g., a recent article about that [5].

[3] When it is to be maximized instead, it is called *reward*, *profit*, *utility* or *fitness* function, depending on the context.

[4] The *Bayes-risk* is the expectation of the risk over the evidence: $R(\hat{\boldsymbol{x}}) \equiv \int dP(\boldsymbol{y})\, R(\hat{\boldsymbol{x}}, \boldsymbol{y})$.

We therefore compute

$$\hat{\boldsymbol{x}}_{\rm opt}(\boldsymbol{y}) \equiv \operatorname*{argmin}_{\hat{\boldsymbol{x}}} p_B(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \operatorname*{argmin}_{\hat{\boldsymbol{x}}} \big\{1 - P(\hat{\boldsymbol{x}}|\boldsymbol{y})\big\} = \operatorname*{argmax}_{\hat{\boldsymbol{x}}} P(\hat{\boldsymbol{x}}|\boldsymbol{y}).$$

The optimal estimator to minimize the block-error-rate is therefore the *MAP estimator*, where MAP stands for maximum a-posteriori[5]. The associated (optimal) risk is

$$p_B(\boldsymbol{y}) \equiv p_B(\hat{\boldsymbol{x}}_{\rm opt}(\boldsymbol{y}), \boldsymbol{y}).$$

We may also be interested in the average performance of this estimator/algorithm over the data generative process, i.e., over the evidence:

$$p_B \equiv \sum_{\boldsymbol{y}} P(\boldsymbol{y})\, p_B(\boldsymbol{y}).$$

Another loss more appropriate when the parameters are continuous (in which case perfect reconstruction is generally doomed) and that appears often in the signal processing literature is the square error loss $E(\hat{\boldsymbol{x}}, \boldsymbol{x}) \equiv \frac{1}{n}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$. The associated risk is the mean-square error (MSE):

$$\mathrm{MSE}(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \int dP(\boldsymbol{x}|\boldsymbol{y})\, \frac{1}{n}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2.$$

Let us minimize this risk (which is convex in the estimator):

$$\nabla_{\hat{\boldsymbol{x}}} \mathrm{MSE}(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \boldsymbol{0} \quad \Rightarrow \quad \int dP(\boldsymbol{x}|\boldsymbol{y})(\boldsymbol{x} - \hat{\boldsymbol{x}}) = \boldsymbol{0}.$$

This yields the minimum mean-square error (MMSE) estimator, which is nothing else than the posterior mean:

$$\hat{\boldsymbol{x}}_{\rm opt}(\boldsymbol{y}) \equiv \operatorname*{argmin}_{\hat{\boldsymbol{x}}} \mathrm{MSE}(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \int dP(\boldsymbol{x}|\boldsymbol{y})\, \boldsymbol{x} \equiv \langle \boldsymbol{X} \rangle.$$

We introduced a notation coming from statistical mechanics for posterior expectations: the *Gibbs-bracket* $\langle \boldsymbol{X} \rangle$ (which depends on the data). The smallest possible expected MSE, the (average) MMSE, is then

$$\begin{aligned} \mathrm{MMSE} = \mathrm{MMSE}(\boldsymbol{X}^*|\boldsymbol{Y}) &\equiv \int dP(\boldsymbol{y})\, \mathrm{MSE}(\hat{\boldsymbol{x}}_{\rm opt}(\boldsymbol{y}), \boldsymbol{y}) \\ &= \int dP(\boldsymbol{y}) dP(\boldsymbol{x}|\boldsymbol{y})\, \frac{1}{n}\|\boldsymbol{x} - \langle \boldsymbol{X} \rangle\|_2^2 \\ &= \int dP(\boldsymbol{x}^*) dP(\boldsymbol{y}|\boldsymbol{x}^*)\, \frac{1}{n}\|\boldsymbol{x}^* - \langle \boldsymbol{X} \rangle\|_2^2. \end{aligned}$$

[5]The optimal estimator is equivalently obtained by minimizing the risk $R(\hat{\boldsymbol{x}}, \boldsymbol{y})$ or the Bayes-risk $R(\hat{\boldsymbol{x}})$.

More compactly,

$$\mathrm{MMSE}(\boldsymbol{X}^*|\boldsymbol{Y}) = \frac{1}{n}\mathbb{E}\|\boldsymbol{X}^* - \langle \boldsymbol{X}\rangle\|_2^2$$
$$= \frac{1}{n}\mathbb{E}_{\boldsymbol{Y}}\mathrm{Var}(\boldsymbol{X}|\boldsymbol{Y}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{Y}}\big\langle \|\boldsymbol{X} - \langle \boldsymbol{X}\rangle\|_2^2\big\rangle, \quad (4)$$

where the first $\mathbb{E} = \mathbb{E}_{\boldsymbol{X}^*,\boldsymbol{Y}} = \mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}$. We have introduced the $*$ notation to distinguish between a sample from the posterior $\boldsymbol{X} \sim P(\cdot\,|\boldsymbol{y})$, and the (random) ground-truth signal to infer $\boldsymbol{X}^* \sim P$. If we want to emphasize that the MMSE is the expected posterior variance we write $\frac{1}{n}\mathbb{E}_{\boldsymbol{Y}}\mathrm{Var}(\boldsymbol{X}|\boldsymbol{Y})$. If instead we prefer to think of it as the mean-square deviation between the optimal estimator and the ground-truth we prefer the notation $\mathrm{MMSE}(\boldsymbol{X}^*|\boldsymbol{Y})$. But this is the same. We will keep this notation in the remaining in order to avoid confusions.

As we will see soon the MMSE, in addition of being relevant in most applications, has the great advantage of being more easily accessible/computed thanks to a simple relation that links it directly to the main quantity of interest: the *mutual information*, or *free energy* (these are linked by an additive constant). Also it is often the case that when one error metric worsen at a *phase transition* point (e.g., as the noise level increases, or the amount of data decreases), the others become bad too. This is because phase transitions are intrinsic of the problem at hand as they depend only on the mutual information. The optimal errors are just *observables* and the dramatic change of the behavior of observables happens generally at the same point. E.g., at 0 degree celsius not only the correlation lenght between molecules of water abruptly changes, but also their mean displacement etc.

A deep consequence of the Bayes formula: the "Nishimori identity". Let (X, Y) be a couple of random variables (that can be vectors etc) with joint distribution P_{XY} and conditional distribution $P_{X|Y}$. Let $k \geq 1$ and let $X^{(1)}, \dots, X^{(k)}$ be i.i.d. random variables with distribution $P_{X|Y}$. Let us denote $\mathbb{E}$ the expectation w.r.t. P_{XY} and $\langle - \rangle$ the expectation w.r.t. the product measure $P_{X|Y}^{\otimes\infty}$. Then, for all continuous bounded function g we have[6]

$$\mathbb{E}\big\langle g(Y, X, X^{(2)}, \dots, X^{(k)})\big\rangle = \mathbb{E}\big\langle g(Y, X^{(1)}, X^{(2)}, \dots, X^{(k)})\big\rangle. \quad (5)$$

Proof. This directly follows from Bayes formula $P_{XY} = P_{X|Y}P_Y = P_{Y|X}P_X$. It is equivalent to sample the couple (X, Y) according to its joint distribution or to sample first Y according to its marginal distribution and then to sample X conditionally on Y from the conditional distribution. Thus the two $(k+1)$-tuples $(Y, X, X^{(2)}, \dots, X^{(k)})$ and $(Y, X^{(1)}, X^{(2)}, \dots, X^{(k)})$ have the same law.

[6] This identity has been abusively called "Nishimori identity" in the statistical physics literature despite that it is a simple consequence of Bayes formula. The "true" Nishimori identity concerns models with one extra feature, namely a gauge symmetry which allows to eliminate the input signal, and the expectation over the signal $\boldsymbol{X}$ in expressions of the form $\mathbb{E}\langle - \rangle$ can therefore be dropped.

In equations,

$$\begin{aligned}\mathbb{E}\big\langle g(Y, X, X^{(2)}, \dots, X^{(k)})\big\rangle \\ &\equiv \mathbb{E}_{XY}\mathbb{E}_{X^{(2)}|Y}\dots\mathbb{E}_{X^{(k)}|Y}\, g(Y, X, X^{(2)}, \dots, X^{(k)}) \\ &= \mathbb{E}_{Y}\mathbb{E}_{X|Y}\mathbb{E}_{X^{(2)}|Y}\dots\mathbb{E}_{X^{(k)}|Y}\, g(Y, X, X^{(2)}, \dots, X^{(k)}) \\ &= \mathbb{E}_{Y}\mathbb{E}_{X^{(1)}|Y}\mathbb{E}_{X^{(2)}|Y}\dots\mathbb{E}_{X^{(k)}|Y}\, g(Y, X^{(1)}, X^{(2)}, \dots, X^{(k)}) \\ &\equiv \mathbb{E}\big\langle g(Y, X^{(1)}, X^{(2)}, \dots, X^{(k)})\big\rangle.\end{aligned}$$

□

This seemingly innocent identity is actually absolutely key. It is the origin of a tremendous number of simplifications that allow to carry a complete analysis of high-dimensional inference problems in the *Bayesian optimal setting*. In an inference setting this means that the posterior $P_{X|Y}$ is known. In the proof above this is used at the last step to replace X by $X^{(1)} \sim P_{X|Y}$. It $X^{(1)}$ was drawn from another distribution than the posterior $P_{X|Y}$ the proof could not be carried out: this would create an assymetry between the $X^{(i)}$'s and X. Instead in the Bayesian optimal setting this identity says that, inside an average over everything $\mathbb{E}\langle - \rangle$, the signal X can be replaced by a sample from the posterior $X^{(1)}$: in expectation, the signal and a posterior sample play totally symmetric roles. As we will see, this deep symmetry is at the root of important concentration inequalities (of the MMSE and *overlap*, see below).

Identity (4) showing $\mathbb{E}\big\langle \|\boldsymbol{X} - \langle \boldsymbol{X}\rangle\|_2^2\big\rangle = \mathbb{E}\|\boldsymbol{X}^* - \langle \boldsymbol{X}\rangle\|_2^2$ was actually our first application of the Nishimori identity; indeed $\mathbb{E}\big\langle \|\boldsymbol{X} - \langle \boldsymbol{X}\rangle\|_2^2\big\rangle$ is just a function of $\boldsymbol{X} \sim P(\cdot\,|\boldsymbol{y})$ and of $\langle \boldsymbol{X}\rangle$ which only depends on $\boldsymbol{y}$, so $\boldsymbol{X}$ can be replaced by the signal $\boldsymbol{X}^*$. Another simplification thanks to it is

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|\boldsymbol{X}^* - \langle \boldsymbol{X}\rangle\|_2^2 \overset{\text{N}}{=} \frac{1}{n}\mathbb{E}\big\langle \|\boldsymbol{X} - \langle \boldsymbol{X}\rangle\|_2^2\big\rangle &= \frac{1}{n}\big(\mathbb{E}\big\langle \|\boldsymbol{X}\|_2^2\big\rangle - \mathbb{E}\|\langle \boldsymbol{X}\rangle\|_2^2\big) \\ &\overset{\text{N}}{=} \frac{1}{n}\big(\mathbb{E}\|\boldsymbol{X}^*\|_2^2 - \mathbb{E}\big\langle \boldsymbol{X}^{(1)}\cdot\boldsymbol{X}^{(2)}\big\rangle\big) \\ &\overset{\text{N}}{=} \rho - \mathbb{E}\langle Q\rangle, \qquad (6)\end{aligned}$$

where the *overlap* between a posterior sample and the signal is

$$Q(\boldsymbol{X}, \boldsymbol{X}^*) = Q \equiv \frac{1}{n}\boldsymbol{X}^*\cdot\boldsymbol{X}.$$

Above $\boldsymbol{X}^{(1)} = \boldsymbol{X}$ and $\boldsymbol{X}^{(2)}$ are i.i.d. samples from the posterior, called *replicas*, and a Gibbs-bracket sandwiching multiple replicas is the expectation w.r.t. the product measure $P(\cdot\,|\boldsymbol{y})^{\otimes\infty}$. In the derivation of (6) we used

$$\mathbb{E}\|\langle \boldsymbol{X}\rangle\|_2^2 = \mathbb{E}\big\langle \boldsymbol{X}^{(1)}\cdot\boldsymbol{X}^{(2)}\big\rangle \overset{\text{N}}{=} \mathbb{E}\big[\boldsymbol{X}^*\cdot\langle \boldsymbol{X}\rangle\big] \equiv n\mathbb{E}\langle Q\rangle.$$

1.2 Surprise, Shannon entropy and mutual information

We want to precisely quantify when statistical inference is possible or not by estimating if the data contains enough information about the model parameters. But what is information, and how to quantify it? The answer has been understood by Claude Shannon in 1948 in his seminal paper [6] who started information theory: the entropy. Computing entropies and relatives will be one of our main goal, so let us start by understand why it is the proper definition of information content conveyed by a random variable.

1.2.1 Shannon entropy and its properties

Surprise and Shannon entropy. For simplicity let us consider a discrete setting. Let $X \sim P$ a discrete r.v. with possible outcomes in $\mathcal{X} = (x_1, x_2, \dots, x_{|\mathcal{X}|})$; the notation $X \sim P$ means that the probability distribution of X is P. X can be a vector, on any other type or random object. The *Shannon entropy* $H(X) = H(P)$ of the random variable X, or equivalently of the *ensemble* $\mathcal{E}_X = (X, \mathcal{X}, P)$, is defined as[7]:

$$H(X) = H(P) \equiv \sum_{x \in \mathcal{X}} P(x) \ln \frac{1}{P(x)} = \sum_{i=1}^{|\mathcal{X}|} P_i \ln \frac{1}{P_i}. \tag{7}$$

It is the expectation with respect to P of the *information content* $h(x) = h(P(x))$, or *surprise*, of the outcome x of the r.v. X:

$$h(x) = h(P(x)) \equiv \ln \frac{1}{P(x)}.$$

If the outcome x has low probability then observing it is quite surprising, and it brings a lot of information as it was not expected: $h(x)$ is high. If instead $P(x)$ is close to 1 it is not surprising to observe x, so this outcome brings low information: $h(x)$ is low. Said differently: if the outcome of a random variable is very probable, it is no surprise (and generally uninteresting) when it happens, because it was expected. However, if an outcome is unlikely to occur, it is much more informative if it happens to be observed. The term information content must be understood as a *potential* information gain if x is observed. Here the information content and entropy are expressed in "nats" (for "natural units"), because the logarithm is in natural basis. When using the $\log_2$ they are expressed in "bits".

Imagine you are in the desert and suddenly it rains like hell. Worst, it rains cows that play piano! What? It is amazingly surprising no? The probability of this event is actually so low that it brings an enormous amount of information; in this case it should lead you to the conclusion that you are dreaming. If instead your are in the desert and its super sunny and hot, it is not surprising at all; this does

[7] Do not get confused between the H of the Shannon entropy and the calligraphic $\mathcal{H}$ used for Hamiltonians later on.

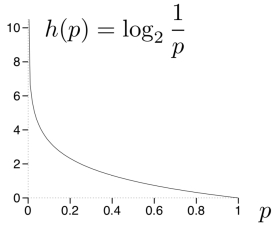
not bring more information than what you already know, and if you are dreaming, it is unlikely that this observation will help you realize it. Another example: the knowledge that some particular number will not be the winning one of a lottery provides very little information, because any particular chosen number will almost certainly not win. However, knowledge that a particular number will win a lottery has high informational value because it communicates the outcome of a very low probability event.

Be focused here. The entropy can also be interpreted as a *measure of unpredictability* of X, or of *uninformation/lack of knowledge* about what X's outcome will be: the more surprising are the outcomes in expectation, the more unpredictable is the actual outcome, which also mean the less we know about x *before* observing it. *$H(X)$ quantifies the amount of missing information necessary to determine the outcome of X before observing it.* This can be confusing because previously we said that $H(X)$ is an expected information content, while now we speak about a measure of uninformation. There is no paradox: an information $H(X)$ is *gained* in expectation when *x is actually observed.* But *prior* to observing the outcome, $H(X)$ is a measure of uninformation about it. Put differently: observing the outcome x *converts* in average $H(X)$ units of uninformation into information. So it just a matter of conceptually placing ourselves *before* x is observed –in which case the interpretation as a measure of uninformation may be more natural–, or *after* x is observed –where the interpretation as an expected information content seems to fit better. But at the end this is the same thing.

An example might help: the outcome of a toss of a fair coin $X_{\text{fair}} \sim \text{Ber}(1/2)$ is much more unpredictable than the outcome of a strongly biased coin $X_{\text{bias}} \sim \text{Ber}(9/10)$, or equivalently our lack of knowledge about what will be x_{fair} is higher: we are more uninformed. But when *observing* the outcome of the fair coin, we then *gain* more information than with the unfair one, because it is in average more surprising. In the first case, which has entropy $H(X_{\text{fair}})$ of one bit, betting on one side or the other is the same statistically. While in the second case, where $H(X_{\text{bias}}) = \frac{9}{10}\log_2\frac{10}{9} + \frac{1}{10}\log_2 10 \approx 0.47$, the outcome is much more predictable, we are less uninformed (= more informed); it would be an error not to bet on the outcome $x_{\text{bias}} = 1$.

The Shannon entropy $H(X)$ of an ensemble $\mathcal{E}_X$,
or equivalently of the random variable X, quantifies:
i) Its average information content,
i.e., the expected information gain when observing x.
ii) The average uninformation/lack of knowledge
about the outcome x prior to observe it.
iii) Its unpredictability.
The higher the entropy of X, the less "structured" its distribution is.

Finally, if expressed in bits, $H(X)$ is the expected number of binary "yes/no" ques-



p	$h(p)$	$H_2(p)$
0.001	10.0	0.011
0.01	6.6	0.081
0.1	3.3	0.47
0.2	2.3	0.72
0.5	1.0	1.0

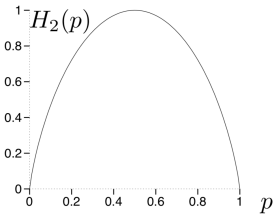


Figure 1: From [4]. Information content (in bits as the logarithm is in base 2) of an event with probability p. The less probable an outcome is, the greater its information content. On the right is the binary entropy function $H_2(p)$, with its maximum at $1/2$.

tions required to determine the outcome *before* it is observed, or equivalently, the expected number of binary questions that the oucome x answered *after* being observed.

"Good" properties of the entropy. We will mathematically justify, based on Shannon's source-coding theorem, that $H(X)$ is indeed *the* proper definition of information content of X. But at the moment let us admit it, and give some additional properties that stenghten this claim. Let $X \sim P$.

- $H(X) \geq 0$ with equality if and only if $P_i = 1$ for one i. There is no such thing as negative information, and a deterministic variable convey no information.
- $H(X)$ is maximized if P is uniform: $H(X) \leq \ln |\mathcal{X}|$ with equality if and only if $P_i = \frac{1}{|\mathcal{X}|}$ for all i.

So the uniform distribution (the less structured of them all) has maximum entropy, while the trivial distribution giving full weight to a single event has 0 entropy. As the entropy of the uniform distribution increases with $|\mathcal{X}|$, casting a die has higher entropy than tossing a coin because each outcome of a die toss has smaller probability ($1/6$ to be compared to the $1/2$ of the fair coin; each outcome of a die is more surprising as there are more of them).

Notice another nice property of the information content function $h(p) = -\ln p$. Imagine learning the outcome x and y of two independent random variables, X and Y. Intuitively, we might want any measure of the "amount of information gained" to have the property of additivity: for independent random variables X and Y, the information gained when we learn the outcome of both should equal the sum of the information gained if x alone were learned and the information gained if y alone were learned. And indeed:

- $h(x, y) = h(x) + h(y)$.
- $P_{XY} = P_X \otimes P_Y \;\Rightarrow\; H(X, Y) = H(X) + H(Y)$.

Here

$$H(X,Y) = H(P_{XY}) \equiv \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x,y)\ln\frac{1}{P(x,y)}$$

is the entropy of the joint distribution, $h(x,y) = h(P(x,y)) = -\ln P(x,y)$. In words, *entropy is additive for independent variables.* Define also the *conditional entropy* of X given Y:

$$H(X|Y) \equiv \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(y)P(x|y)\ln\frac{1}{P(x|y)}.$$

It is equal to the expected information revealed by evaluating the outcome of X *given* that you know already the outcome of Y. Or equivalently, it is the remaining amount of unpredictability of X given that Y has already been observed.

Other important properties of the entropy that confirm its interpretation are:

- The entropy $H(X,Y)$, i.e., the amount of information revealed by simultaneously evaluating (X,Y) equals the information revealed by conducting two consecutive experiments: first evaluating the value of Y, then revealing the value of X *given* that you know the value of Y (or the other way around). This is called the *chain rule for entropy*:

  $$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X).$$

- If f is a function then $H(f(X)|X) = 0$. Applying that to the previous formula yields $H(X) + H(f(X)|X) = H(f(X)) + H(X|f(X))$, so:

  $$H(f(X)) \leq H(X),$$

  with equality if f is invertible. The entropy may only decrease when X is passed through a function.

- If X and Y are independent then knowing the value of Y doesn't influence our knowledge of the value of X (since the two don't influence each other by independence):

  $$P_{XY} = P_X \otimes P_Y \;\Rightarrow H(X|Y) = H(X),\; H(Y|X) = H(Y).$$

- The entropy of two simultaneous events is no more than the sum of the entropies of each individual event, and are equal if the two events are independent:

  $$H(X,Y) \leq H(X) + H(Y).$$

- The entropy seen as a function of the probability distribution P is concave: for all $P^{(1)}$ and $P^{(2)}$ and $\lambda \in [0,1]$,

  $$H(\lambda P^{(1)} + (1-\lambda)P^{(2)}) \geq \lambda H(P^{(1)}) + (1-\lambda)H(P^{(2)}).$$

  Note that $\lambda P^{(1)} + (1-\lambda)P^{(2)}$ is a probability distribution.

1.2.2 Entropy from lossy compression: Shannon's source-coding theorem

We argued that the entropy is a proper measure of expected information content of a random variable, because it verifies many properties that are "natural" for an information measure. Is there a more principled way to show that this is indeed the proper quantity? Yes, through the notion of *lossy compression*. In this section we assume that we do not know yet that the proper definition of information content is the Shannon entropy $H(X)$: this is what we want to prove.

Lossy compression and essential bit content. Consider the ensemble $\mathcal{E}_X = (X, \mathcal{X}, P)$. One naive way to quantify its information content is through its *raw bit content* (here we use directly the $\log_2$ basis to have units in bits)

$$H_0(X) \equiv \log_2 |\mathcal{X}|.$$

This is the number of binary variables/bits necessary to code (i.e., map one-to-one) all outcomes in $\mathcal{X}$ to binary strings, that we call *codewords*; indeed $|\mathcal{X}| = 2^{H_0(X)}$. $H_0(X)$ is a lower bound to the number of binary questions that are always guaranteed to identify an outcome from the ensemble X[8]. $H_0(X)$ is an additive quantity for independent variables, as should be a proper information content measure. This measure of information content does not include any probabilistic element, and the encoding rule it corresponds to does not "compress" the source data, it simply maps each outcome $x \in \mathcal{X}$ to a constant-length binary string/codeword. To better quantify information we need somehow to take into account the *probabilities* of the different outcomes. One simple way to do so is to simply remove from the alphabet $\mathcal{X}$ the less probable outcomes. E.g., one could remove from in the standard ASCII alphabet the symbols $\{!, @, \#, \%, *, >, <, \backslash, /, \{, \}, [,]\}$; still you could understand a vast majority of the texts. But when removing symbols from the original alphabet $\mathcal{X}$, we increase the risk that some outcomes x cannot be described anymore in the compressed alphabet (or equivalently that multiple outcomes are associated to the same binary string/codeword, so that there are possible confusions because the mapping is not bijective). The term "lossy" therefore means that we allow *some* loss of information. We will see that how much loss should be allowed is a tricky question.

We introduce the *risk* δ we are taking when using this compression method: δ is the probability that there will be no name for an outcome $x \in \mathcal{X}$. E.g., $\mathcal{X} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}, \mathrm{f}, \mathrm{g}, \mathrm{h}\}$, and $P_X = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$. The raw bit content of this ensemble is 3 bits, corresponding to $2^3 = 8$ binary names/strings. But notice that $P(x \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}) = \frac{15}{16}$. So if we are willing to run a risk of $\delta = 1/16$ of not having a name for x, then we can get by with four names $\{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$; half

[8] The raw bit content is related to the microcanonical entropy $S(E) = \ln \mathcal{Z}(E)$ in the microcanonical ensemble of statistical mechanics, where all configurations have same energy E and $\mathcal{Z}(E)$ simply counts the number of configurations with energy E. Expressed in base 2 the microcanonical entropy of an isolated system is therefore the number of bits necessary to code all its valid configurations in binary form.

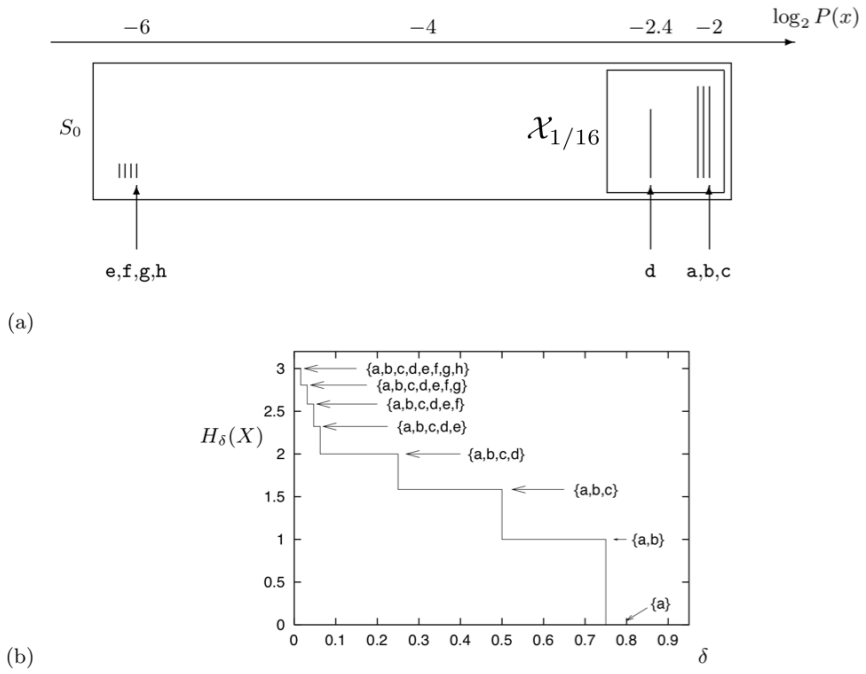


Figure 2: From [4]. (a) The outcomes of X ranked by probability. (b) The essential bit content $H_\delta(X)$. The labels show the smallest sufficient set as a function of δ. Note $H_0(X) = 3$ bits and $H_{1/16}(X) = 2$ bits.

as many names as are needed if every $x \in \mathcal{X}$ is required to have a name. Let us now formalize this idea.

To make a compression strategy with risk δ, we make the smallest possible subset $\mathcal{X}_\delta$ such that the probability that x is not in $\mathcal{X}_\delta$ is less than or equal to δ, i.e., $P(x \notin \mathcal{X}_\delta) \leq \delta$. So the *smallest δ-sufficient subset $\mathcal{X}_\delta$ is the smallest $\mathcal{X}_\delta \subseteq \mathcal{X}$ satisfying*

$$P(x \in \mathcal{X}_\delta) \geq 1 - \delta.$$

The subset $\mathcal{X}_\delta$ can be constructed by ranking the elements of $\mathcal{X}$ in order of decreasing probability and adding successive elements starting from the most probable elements until the total probability is $\geq 1 - \delta$. We can then create a data compression code by assigning a binary name to each element of the smallest sufficient subset. The process we described of defining a new compressed alphabet in order to describe a random variable is called *source-coding*, and the new alphabet, here $\mathcal{X}_\delta$ or equivalently its binary representation, is called *code*[9]. Note

[9]The terminology is similar to the one used in communication, i.e., in the *noisy channel-coding problem*. Note that in channel-coding we define a code with *additional redundancy* to increase robustness to the channel noise, while in source-coding we at contrary try to get rid as much as possible of the redundancy in order to extract the "pure" (non-redundant) information.

that representing $\mathcal{X}_\delta$ in binary form is not necessary, it is just sometimes more convenient to think in terms of binary strings. Going from one representation to another in a bijective way does not change the information content.

For each value of δ we can then define a new measure of information content as the raw bit content of $\mathcal{X}_\delta$: the *δ-essential bit content* of X is

$$H_\delta(X) \equiv \log_2 |\mathcal{X}_\delta|. \tag{8}$$

Note that $H_0(X)$ is the special case of $H_\delta(X)$ with $\delta = 0$ (if $P(x) > 0$ for all $x \in \mathcal{X}$). This quantity seems like a sound definition of information content of X: we have compressed X by reducing its alphabet, allowing a δ-probability of error. If δ is small, the δ-sufficient subset is enough to describe properly X, so that all elements in $\mathcal{X}_\delta$ must contain "pure" information, quantified by the δ-essential bit content $H_\delta(X)$. Unfortunately this definition suffers an important caveat: in general $H_\delta(X)$ strongly depends on δ, so how to properly set δ? Which value really selects the "pure" information? There is not definite answer, see Fig. 2. This is not desirable for a "fundamental" measure of information.

Is the notion of compression therefore not appropriate to quantify information content? Actually it is, but we have change a bit the setting; this is where the genius of Shannon enters into the game.

Source-coding. Instead of just considering the random variable X from the ensemble $\mathcal{E}_X = (X, \mathcal{X}, P)$, consider now a *source* that generates a string of *independent* outcomes from the *same* X: $\boldsymbol{x}^n = (x_1, x_2, \ldots, x_n)$. This string is therefore the outcome of a random variable $\boldsymbol{X}^n = (X_1, X_2, \ldots, X_n) \sim P^{\otimes n}$ taking values in $\mathcal{X}^n$, where X_i are independent copies of X. E.g., if X is a biased coin $X \sim \mathrm{Ber}(p)$, then strings will be made of zeros and ones with, *typically*, a number np of ones, see Fig. 3. The motivation in studying this source is that $i)$ its expected information content, whatever it means, must be n times the one of X alone, because the information content must be additive for independent variables X_i's; $ii)$ as n will get larger, simplifications will occur thanks to the *law of large numbers*[10].

The question then becomes: how much can we compress the random process/string $\boldsymbol{X}^n$, so that (almost no) information is lost? Said differently, what is the minimal number of binary symbols $H_\delta(\boldsymbol{X}^n) = \log_2 |(\mathcal{X}^n)_\delta|$[11] necessary to represent this random string when the risk δ is small? Indeed, if the random string is *maximally compressed* up to only a small error probability, then the number of symbols necessary in the compressed representation sounds like a very reasonable measure of information content. Because if less symbols than this were used to represent the string, there would be a high risk of making errors and therefore

[10] The law of large numbers is responsible for the validity of statistical experiments. Without this law, we could never verify statistical properties of a system by performing many experiments. In particular, quantum mechanics would be free of any physical meaning.

[11] The parentheses are here to emphasize that $(\mathcal{X}^n)_\delta$ is the δ-sufficient subset associated with $\mathcal{X}^n$; this is not $\mathcal{X}_\delta \times \ldots \times \mathcal{X}_\delta = (\mathcal{X}_\delta)^n$.

$\mathbf{x}$	$\log_2(P(\mathbf{x}))$
...1.....1.....1....1.1.....1.....1.....1.....1.....11...	-50.1
.....1.....1.....1.....1...1.....1.....1.....1....	-37.3
.....1....1..1..1....11..1.1.....11.....1...1.1..1...1.....1.	-65.9
1.1...1.....1.....11.1..1.....1..1.....1.....1..1.11.....	-56.4
...11.....1...1.....1.1.....1.....1....1...1.....1.....1.....	-53.2
.....1.....1.....1.1.....1.....1.....1...1.....1.....	-43.7
.....1.....1.....1...1.....1.....1.....1.....1..11.....	-46.8
.....1..1..1.....111.....1.....1.....1.....1.1...1...1.....1	-56.4
.....1.....1.....1.....1.....1....1.....1...	-37.3
.....1.....1.....1.....1.....1..1.1.1.1..1.....1.	-43.7
1.....1.....1...1.....1.....1....1....1.....1..11..1.1...1.....	-56.4
.....11.1.....1.....1.....1.....1.....1.....	-37.3
.1.....1...1.1.....1.....11.....1.1...1.....1.....11.....	-56.4
.....1...1..1.....1..11.1.1.1...1.....1.....1.....1..1.....	-59.5
.....11.1.....1....1..1.....1.....1.....1.....1.....	-46.8
.....	-15.2
11111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111	-332.1

Figure 3: From [4]. The top 15 strings are samples from $\boldsymbol{X}^n$, where $X \sim \mathrm{Ber}(p)$. They all look typical. The bottom two are the most and least probable strings in this ensemble. The final column shows the log-probabilities of the strings, to be compared with the entropy $H(\boldsymbol{X}^{100}) = 46.9$ bits.

information would be lost. So each symbol in the compressed form must carry "pure" information. This reasonning applied to X alone is precisely what lead us to the definition of δ-essential bit content $H_\delta(X)$. We've seen that this quantity suffered from its strong dependence in δ. The magic is that as $n \to +\infty$, as opposed to the case $n = 1$ discussed previously, the maximal compression level of $\boldsymbol{X}^n$ up to some risk δ –and therefore its δ-essential bit content $H_\delta(\boldsymbol{X}^n)$– will (almost) not depend on δ; this suddenly makes this object much more attractive as a definition of information content.

Let us now state Shannon's source coding theorem, that justifies why Shannon entropy is the proper definition of information content. Remember that at this stage the entropy $H(X) = -\sum_{x\in\mathcal{X}} P(x)\ln P(x)$ is just a mathematical definition, we did not yet justify its meaning of information content; our best candidate at the moment is the δ-essential bit content.

Shannon's source-coding theorem. Let $\mathcal{E}_X = (X, \mathcal{X}, P)$ be an ensemble with entropy $H(X)$ bits (or nats depending of the basis of the log). Let $\mathcal{X}^n \ni \boldsymbol{X}^n = (X_1, X_2, \ldots, X_n) \sim P^{\otimes n}$. Given arbitrarily small $\epsilon > 0$ and any $0 < \delta < 1$, there exists a positive integer $n^*(\epsilon, \delta)$ such that for $n > n^*(\epsilon, \delta)$ the δ-essential

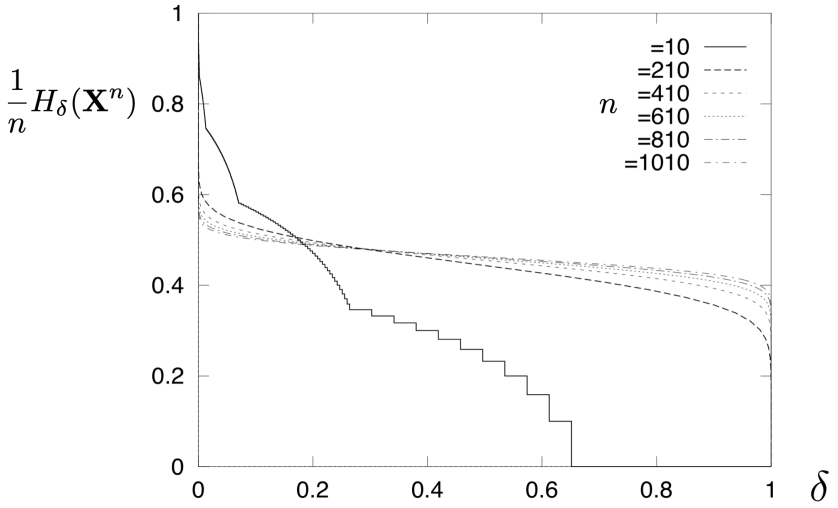


Figure 4: From [4]. Convergence of the δ-essential bit content per bit $\frac{1}{n}H(\boldsymbol{X}^n)$ for increasing n values, and where $X \sim \mathrm{Ber}(p) = 0.1$: we clearly observe that $\frac{1}{n}H(\boldsymbol{X}^n)$ becomes an increasingly flat function almost independent of δ (except very close to the border), which constant value tends to $H(X) = -0.1\log_2 0.1 - 0.9\log_2 0.9 \approx 0.47$ bits, except close to the borders $\delta = 0$ or 1.

bit content of the length-n string $\boldsymbol{X}^n$ verifies

$$\left|\frac{1}{n}H_\delta(\boldsymbol{X}^n) - H(X)\right| < \epsilon.$$

Let us interpret this absolutely fundamental result. As long as we are allowed a tiny probability of error δ, compression down to $nH(X)$ bits is possible. Even if we are allowed a large probability of error, we still can compress only down to $nH(X)$ bits. This theorem therefore settles the question: our best candidate of information content definition $\frac{1}{n}H_\delta(\boldsymbol{X}^n) = \frac{1}{n}\log_2|(\mathcal{X}^n)_\delta|$, which is the δ-essential bit content *per bit*, can be made arbitrarily close to $H(X)$ which is *independent of* δ! This solves the only problem we had with H_δ, its dependence on δ; as $n \to +\infty$, $\frac{1}{n}H_\delta(\boldsymbol{X}^n)$ becomes independent of δ and tends to $H(X)$, see Fig. 4. Because by construction the random string $\boldsymbol{X}^n$ contains in expectation the same information per symbol than X, the expected information content of X is then indeed given by $H(X)$.

The proof idea is actually simple. The point is that, as n gets larger, by the law of large numbers almost all outcomes of $\boldsymbol{X}^n$ are *typical*, so that only the typical sequences need to be encoded during the compression. Let us consider for simplicity Bernoulli variables $X \sim \mathrm{Ber}(p)$. All typical sequences have approximately the same number np of ones and $n(1-p)$ of zeros: the probability that the outcome is a sequence with exactly $R = r$ ones is a binomial distribution $R \sim \mathrm{Bin}(n, p)$. The relative fluctuation of R is $O(1/\sqrt{n})$ so R concentrates onto

its mean when n gets large[12]. This implies that the only possible outcomes $\boldsymbol{x}^n$ are those with R values very close to np: this informally defines the *typical set*. The same argument extends to more general (non binary) alphabet. So the probability of a typical sequence made of nP_1 symbols $\mathcal{X}_1$, nP_2 symbols $\mathcal{X}_2$, ect, is

$$P(\boldsymbol{x}_{\rm typ}) = \prod_{i=1}^{n} P(x_{{\rm typ},i}) \approx \prod_{j=1}^{|\mathcal{X}|} P_j^{nP_j} \equiv P_{\rm typ}. \tag{9}$$

What is the information content/surprise in bits of a typical outcome?

$$h(\boldsymbol{x}_{\rm typ}) = \log_2 \frac{1}{P(\boldsymbol{x}_{\rm typ})} \approx n \sum_{j=1}^{|\mathcal{X}|} P_j \log_2 \frac{1}{P_j} = nH(X). \tag{10}$$

So the proof strategy is: i) as n gets large only typical sequences/outcomes are observed; they carry almost all the probability mass of $\boldsymbol{X}^n$. So when defining the smallest δ-sufficient subset $(\mathcal{X}^n)_\delta$ we need only to code these typical outcomes; doing so the error probability δ is small. The number of typical outcomes is exponentially large in n (this follows from the asymptotic equipartition principle), so even if we allow a risk δ very close to 1 (but independent of n) and therefore only code a small fraction of the typical sequences, there are still approximately as many at leading (exponential) order as n get large. So independently of $0 < \delta < 1$ the number $|(\mathcal{X}^n)_\delta|$ of typical sequences to code is the same at leading order. The question becomes: can we count them, i.e., evaluate $|(\mathcal{X}^n)_\delta|$? ii) By definition all typical sequences have approximately the same probability $P_{\rm typ}$, and they carry almost all the mass, so $\sum_{\{{\rm typical}\, \boldsymbol{x}^n\}} P(\boldsymbol{x}^n) \approx \#_{\rm typ} P_{\rm typ} \approx 1$, where $\#_{\rm typ} = |(\mathcal{X}^n)_\delta|$ is the number of typical sequences. This implies that there are approximately $\#_{\rm typ} \approx 1/P_{\rm typ} \approx 2^{nH(X)}$ typical sequences (from (9), (10)); we can thus approximately count them. This allows to estimate the expected information content per bit as $\frac{1}{n}\log_2 |(\mathcal{X}^n)_\delta| \approx H(X)$, which is the same as the expected information content of X by construction of $\boldsymbol{X}^n$.

Proof of the source-coding theorem. Et voila! □

1.2.3 Mutual information, and I-MMSE formula

Differential entropy and mutual information. In a continuous setting one can also define the *differential entropy*, as did Shannon, by simply replacing sums with integrals in the entropy definition:

$$H(X) \equiv \int dP(x) \ln \frac{1}{P(x)}.$$

[12] Relative fluctuations of the order $O(1/\sqrt{n})$ of macroscopic quantities like R are typical of complex systems treated in statistical mechanics. That the relative fluctuations vanish is the reason why such random systems can be analyzed and described by asymptotically (as $n \to +\infty$) deterministic observables, converging on their ensemble mean.

Unfortunately, since probability density functions can be greater than 1, the differential entropy loses an important natural properties of information measure: its positivity. E.g., the uniform distribution $\mathcal{U}([0, 1/2])$ has differential entropy $-\ln 2$. In practice this is not an issue as one can always conceptually discretize things so all our understanding in the discrete setting remains valid. The "proper" extension of Shannon entropy to the continuous setting was provided by Jaynes who defined the notion of *limiting density of discrete points* which we won't use. A better behaved quantity that maintains all its properties when going from discrete to continuous is the *mutual information* $I(X;Y) = I(Y;X)$ between two random variables. It is defined as the Kullback-Leibler divergence between their joint distribution and the product of their marginals:

$$\begin{aligned} I(X;Y) \equiv \mathrm{D}_{\mathrm{KL}}(P_{XY} \| P_X \otimes P_Y) &= \int dP_{XY}(x,y) \ln \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \geq 0. \end{aligned}$$

By working with differences of differential entropies, we recover the desirable property, for a measure of information, of non-negativity in the continuous case. The mutual information is interpreted as a measure of the mutual dependence of X and Y. It quantifies the "amount of information" obtained about one random variable through observing the other one. And indeed it cancels if and only if the variables are independent:

$$I(X;Y) = 0 \Leftrightarrow P_{XY} = P_X \otimes P_Y.$$

Another important property: for any measurable functions g_1 and g_2,

$$I(g_1(X); g_2(Y)) \leq I(X;Y),$$

with equality if both functions are invertible. This is linked to the *data processing inequality*: let X, Y and Z be random variables, where Z may depend on Y only (i.e., $P_{Z|XY} = P_{Z|Y}$). Said differently $X \to Y \to Z$ is a Markov chain. Then

$$I(X;Z) \leq I(X;Y).$$

This means than no transformation of the data can create information. This inequality is probably as fundamental as the principle of conservation of energy of isolated systems in physics.

In a inference problem where we want to recover the parameters $\boldsymbol{x}$ from the data $\boldsymbol{y}(\boldsymbol{x})$ the last form has a particularly nice interpretation: $H(Y) - H(Y|X)$ is the total information carried by the data minus the remaining unpredictability/uninformation about the data when the signal is known, which is therefore

the noise contribution. E.g., in a Gaussian denoising model,

$$y = \sqrt{\lambda}\, x^* + z \quad \Rightarrow \quad H(Y|X^*) = H(Z) = \frac{1}{2}\ln(2\pi e).$$

The mutual information is thus the information carried by the data purely about the signal. As such it quantifies the information-theoretic limits of inference, and computing it will be our main task. In the particular case of the Gaussian denoising model the explicit expression of the mutual information reads

$$\begin{aligned}
&I(X^*; Y = \sqrt{\lambda}\, X^* + Z) \\
&\quad = H(Y) - H(Z) \\
&\quad = -\int dy dP_X(x^*) \frac{e^{-\frac{1}{2}(y-\sqrt{\lambda}x^*)^2}}{\sqrt{2\pi}} \ln \int dP_X(x) \frac{e^{-\frac{1}{2}(y-\sqrt{\lambda}x)^2}}{\sqrt{2\pi}} - \frac{\ln(2\pi e)}{2} \\
&\quad = -\int dz dP_X(x^*) \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \ln \int dP_X(x) \frac{e^{-\frac{1}{2}(\sqrt{\lambda}x^*+z-\sqrt{\lambda}x)^2}}{\sqrt{2\pi}} - \frac{\ln(2\pi e)}{2} \\
&\quad = -\mathbb{E} \ln \int dP_X(x) e^{\lambda X^* x + \sqrt{\lambda} Z x - \frac{\lambda}{2}x^2} + \frac{1}{2}\mathbb{E}[(\sqrt{\lambda}X^* + Z)^2] - \frac{1}{2},
\end{aligned}$$

which finally gives

$$I(X^*; \sqrt{\lambda}\, X^* + Z) = \frac{\lambda\rho}{2} - \mathbb{E} \ln \int dP_X(x) e^{\lambda X^* x + \sqrt{\lambda} Z x - \frac{\lambda}{2}x^2}. \tag{11}$$

The conditional mutual information is defined as

$$\begin{aligned}
I(X;Y|Z) &\equiv \mathbb{E}_Z \, \mathrm{D}_{\mathrm{KL}}(P_{XY|Z} \| P_{X|Z} \otimes P_{Y|Z}) \\
&= \int dP_{XYZ}(x,y,z) \ln \frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z) P_{Y|Z}(y|z)} \\
&= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\
&= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\
&= H(X|Z) - H(X|Y,Z) \\
&= H(Y|Z) - H(Y|X,Z) \geq 0.
\end{aligned}$$

Finally, the *chain rule for mutual information* (which follows from the definition of the conditional entropy) reads:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z).$$

The I-MMSE formula for inference under Gaussian noise. As already mentioned from the mutual information one can easily deduce at least one error metric: the MMSE (4). Consider inference of a generic n-dimensional signal $\boldsymbol{x}^*$, an

outcome of $\boldsymbol{X}^* \sim P$, from data corrupted by Gaussian noise with signal-to-noise ratio λ:

$$\boldsymbol{y} = \sqrt{\lambda}\, \boldsymbol{x}^* + \boldsymbol{z}, \tag{12}$$

where $\boldsymbol{z}$ is the outcome of standard Gaussian vector with identity covariance $\mathcal{N}(0, \mathrm{I}_n)$. The *I-MMSE formula* linking the mutual information and the (average) MMSE then reads

$$\begin{aligned}\frac{d}{d\lambda} I(\boldsymbol{X}^*; \boldsymbol{Y}) &= \frac{1}{2}\mathrm{MMSE}(\boldsymbol{X}^*|\boldsymbol{Y}) \\ &= \frac{1}{2}\mathbb{E}\|\boldsymbol{X}^* - \langle \boldsymbol{X}\rangle\|_2^2 = \frac{1}{2}\mathbb{E}\big\langle \|\boldsymbol{X} - \langle \boldsymbol{X}\rangle\|_2^2\big\rangle, \end{aligned} \tag{13}$$

where recall that we distinguish between $\boldsymbol{X} \sim P(\cdot\,|\boldsymbol{y})$ and $\boldsymbol{X}^*$ the ground-truth signal despite they play symmetric roles by the Nishimori identity.

Proof. We have

$$I(\boldsymbol{X}^*; \boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}^*) = H(\boldsymbol{Y}) - H(\boldsymbol{Z})$$

with noise contribution $H(\boldsymbol{Z}) = \frac{n}{2}\ln(2\pi e)$. Let us then compute

$$\begin{aligned}\frac{d}{d\lambda} H(\boldsymbol{Y}) &= -\frac{d}{d\lambda}\int dP(\boldsymbol{x}^*) d\boldsymbol{y} \frac{e^{-\frac{1}{2}\|\boldsymbol{y}-\sqrt{\lambda}\boldsymbol{x}^*\|^2}}{(2\pi)^{n/2}} \ln \int dP(\boldsymbol{x}) \frac{e^{-\frac{1}{2}\|\boldsymbol{y}-\sqrt{\lambda}\boldsymbol{x}\|^2}}{(2\pi)^{n/2}} \\ &= -\frac{d}{d\lambda}\int dP(\boldsymbol{x}^*) d\boldsymbol{z} \frac{e^{-\frac{1}{2}\|\boldsymbol{z}\|^2}}{(2\pi)^{n/2}} \ln \int dP(\boldsymbol{x}) \frac{e^{-\frac{1}{2}\|\boldsymbol{z}-\sqrt{\lambda}(\boldsymbol{x}-\boldsymbol{x}^*)\|^2}}{(2\pi)^{n/2}} \\ &= \frac{1}{2\sqrt{\lambda}}\mathbb{E}_{\boldsymbol{X}^*,\boldsymbol{Z}}\big\langle (\boldsymbol{Z} + \sqrt{\lambda}(\boldsymbol{X}^* - \boldsymbol{X}))\cdot(\boldsymbol{X}^* - \boldsymbol{X})\big\rangle, \end{aligned} \tag{14}$$

where the Gibbs-bracket $\langle - \rangle$ is the expectation acting on replica $\boldsymbol{X}$ with distribution

$$P(\boldsymbol{x}|\boldsymbol{y}(\boldsymbol{x}^*, \boldsymbol{z})) = \frac{P(\boldsymbol{x}) e^{-\frac{1}{2}\|\boldsymbol{z}-\sqrt{\lambda}(\boldsymbol{x}-\boldsymbol{x}^*)\|^2}}{\int dP(\boldsymbol{x}') e^{-\frac{1}{2}\|\boldsymbol{z}-\sqrt{\lambda}(\boldsymbol{x}'-\boldsymbol{x}^*)\|^2}}.$$

Now we use a very useful Gaussian integration by part formula, sometimes called Steins's lemma. It says that, for any bounded function $\boldsymbol{g} : \mathbb{R}^n \mapsto \mathbb{R}^n$ of a standard Gaussian random vector $\boldsymbol{Z} \sim \mathcal{N}(0, \mathrm{I}_n)$ we have

$$\mathbb{E}[\boldsymbol{Z}\cdot\boldsymbol{g}(\boldsymbol{Z})] = \mathbb{E}\,\nabla\cdot\boldsymbol{g}(\boldsymbol{Z}) = \mathbb{E}\,\mathrm{div}\,\boldsymbol{g}(\boldsymbol{Z}).$$

Recall we use the $\cdot$ notation for the scalar product. This formula applied to a Gibbs-brackets associated to a general Hamiltonian depending on Gaussian noise

yields

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Z}\cdot\langle\boldsymbol{h}(\boldsymbol{X})\rangle] &= \mathbb{E}\,\nabla_{\boldsymbol{Z}}\cdot\frac{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}\boldsymbol{h}(\boldsymbol{x})}{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}} \\
&= -\mathbb{E}\,\frac{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}\boldsymbol{h}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{Z}}\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}} \\
&+\mathbb{E}\Big[\frac{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}\boldsymbol{h}(\boldsymbol{x})}{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}}\cdot\frac{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}\nabla_{\boldsymbol{Z}}\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}{\int dP(\boldsymbol{x})e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Z})}}\Big] \\
&= -\mathbb{E}\big\langle\boldsymbol{h}(\boldsymbol{X})\cdot\nabla_{\boldsymbol{Z}}\mathcal{H}(\boldsymbol{X};\boldsymbol{Z})\big\rangle+\mathbb{E}\big[\big\langle\boldsymbol{h}(\boldsymbol{X})\big\rangle\cdot\big\langle\nabla_{\boldsymbol{Z}}\mathcal{H}(\boldsymbol{X};\boldsymbol{Z})\big\rangle\big].
\end{aligned}$$

Applied to (14), where the "Hamiltonian" is $\frac{1}{2}\|\boldsymbol{z}-\sqrt{\lambda}(\boldsymbol{x}-\boldsymbol{x}^*)\|^2$, it gives

$$\begin{aligned}
\frac{d}{d\lambda}H(\boldsymbol{Y}) &= \frac{1}{2}\mathbb{E}\big[\big\langle\|\boldsymbol{X}^*-\boldsymbol{X}\|^2\big\rangle+\frac{1}{\sqrt{\lambda}}\nabla_{\boldsymbol{Z}}\cdot\langle\boldsymbol{X}^*-\boldsymbol{X}\rangle\big] \\
&= \frac{1}{2}\mathbb{E}\big[\big\langle\|\boldsymbol{X}^*-\boldsymbol{X}\|^2\big\rangle-\frac{1}{\sqrt{\lambda}}\big\langle(\boldsymbol{X}^*-\boldsymbol{X})\cdot(\boldsymbol{Z}+\sqrt{\lambda}(\boldsymbol{X}^*-\boldsymbol{X}))\big\rangle \\
&\qquad+\frac{1}{\sqrt{\lambda}}\big\langle(\boldsymbol{X}^*-\boldsymbol{X})\big\rangle\cdot\big\langle\boldsymbol{Z}+\sqrt{\lambda}(\boldsymbol{X}^*-\boldsymbol{X})\big\rangle\big] \\
&= \frac{1}{2}\mathbb{E}\|\boldsymbol{X}^*-\langle\boldsymbol{X}\rangle\|^2 \\
&\overset{\text{N}}{=} \frac{1}{2}\mathbb{E}\big\langle\|\boldsymbol{X}-\langle\boldsymbol{X}\rangle\|_2^2\big\rangle
\end{aligned}$$

which is also equal to $\frac{d}{d\lambda}I(\boldsymbol{X}^*;\boldsymbol{Y})$, and N stands for "Nishimori". □

1.3 Statistical mechanics 101, and links with Bayesian inference

What is statistical mechanics? A high-level tentative. Statistical physics has been developed in the beginning of the 20th century in order to describe matter in its various forms, such as solid, fluid or gaseous states. At the same time was developed another pillar of modern physics[13], namely quantum mechanics, that allows to fully describe the dynamics of atoms and molecules a the *microscopic level* (through the Schrödinger equation). So you might wonder: *why should we need yet another theory in order to describe matter if quantum mechanics already does precisely that?* That is a fair question. A first answer is practical: imagine you are interested in describing the precise dynamics of all molecules of water in a given drop of water. Even if you would have access to a (classical) computer with a CPU as large as the observable universe, and that could compute for as long

[13]The three pillars of modern physics are 1) Einstein's theory of relativity that describes the first (and weaker) fundamental physical force, namely the gravity; 2) quantum mechanics that describes in a unified way the three remaining fundamental forces: the electromagnetic force, the strong nuclear force (responsible of the stability of the atomic nucleus through preventing protons to repeal each other in the atomic nucleus in spite that they have same electric charge), and the weak nuclear force (responsible of the radioactivity); 3) statistical physics which describes complex systems.

as the age of the universe (13.8 billion years), yet you could not solve the exact equations of quantum mechanics that, *in principle*, describe exactly the motion of the molecules (i.e., positions and velocities). And even *assuming you could* compute that by solving exactly the Schrödinger equation describing all the atoms in this drop of water, storing all this information on a physical memory (on a modern computer memory technology like SSD, or even on a highly optimized medium like DNA) would require a volume larger than the observable universe; this is doomed[14]...

A second answer is more "philosophical": do we really care about knowing the precise motion of each and every single atom forming the system of interest, like this drop of water? In general the answer is no. What we care about are *macroscopic quantities*, i.e., that are "averaged" over the atoms. For example, what you could be interested in is the density of particles in some medium, or how "disordered" the atoms are (note that the notion of disorder in meaningless at the single atom level). This will give you meaningful information about the state of matter: very disordered and of low density = a gas, still disordered enough and of high density = a fluid, ordered state of high density = a solid. Another example: a material is magnetic if its magnetization, which measures the (averaged over the atoms) "alignment" of the spins, is non-zero. Here again we do not care about the microscopic details of which atom has a upward spin, and which ones are oriented downwards.

What we care about are averaged quantities which describe the system as a whole, i.e., at a macroscopic level, not microscopic one.

This is also because these are the quantities that can be experimentally measured with an apparatus (as opposed to single atoms motions which are hard to track): density, temperature, viscosity, magnetization, concentration etc.

A final answer is summarized by a quote[15] from Philip Anderson, winner of the physics nobel prize in 1977 for his investigations into the electronic structure of magnetic and disordered systems, which allowed for the development of electronic switching and memory devices in computers:

More is different.

This means what it says: complex systems, i.e., systems made of a large number of interacting components/entities/variables (like the atoms in matter), such as (but not restricted to, as we will see) solids, fluids or gas *cannot be described as a collection of the behaviors of its individual components.* Another way to phrase it is: *the whole is more than the sum of its parts.* It is impossible to describe a complex

[14] Note that the numbers appearing in statistical mechanics are *order of magnitudes* larger that those appearing in astrophysics and cosmology. For examples there are typically 10^{24} molecules of water in a drop of water, that is one million billion billion of them. Check the definition of the Avogadro number on Wikipedia.

[15] Actually this is the name of one of its articles.

system by first individually analyzing, even very precisely, all its components and to then try to combine all this knowledge together: *a complex system has to be considered as a whole.*

This leads to another fundamental concept:

Emergent phenomena.

Emergence occurs when an entity is observed to have properties its parts do not have on their own. These properties or behaviors emerge only when the parts interact in a wider whole. You cannot explain why a solid is what it is (a highly structured system with particular physical properties like conductivity, resistance to mechanical stress, optical properties etc) from the properties of the atoms that form it. You cannot understand the dynamics of bird flocks from the understanding of a single bird behavior. Weather cannot be predicted from the knowledge of single air molecules dynamics. A financial crisis cannot simply be explained by the behavior of individuals agents. The success of algorithms to solve complex tasks cannot be reduced to tracking how single bits are processed. And so on and so forth.

The concept of emergence is intimately linked to another absolutely fundamental concept. Its understanding and quantification in information processing tasks will be our main goal:

Phase transitions.

A phase transition, which is an emergent phenomenon, is when a complex system experiences a quite sudden change of certain macroscopic/global properties –called observables– when some external parameter(s) –called control parameter(s)– is varied (by an external operator, like a physicist in a lab doing experiments and changing the temperature or the pressure to see how a medium behave, or a programmer testing its computer code for various parameters). You already know what is a phase transition, you experience it daily when you cook your pastas: going from a solid state (i.e., ice) to liquid *precisely* at 0 degree Celsius (at atmospheric pressure), or from liquid to gas at 100 degrees, are two of the very many possible phase transitions that occur in nature. In this case the phase of matter is described by observables like the density and level of ordering of the atoms, and the control parameter is the temperature. There exist various types of phase transition; sometimes they are quite smooth (these are called phase transition of the "second order" type), and sometimes very sharp and discontinuous (of the "first order" type).

Other more "exotic" examples of phase transitions could be: the recovery of a souvenir by the brain once enough stimuli in the direction of the memorized pattern are provided (a simple model of associative memory is the hopfield model). Here the observable is the level of recovery of the souvenir and the control parameter is the amount of stimuli; a crack in the financial market, where suddenly all prices drop all together; the sudden "rigidity" transition that happens when

you randomly pack enough balls in a box (this is called the "jamming transition", and this is related to computer memory optimization or error correcting codes in communication). Or in information processing tasks: there is a critical noise level in a communication channel above which communication becomes impossible. This limit is called the *Shannon capacity* and really is nothing but a phase transition. The observable here is the quality of recovery of the transmitted signal/information, the control parameter is the noise (i.e., corruption) level, or the rate of information transmission. A final one. Say you want to train an algorithm that, given a large amount of labeled training examples, is able to distinguish pictures of dogs and cats. There exists a minimum number of training examples below which, no matter the power of the computer, the algorithm will never be able to properly classify the images. The observable is the classification performance of the algorithm, the control parameter is the size of the training set.

"Statistical mechanics" therefore means that we study the motion (i.e., derive the mechanics) of averaged quantities describing potentially very complex systems, and this thanks to statistics. In other words:

Statistical mechanics allows to predict macroscopic/global properties of a complex system from the knowledge of the governing equations at the microscopic level (i.e., from the local properties of its simple components). It links the microscopic, overall unpredictable, world and the macroscopic observable one using statistics.

Here we are interested in "information processing" tasks. This means that *the systems of interest are algorithms dealing with large amount of data.* Again, for these processes like for the atoms in a solid, we are not really interested in the fine details, but more by meaningful averaged "global" quantities that describe the system at the scale of interest. In information processing, the macroscopic quantity of interest is often the performance of an algorithm in solving the task it has been designed for.

Equilibrium. A system is at equilibrium when there are no global fluxes and all macroscopic quantities remain unchanged. Mathematically this implies:

A random complex system is at equilibrium if it explores its allowed configurations (i.e., microscopic states) following a probability distribution taking a special form, called "Gibbs-Boltzmann" distribution.

Consider a random system described by the ensemble $(\boldsymbol{X}, \mathcal{X}^n, P(\cdot\,; \boldsymbol{y}, \beta))$, where $\boldsymbol{X} = (X_1, \dots, X_n)$ is a n-dimensional random vector with possible outcomes $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$. The n-dimensional vector $\boldsymbol{x}$ is a *configuration/microscopic state* (i.e., in configuration $\boldsymbol{x}$ its first component, like a spin or an atomic position in physics, takes value x_1, the second x_2 and so forth). In physics the X_i's are also called *degrees of freedom*, or *spins*. The system may also depend on other

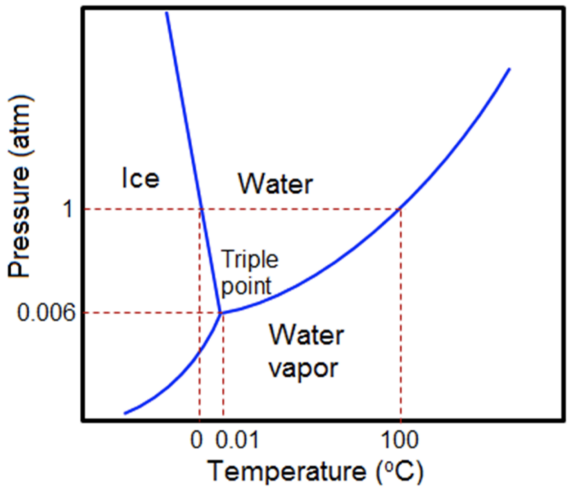


Figure 5: Phase diagram of water in the (Pressure, Temperature) plane, the two control parameters in this case, with boundaries delimited by phase transitions.

fixed, or *quenched*, parameters $\boldsymbol{y}$ (i.e., that do not fluctuate according to the Gibbs-Boltzmann distribution defined below). The quenched $\boldsymbol{y}$ are also called *disorder*.

This random system is at *equilibrium* if its probability distribution to be observed in the microscopic configuration $\boldsymbol{x}$ takes the *Gibbs-Boltzmann* form:

$$P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{y}, \beta) = P(\boldsymbol{x}; \boldsymbol{y}, \beta) = \frac{\exp\{-\beta \mathcal{H}(\boldsymbol{x}; \boldsymbol{y})\}}{\mathcal{Z}(\boldsymbol{y}, \beta)}. \tag{15}$$

The normalization constant, also called *partition function*, is

$$\mathcal{Z}(\boldsymbol{y}, \beta) = \sum_{\boldsymbol{x} \in \mathcal{X}^n} \exp\{-\beta \mathcal{H}(\boldsymbol{x}; \boldsymbol{y})\}. \tag{16}$$

The partition function essentially contains all relevant information about the system. $\mathcal{X}^n$ is the ensemble of allowed configurations for the system, the *configuration space*, such as the positions for the atoms in a the medium, in which case $\mathcal{X}^n = \mathbb{R}^n$ and the sum in the partition function is replaced by an integral; or for magnetic materials made of spins $\mathcal{X}^n = \{-1, 1\}^n$, etc.

The function $\mathcal{H}$ is the *Hamiltonian* of the system, i.e., its *energy/cost function*: $\mathcal{H}(\boldsymbol{x}; \boldsymbol{y})$ is the energy of the system when in microscopic state/configuration $\boldsymbol{x}$. The Hamiltonian defines the system. It fully dictates the (random) behavior of the system through the Gibbs-Boltzmann distribution.

The Gibbs-Boltzmann distribution follows from a *maximum entropy principle*. It is the solution of the following constrained optimization problem: Find P such that $H(P)$ is maximized under $\sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x}) = 1$ as well as the fixed average energy constraint $\langle \mathcal{H}(\boldsymbol{X}; \boldsymbol{y}) \rangle_P \equiv \sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x}) \mathcal{H}(\boldsymbol{x}; \boldsymbol{y}) = E$.

Finally β is the *inverse temperature* and plays the role of Lagrange multiplier in solving the constraint optimization problem above. Physically speaking, it is a

control parameter that allows to tune the "amount of randomness" in the system. This randomness is, in physical systems, due to thermal fluctuations, i.e., uncontrolled interactions of the system with the outside world/environment (also called "thermal bath"). We can see that by looking at the two extremes: when $\beta \to +\infty$ (or equivalently the temperature approaches 0_+) then the Gibbs-Boltzmann distribution simply becomes

$$\lim_{\beta\to+\infty} P(\boldsymbol{x};\boldsymbol{y},\beta) = \frac{1}{\mathcal{Z}(\boldsymbol{y},+\infty)}\mathbf{1}\big(\boldsymbol{x} \in \{\mathrm{argmin}_{\boldsymbol{x}'}\, \mathcal{H}(\boldsymbol{x}';\boldsymbol{y})\}\big) \tag{17}$$

where $\mathbf{1}(A)$ is the indicator function. In the zero temperature limit the only allowed configurations are the ones with minimum energy; these are called *ground states*. The system is trapped in one of these ground states forever, it becomes deterministic ("frozen") as there is not anymore thermal energy from the outside that can flow in the system to perturb it through thermal fluctuations. In this case the partition function $\mathcal{Z}(\boldsymbol{y},+\infty)$ simply counts the number of ground states. In general, finding the set of ground states or even a single one is not an easy optimization problem at all.

If instead the temperature diverges, i.e., $\beta \to 0_+$, this probability distribution becomes uniform over all configurations:

$$\lim_{\beta\to 0_+} P(\boldsymbol{x};\boldsymbol{y},\beta) = \frac{1}{\mathcal{Z}(\boldsymbol{y},0)} = \frac{1}{|\mathcal{X}^n|}. \tag{18}$$

In this case the partition function tends to the cardinal/volume of the configuration space (that generally grows exponentially with the number of variables n). Tuning the inverse temperature β from 0_+ to $+\infty$ allows to drive the system from totally random (like a gas at very high temperature, where atoms essentially never interact) to a purely deterministic one (like water frozen at almost the absolute zero, where atoms are trapped in an ordered crystalline structure and nothing moves anymore). In between exists a full *phase diagram*, which encodes in which *macroscopic state* –states described by macroscopic global quantities that are averages over the microscopic configurations– is the system as a function of external parameters, with boundaries delimited by phase transitions.

A quantity of paramount importance in statistical mechanics is the *free energy*:

$$F_n(\boldsymbol{y}) \equiv -\frac{1}{n\beta}\ln \mathcal{Z}(\boldsymbol{y}).$$

This is because from this quantity we can then locate the phase transitions: these are points where the large n limit of the free energy is non-analytic. From the free energy we can also derive all relevant macroscopic/global quantities of interest (average energy density, magnetization, etc), by simply taking derivatives with respect to control parameters (β, external magnetic field, etc). We will shortly show that this object is, for properly defined models, well behaved in the sense that i) it is self-averaging, i.e., it concentrates onto its mean:

$$\lim_{n\to+\infty} \mathbb{E}[(F_n(\boldsymbol{Y}) - f_n)^2] = 0,$$

where the average free energy, also called *quenched free energy*, is

$$f_n \equiv \mathbb{E}\, F_n(\boldsymbol{Y}) = -\frac{1}{n\beta}\mathbb{E} \ln \mathcal{Z}(\boldsymbol{Y}).$$

When this is true it justifies an *ensemble analysis*: instead of considering a particular realization of the parameters $\boldsymbol{y}$ defining the Hamiltonian of the model, we can average over $\boldsymbol{Y}$ and this will give the same results. This is the reason why statistical mechanics makes sense. Moreover $ii)$ its *thermodynamic limit* exists:

$$f = \lim_{n\to+\infty} f_n.$$

These two points imply that

$$\lim_{n\to+\infty} \mathbb{E}[(F_n(\boldsymbol{Y}) - f)^2] = 0,$$

and therefore convergence in probability: for any ϵ,

$$\lim_{n\to+\infty} P(|F_n(\boldsymbol{Y}) - f| > \epsilon) = 0.$$

Bayesian inference as a disordered statistical mechanical problem. A posterior distribution (2) can be written as a Gibbs-Boltzmann distribution with $\beta = 1$ and Hamiltonian

$$\mathcal{H}(\boldsymbol{x}; \boldsymbol{y}) = -\ln P(\boldsymbol{x}) - \ln P(\boldsymbol{y}|\boldsymbol{x}).$$

The second term above is the *log-likelihood*. The partition function is then the evidence:

$$\mathcal{Z}(\boldsymbol{y}) = P(\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}).$$

Finally the average free energy is therefore equal to the differential entropy of the data (or Shannon entropy if the data is discrete) divided by the number of parameters to infer:

$$f_n \equiv -\frac{1}{n}\mathbb{E} \ln \mathcal{Z}(\boldsymbol{Y}) = \frac{1}{n}\int d\boldsymbol{y}\mathcal{Z}(\boldsymbol{y}) \ln \frac{1}{\mathcal{Z}(\boldsymbol{y})} = \frac{1}{n}H(\boldsymbol{Y}). \tag{19}$$

So we see that statistical physics, and Bayesian inference and information theory are deeply related. The data $\boldsymbol{y}$, or equivalently the ground-truth signal and noise $\boldsymbol{x}^*$, $\boldsymbol{z}$ play the role of quenched disorder in the statistical mechanics analogy. The mutual information is therefore related to the free energy through a simple additive term:

$$\frac{1}{n}I(\boldsymbol{X}^*; \boldsymbol{Y}) = \frac{1}{n}H(\boldsymbol{Y}) - \frac{1}{n}H(\boldsymbol{Y}|\boldsymbol{X}^*) = f_n - \frac{1}{n}H(\boldsymbol{Y}|\boldsymbol{X}^*). \tag{20}$$

The noise contribution $\frac{1}{n}H(\boldsymbol{Y}|\boldsymbol{X}^*)$ is generally simple to compute because we often restrict ourselves to settings where the noise is independent for each data point, so that $\frac{1}{n}H(\boldsymbol{Y}|\boldsymbol{X}^*) = \frac{m}{n}H(Y_1|\boldsymbol{X}^*) = \frac{m}{n}H(Z_1)$ where Z_1 represents the noise in the process (not necessarily Gaussian), m is the number of conditionnally (on $\boldsymbol{X}^*$) independent data points. So the real task is to compute the entropy density of the evidence $\frac{1}{n}H(\boldsymbol{Y})$, i.e., the free energy exactly like in physics.

2 Information-theoretic limits

In this section we will focus our efforts on establishing the information-theoretic limits of inference in the Wigner spike model, a simple probabilistic model of principal component analysis. This model is rich enough so that it contains all features of more complex inference problems, and remains simple enough to fully analyze it in a finite amount of time. We will study it through the lens of various advanced mean-field techniques, that all take roots in the statistical mechanics of disordered systems.

Spiked Wigner model. We consider a signal-vector $\boldsymbol{x}^* = (x_i^*)_{i=1}^n$ with bounded components. Its entries x_i^* were drawn i.i.d. from the same prior P_X supported, without loss of generality, on $[-1, 1]$ and with second moment ρ. You can keep in mind as a running example the simple case $P_X = \mathrm{Ber}(\rho)$ where $\rho \in (0, 1]$ controls the sparsity of the vector. The symmetric data-matrix $\boldsymbol{y} = (y_{ij})_{i,j=1}^n$ is obtained through the following generative process, or "observation model":

$$y_{ij} = \sqrt{\frac{\lambda}{n}} x_i^* x_j^* + z_{ij}, \quad 1 \le i < j \le n, \tag{21}$$

where the noise-matrix enries z_{ij} are i.i.d. outcomes of a standard Gausian $\mathcal{N}(0, 1)$ for $i < j$, with $z_{ij} = z_{ji}$. The inference task is to estimate the rank-one "spike" $(x_i^* x_j^*)$ from the data (y_{ij}) corrupted by a Wigner noise matrix (z_{ij}). Some applications of the Wigner spiked model include:

- *Sparse PCA:* In the simplest case the prior $P_X = \mathrm{Ber}(\rho)$. The task is to estimate the hidden sparse, low-rank representation $\boldsymbol{x}^* \otimes \boldsymbol{x}^*$ (and then $|\boldsymbol{x}^*|$ by eigenvalue decomposition) of $\boldsymbol{y}$.
- *Submatrix localization:* Again $P_X = \mathrm{Ber}(\rho)$. One has then to extract a submatrix of $\boldsymbol{y}$ of size $\rho n \times \rho n$ with larger mean.
- *Community detection in the Stochastic Block Model:* Recovering two communities of size ρn and $(1-\rho)n$ in a dense SBM of n vertices is "equivalent" to the spiked Wigner model with

$$P_X = \rho \delta_{\sqrt{(1-\rho)/\rho}} + (1-\rho)\delta_{-\sqrt{\rho/(1-\rho)}}. \tag{22}$$

- $\mathbb{Z}/2$ *synchronization:* The prior is Rademacher $P_X = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$. The task is to determine the nodes states $\boldsymbol{x}^* \in \{-1, 1\}^n$ (up to a global sign) from noisy pairwise products $\boldsymbol{y}$.

There exists a non-symmetric version called *spiked Wishart model*:

$$y_{ij} = \sqrt{\frac{\lambda}{n}} u_i^* v_j^* + z_{ij}, \quad i \in \{1, \dots, n\},\ j \in \{1, \dots m = \Theta(n)\}.$$

Some applications of it include

- *Sparse PCA/spiked covariance model:* The signal is $\boldsymbol{u}^*$ with $P_U = \mathrm{Ber}(\rho)$. The matrix $\boldsymbol{v}^*$ with $P_V = \mathcal{N}(0, 1)$ is then interpreted as a noise matrix. The observations $\boldsymbol{y}_j$, i.e., the columns of $\boldsymbol{y}$ are then i.i.d. outcomes of

  $$\mathcal{N}\Big(0, \mathrm{I}_n + \frac{\lambda}{n} \boldsymbol{u}^* \otimes \boldsymbol{u}^*\Big).$$

  The goal is to recover the spike $\boldsymbol{u}^* \otimes \boldsymbol{u}^*$ of the covariance.

- *High-dimensional clustering of m noisy n-dimensional points in k clusters:* Consider the rank-k version of the spiked Wishart model where $\boldsymbol{u}^* \in \mathbb{R}^{n \times k}$, $\boldsymbol{v}^* \in \mathbb{R}^{m \times k}$:

  $$y_{ij} = \sqrt{\frac{\lambda}{n}} \sum_{\ell=1}^{k} u_{i\ell}^* v_{j\ell}^* + z_{ij}.$$

  Then the columns of $\boldsymbol{u}^*$ are n-dimensional vectors representing centers of k different clusters. The lines of $\boldsymbol{v}^*$ are uniform permutations of the k-dimensional vector with a single one $(1, 0, \dots, 0)$ selecting the cluster to which belong the noisy point $\boldsymbol{y}_j$ with $j \in \{1, \dots, m\}$ (the columns of $\boldsymbol{y}$).

In the spiked Wigner model we have very many $\Theta(n^2)$ observations for reconstructing few parameters, i.e., $\Theta(n)$, but each observation is mostly noise because of the $1/\sqrt{n}$ scaling. The total signal-to-noise ratio (SNR) per parameter: $\#\,\text{observations} \cdot \mathrm{SNR}_{\mathrm{obs}} / \#\,\text{parameters to infer}$, where $\mathrm{SNR}_{\mathrm{obs}}$ denotes the SNR per observation, needs to be $\Theta(1)$ to make the inference problem not trivial nor impossible. In the present case we have access to $n(n-1)/2$ independent observations and $\mathrm{SNR}_{\mathrm{obs}} = \mathbb{E}[(X_1 X_2)^2]\lambda/n = \rho^2\lambda/n$. Therefore we check

$$\frac{\frac{n(n-1)}{2} \times \frac{\rho^2\lambda}{n}}{n} = \frac{\rho^2\lambda}{2} + O(1/n) = \Theta(1)\,. \tag{23}$$

This explains the presense of the scaling $1/\sqrt{n}$ in the observation model (21). Note that any other scaling would make the estimation task either trivial if the total SNR per parameter tends to infinity, or impossible if it tends to zero.

We suppose that we are in a *Bayesian optimal setting* where the prior P_0 as well as the noise distribution are known so that the true posterior is known.

$$P(\boldsymbol{x}|\boldsymbol{y}) \propto \prod_{i=1}^{n} P_X(x_i) \frac{1}{(2\pi)^{\frac{n(n-1)}{4}}} \exp\Big\{ -\frac{1}{2} \sum_{1\le i<j\le n} \Big(y_{ij} - \sqrt{\frac{\lambda}{n}} x_i x_j\Big)^2 \Big\} \tag{24}$$

$$= \frac{1}{\mathcal{Z}_n(\boldsymbol{y})} \prod_{i=1}^{n} P_X(x_i) \exp\{-\mathcal{H}(\boldsymbol{x};\boldsymbol{y})\}, \tag{25}$$

where, after simplifying all the x-independent terms with the normalization, the Hamiltonian and partition function are

$$\begin{aligned} \mathcal{H}(\boldsymbol{x};\boldsymbol{y}) &= \sum_{i<j} \Big(\frac{\lambda}{2n} x_i^2 x_j^2 - y_{ij} \sqrt{\frac{\lambda}{n}} x_i x_j \Big) \\ &= \sum_{i<j} \Big(\frac{\lambda}{2n} x_i^2 x_j^2 - \frac{\lambda}{n} x_i^* x_j^* x_i x_j - \sqrt{\frac{\lambda}{n}} z_{ij} x_i x_j \Big), \\ \mathcal{Z}_n(\boldsymbol{y}) &= \int \prod_{i=1}^{n} dP_X(x_i) \exp\{-\mathcal{H}(\boldsymbol{x};\boldsymbol{y})\}. \end{aligned} \tag{26}$$

We used the definition of $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}^*, \boldsymbol{z})$ to express the Hamiltonian as a function of the independent variables. Due to these convenient simplifications, you can easily check that the more convenient free energy expression

$$f_n \equiv -\frac{1}{n} \mathbb{E} \ln \mathcal{Z}_n(\boldsymbol{Y}) = -\frac{1}{n} \mathbb{E} \ln \mathcal{Z}_n(\boldsymbol{X}^*, \boldsymbol{Z}),$$

(here we emphasize that we can equivalently express the free energy as a function of the data or of the independent signal and noise) is related to the Shannon entropy of the data and the mutual information through

$$\begin{aligned} \frac{1}{n} H(\boldsymbol{Y}) &= f_n + \frac{n-1}{4} \ln(2\pi e) + \frac{\rho^2 \lambda}{4} \frac{n-1}{n}, \\ \frac{1}{n} I(\boldsymbol{X}^*; \boldsymbol{Y}) &= f_n + \frac{\rho^2 \lambda}{4} \frac{n-1}{n}. \end{aligned} \tag{27}$$

Working with this definition of the free energy f_n rather than $\frac{1}{n} H(\boldsymbol{Y})$ will slightly simplify computations. As before, the expectation w.r.t. the posterior is, for any bounded function g,

$$\langle g(\boldsymbol{X}) \rangle \equiv \mathbb{E}[g(\boldsymbol{X})|\boldsymbol{y}] = \int dP(\boldsymbol{x}|\boldsymbol{y})\, g(\boldsymbol{x}).$$

The notation $\mathbb{E}$ is the expectation w.r.t. the quenched variables instead. Objects of the form $\langle g(\boldsymbol{X}) \rangle$ are functions of the quenched variable(s) $\boldsymbol{y}$, or equivalently

$\boldsymbol{x}^*$, $\boldsymbol{z}$. We keep the notation $\boldsymbol{X}$ for a sample from the posterior $P(\cdot\,|\boldsymbol{y})$.

Link with the mean-field spin glass, and gauge symmetry in the binary case. The most studied disordered statistical mechanics model is the mean-field spin glass, also called Sherrington-Kirkpatrick (SK) model. Let $\boldsymbol{x} \in \{-1,1\}^n$. The Hamiltonian of the SK model reads

$$\mathcal{H}_{\rm SK}(\boldsymbol{x};\boldsymbol{z}) = -\sum_{1\le i<j\le n} \frac{1}{\sqrt{n}} z_{ij} x_i x_j, \qquad P_{\rm SK}(\boldsymbol{x};\boldsymbol{z}) = \frac{e^{-\beta\mathcal{H}_{\rm SK}(\boldsymbol{x};\boldsymbol{z})}}{\mathcal{Z}_{\rm SK}(\beta,\boldsymbol{z})},$$

where z_{ij} are i.i.d. outcomes of a standard Gaussian random variable $\mathcal{N}(0,1)$. Notice that this Hamiltonian is the same as the one of the spiked Wigner model with binary variables (again only the x-dependent terms are relevant when defining the Hamiltonian):

$$\mathcal{H}(\boldsymbol{x};\boldsymbol{y}) = -\sum_{i<j} y_{ij}\sqrt{\frac{\lambda}{n}} x_i x_j = -\sum_{i<j}\Big(\frac{\lambda}{n} x_i^* x_j^* + \sqrt{\frac{\lambda}{n}} z_{ij}\Big) x_i x_j \tag{28}$$

when only the noise term is present and λ is set to 1. The additional signal-related term $-\sum_{i<j}\frac{\lambda}{n} x_i^* x_j^* x_i x_j$ is called *planted* term, and inference models are *planted models*. The planted term plays the role of external magnetic field that tends to align the spins in the signal direction.

Observe that the Hamiltonian of the spiked Wigner model is invariant under the change of variable $y_{ij} \to y_{ij} x_i^* x_j^*$ and $x_i \to x_i x_i^*$, which implies invariance of the associated Gibbs-Boltzmann distribution and Gibbs-bracket; this change of variable is called a *gauge transformation* because of this invariance. Using this gauge transformation it is an exercise to see that the spiked Wigner model with binary variables $x_i^* = \pm 1$ is perfectly equivalent, i.e. we have equality of free energy and of the expectation of any observable, to the spiked Wigner model with $x_i^* = 1$ for all $i = 1,\dots,n$:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}\big\langle g\big(\boldsymbol{Y},\boldsymbol{X}^*,\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(k)}\big)\big\rangle \\ &\qquad = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{1}}\big\langle g\big(\boldsymbol{Y},\boldsymbol{X}^*=\boldsymbol{1},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(k)}\big)\big\rangle_{\boldsymbol{X}^*=\boldsymbol{1}}\end{aligned}$$

where in the second bracket the signal $\boldsymbol{X}^*$ is set to the all ones vector. By the Nishimori identity this is also equal to

$$\begin{aligned}\cdots &\overset{\rm N}{=} \mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}\big\langle g\big(\boldsymbol{Y},\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(k)}\big)\big\rangle \\ &\qquad = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{1}}\big\langle g\big(\boldsymbol{Y},\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(k)}\big)\big\rangle_{\boldsymbol{X}^*=\boldsymbol{1}}.\end{aligned}$$

The $\boldsymbol{X}^{(a)}$'s are i.i.d. replicas drawn from the posterior associated with the bracket acting on them. This implies, e.g.,

$$\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{1}}\big\langle Q(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})\big\rangle_{\boldsymbol{X}^*=\boldsymbol{1}} \overset{\rm N}{=} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{1}}\big\langle M(\boldsymbol{X})\big\rangle_{\boldsymbol{X}^*=\boldsymbol{1}} \tag{29}$$

where the overlap between replicas $Q(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \frac{1}{n}\sum_i x_i^{(1)} x_i^{(2)}$ and the magnetization $M(\boldsymbol{x}) = \frac{1}{n}\sum_i x_i$. Again by Nishimori (29) is also equal to the expected overlap with the ground truth (or between two replicas by the Nishimori identity) in the model with $x_i^* = \pm 1$

$$\cdots \overset{\text{N}}{=} \mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}\big\langle Q(\boldsymbol{X}, \boldsymbol{X}^*)\big\rangle \overset{\text{N}}{=} \mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}\big\langle Q(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})\big\rangle.$$

In the binary case the gauge symmetry therefore allows to remove the dependence in the planted signal. The planted SK model, i.e., the binary spiked Wigner model, is then equivalent to the original SK model but with $\beta = 1$ and noise that is ferromagnetically biased:

$$\mathcal{H}_{\text{planted SK}}(\boldsymbol{x}; \tilde{\boldsymbol{y}}) = -\sum_{1\le i<j\le n} \tilde{y}_{ij} x_i x_j$$

where $\tilde{y}_{ij}$ are now i.i.d. outcomes of $\mathcal{N}\big(\frac{\lambda}{n}, \frac{\lambda}{n}\big)$. These observations are due to Nishimori, and the consequences such as (29) are the "original Nishimori identities". Identity (5) is the generalization to models without gauge symmetry. Calling in general m and σ^2 the mean and variance of the quenched interactions z_{ij}, in the plan (m, σ^2) the line

$$m = \sigma^2$$

(when $\beta = 1$) is called "Nishimori line". Spin glass models living on the Nishimori line can always be re-interpreted as inference problems in the Bayesian optimal setting. For more general inference problems than ones with a gauge symmetry, the Nishimori line is defined by the parameters values such that (5) is verified, i.e., such that we are in the Bayesian optimal setting: the assumed parameters in the posterior match the true parameters used for generating the data.

2.1 Replica symmetric formula for the mutual information

Replica symmetric formula. The main result we will prove is the following asymptotic formula for the mutual information:

Theorem 1 (Replica symmetric formula)**.** *Let the signal prior P_X be bounded with second moment ρ. The mutual information for the spiked Wigner model verifies*

$$\lim_{n\to+\infty} \frac{1}{n} I(\boldsymbol{X}^*; \boldsymbol{Y}) = \inf_{q\in[0,\rho]} i^{(\text{RS})}(q; \lambda, \rho),$$

where, letting $X^* \sim P_X$ and $Z \sim \mathcal{N}(0,1)$, the replica symmetric potential *is*

$$\begin{aligned} i^{(\text{RS})}(q; \lambda, \rho) &\equiv \frac{\lambda}{4}(q-\rho)^2 + I(X^*; \sqrt{\lambda q}\, X^* + Z) \\ &= \frac{\lambda}{4}(q^2+\rho^2) - \mathbb{E}\ln \int dP_X(x) \exp\Big\{\lambda q x X^* + \sqrt{\lambda q} Z x - \frac{\lambda q}{2} x^2\Big\}. \end{aligned}$$

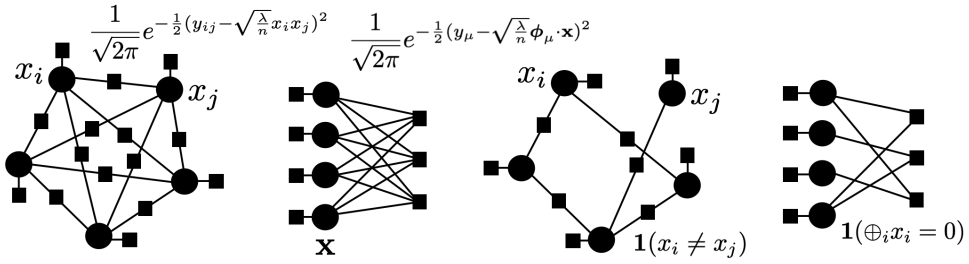


Figure 6: Graphical models/factor graphs associated with different inference and combinatorial optimization problems of a mean-field nature (the two first are dense models, the two last are sparse/tree-like). Starting from left: the factor graph of the spiked Wigner model (the factors connected to a single node represent the independent prior contribution $P_X(x_i)$ encoding, e.g., the domain of the variable etc), of high-dimensional linear regression, of the coloring problem, of a low-density parity-check code.

Decoupling: mean-field interpretation. Looking at this formula something peculiar appears: *the computation of the mutual information of the high-dimensional model* (21) *has been reduced to a simple scalar optimization problem.* In the replica potential appears the mutual information of a much simpler inference problem, namely, denoising under Gaussian noise:

$$y = \sqrt{\lambda q}\, x^* + z.$$

Lets us look at the stationary condition of the replica symmetric potential. For any finite λ it is not difficult to show that the infimum is attained away from the boundaries, so we compute, based on the I-MMSE formula:

$$\frac{di^{(\mathrm{RS})}(q)}{dq} = \frac{\lambda}{2}(q-\rho) + \frac{\lambda}{2}\mathrm{MMSE}(X^*|\sqrt{\lambda q}\, X^* + Z).$$

Cancelling this derivative we obtain the stationary condition:

$$q = \rho - \mathrm{MMSE}(X^*|\sqrt{\lambda q}\, X^* + Z) \overset{\mathrm{N}}{=} \mathbb{E}[X^*\langle X\rangle_q]. \tag{30}$$

We used the consequence (6) of the Nishimori identity. The bracket $\langle -\rangle_q$ is the expectation w.r.t. the measure

$$\frac{P_X(x)\exp\left\{\lambda q x X^* + \sqrt{\lambda q}Zx - \frac{\lambda q}{2}x^2\right\}}{\int dP_X(x')\exp\left\{\lambda q x' X^* + \sqrt{\lambda q}Zx' - \frac{\lambda q}{2}(x')^2\right\}}.$$

Among the solutions of this self-consistency equation the one minimizing the replica symmetric potential, denoted $q_0 = q_0(\lambda, \rho)$, plays a special role: it is linked to the MMSE as we will see next. So we see that the best error reachable for high-dimensional PCA collapses to the analysis of the MMSE of a scalar/decoupled inference problem.

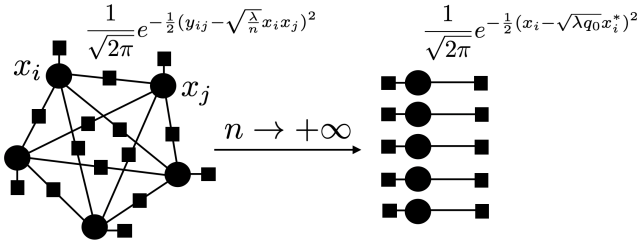


Figure 7: Visualization of the decoupling phenomenon happening in the thermodynamic limit: the left model is the original one, the right one is an asymptotically "equivalent" mean-field model with effective Gaussian factors, with SNR related to the minimizer $q_0(\lambda, \rho)$ of the replica symmetric potential.

This reduction from a high-dimensional model to a low-dimensonal one is typical of mean-field models. Mean-field models, for which such dramatic reduction in complexity is possible, usually belong to one of two possible classes: fully connected models, defined by complete graphical models, and sparse tree-like graphical models, see Fig. 6.

Minimum mean-square error. As a consequence of the replica symmetric formula for the mutual information we obtain by application of the I-MMSE identity a simple formula for the (matrix/spike) MMSE:

Corollary 1 (Minimum mean-square error)**.** *Under the same assumptions as in Theorem 1 and for all (λ, ρ) such that the minimizer of the replica symmetric potential is unique, in which case we denote this unique minimizer*

$$q_0(\lambda, \rho) \equiv \operatorname*{argmin}_{q \in [0,\rho]} i^{(\mathrm{RS})}(q; \lambda, \rho),$$

the spike-MMSE (or matrix-MMSE) verifies:

$$\lim_{n \to +\infty} \frac{1}{n^2} \mathbb{E} \|\boldsymbol{X}^* \otimes \boldsymbol{X}^* - \langle \boldsymbol{X} \otimes \boldsymbol{X} \rangle\|_{\mathrm{F}}^2 = \rho^2 - q_0(\lambda, \rho)^2.$$

Here $\langle \boldsymbol{X} \otimes \boldsymbol{X} \rangle = \mathbb{E}[\boldsymbol{X} \otimes \boldsymbol{X} | \boldsymbol{y}]$ is the spike posterior mean (the MMSE estimator).

Proof. For a sequence of concave functions with a pointwise limit, the limit of the derivative is the derivative of the limit. Let $I \subset \mathbb{R}$ and $(g_n)_{n \geq 0}$ be a sequence of concave functions on I that converge pointwise to g. Then for any $x \in I$ where $g'(x)$ exists we have $\lim_{n \to +\infty} g_n'(x) = g'(x)$. Indeed by concavity

$$\lim_{n \to +\infty} g_n'(x) \geq \lim_{n \to +\infty} \frac{g_n(x + \epsilon) - g_n(x)}{\epsilon} = \frac{g(x + \epsilon) - g(x)}{\epsilon} \xrightarrow[\epsilon \to 0_+]{} g'(x_+).$$

Similarly we obtain $\lim_{n \to +\infty} g_n'(x) \leq g'(x_-)$. If g is differentiable at x then $g'(x_+) = g'(x_-) = g'(x)$, thus the result.

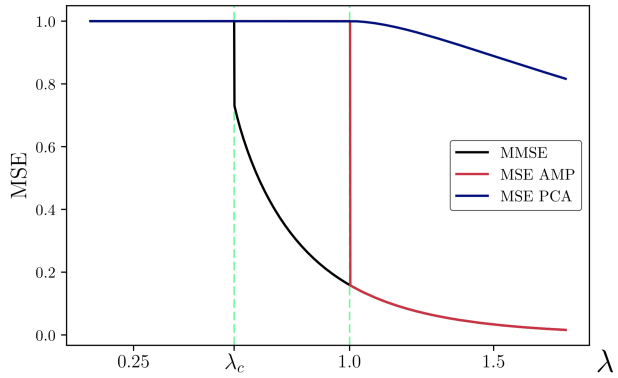


Figure 8: Figure from [7]. Plot of the minimum mean-square error given by Corollary 1, the MSE of the approximate message-passing algorithm and of naive PCA for the spiked Wigner model with prior given by (22).

The sequence of functions we have in mind is the mutual information density $\frac{1}{n}I(\boldsymbol{X}^*;\boldsymbol{Y})$, which are concave in λ. Indeed by the I-MMSE formula

$$\begin{aligned}\frac{d}{d\lambda}\frac{1}{n}I(\boldsymbol{X}^*;\boldsymbol{Y}) &= \frac{1}{2n^2}\sum_{i<j}\mathbb{E}\big[(X_i^*X_j^* - \langle X_iX_j\rangle)^2\big] \\ &= \frac{1}{4n^2}\mathbb{E}\|\boldsymbol{X}^*\otimes\boldsymbol{X}^* - \langle \boldsymbol{X}\otimes\boldsymbol{X}\rangle\|_{\mathrm{F}}^2 + O(1/n). \end{aligned} \tag{31}$$

The MMSE is clearly a non-increasing function of the signal-to-noise ratio λ ("information can't hurt"), which can also be verified by direct computation. Consider (λ,ρ) so that the minimizer of the potential is unique: $q_0 = q_0(\lambda,\rho)$. By Theorem 1 the mutual information converges to $i^{(\mathrm{RS})}(q_0;\lambda,\rho)$. It is not difficult to show by direct computation that $q_0(\lambda,\rho)\in(0,\rho)$. Therefore it must be that the q-derivative of $i^{(\mathrm{RS})}$ cancels at $q_0 = q_0(\lambda,\rho)$ and thus, using the I-MMSE relation,

$$\begin{aligned}\lim_{n\to+\infty}\frac{d}{d\lambda}\frac{1}{n}I(\boldsymbol{X}^*;\boldsymbol{Y}) &= \frac{d}{d\lambda}i^{(\mathrm{RS})}(q_0;\lambda,\rho) \\ &= \frac{\partial}{\partial\lambda}i^{(\mathrm{RS})}(q_0;\lambda,\rho) \\ &= \frac{(\rho-q_0)^2}{4} + \frac{q_0}{2}\mathrm{MMSE}(X^*|\sqrt{\lambda q_0}X^*+Z) \\ &= \frac{(\rho-q_0)^2}{4} + \frac{q_0}{2}(\rho-q_0) \\ &= \frac{\rho^2-q_0^2}{4}\end{aligned}$$

using that $q_0(\lambda,\rho)$ verifies the stationary condition (30). Comparing this result with (31) in the thermodynamic limit $n\to+\infty$ ends the proof. □

Phase diagram and MMSE. Now that we have a rigorous tool to locate the

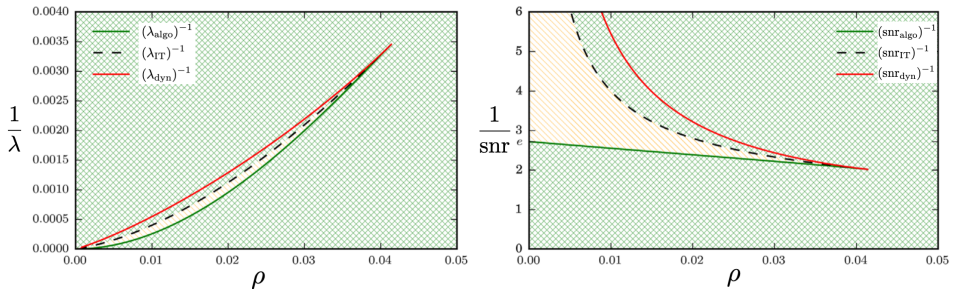


Figure 9: From [8]. Phase diagram of the spiked Wigner model with Bernoulli parameters $X_i \sim \mathrm{Ber}(\rho)$ as a function of the sparsity ρ and inverse of the SNR λ (left) or the inverse of the total SNR given by $\mathrm{snr} \equiv \lambda\rho^2$ (right), see (23). There is no phase transition in the system if $\rho > 0.04139$ and a first order phase transition else. The lower green curve is the algorithmic phase transition of the approximate message-passing algorithm, see section 3, that converges to $(\lambda_{\rm algo})^{-1} = e\rho^2$. The dashed black line is the information theoretic threshold. The upper red curve is the dynamical spinodal where the informative fixed point of state evolution disappears. The orange hashed zone is the hard region in which AMP is sub-optimal (as any known sub-exponential complexity algorithm). In the rest of the phase diagram (green hashed) the AMP provides in the large size limit the Bayesian optimal error.

phase transitions/information theoretic limits, let us plot the phase diagram of PCA.

2.2 A powerful (exact) heuristic: the replica method

2.2.1 "Single-letter" derivation

The replica method has been developed for the study of the thermodynamic properties of disordered statistical mechanical models such as spin glasses. It is based on one of the following equivalent identities, coined *replica trick*:

$$\mathbb{E}\ln \mathcal{Z}(\boldsymbol{Y}) = \lim_{k\to 0_+} \frac{\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]-1}{k} = \lim_{k\to 0_+} \frac{\partial}{\partial k} \ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] = \lim_{k\to 0_+} \frac{\ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{k}$$

where the *replicated partition function* is

$$\mathcal{Z}(\boldsymbol{y})^k \equiv \int dP(\{\boldsymbol{x}_1^k\}) \exp\Big\{ -\sum_{a=1}^{k} \mathcal{H}(\boldsymbol{x}^a;\boldsymbol{y})\Big\}$$

using the notation $\int dP(\{\boldsymbol{x}_1^k\})\cdots = \int_{\mathbb{R}^{nk}} \prod_{a=1}^k \prod_{i=1}^n dP_X(x_i^a)\cdots$, and where the Hamiltonian is (26). This is the partition function of k replicas $\{\mathbb{R}^n \ni \boldsymbol{x}^a \equiv \boldsymbol{x}^{(a)} : a = 1,\dots,k\}$ drawn i.i.d. from the posterior (25) which is $\propto \exp\{-\mathcal{H}(\cdot\,;\boldsymbol{y})\}$.

Choosing your favorite form of the replica trick, the asymptotic free energy (related to the mutual information by an additive constant) reads

$$\lim_{n\to+\infty} f_n = -\lim_{n\to+\infty} \frac{1}{n}\mathbb{E}\ln \mathcal{Z}(\boldsymbol{Y}) = -\lim_{n\to+\infty}\lim_{k\to 0_+} \frac{\ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk}. \tag{32}$$

We therefore compute the moments of the partition function $\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]$ as if $k \in \mathbb{N}$ and then we'll do an analytic continuation to the reals with $k \to 0_+$ and hope for the best.

We compute

$$\begin{aligned}&\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]\\ &\quad= \mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}\int dP(\{\boldsymbol{x}_1^k\})\exp\sum_{i<j}\Big(Y_{ij}\sqrt{\frac{\lambda}{n}}\sum_{a=1}^{k}x_i^a x_j^a - \frac{\lambda}{2n}\sum_{a=1}^{k}(x_i^a x_j^a)^2\Big).\end{aligned}$$

We now integrate the quenched Gaussian variables

$$Y_{ij} \sim \mathcal{N}\Big(\sqrt{\frac{\lambda}{n}}X_i^* X_j^*, 1\Big).$$

Using the Gaussian integration formula (also called Hubbard–Stratonovich formula when seen from right to left)

$$\int_{\mathbb{R}} dz\,\exp\{-az^2 + bz\} = \sqrt{\frac{\pi}{a}}\exp\frac{b^2}{4a}$$

and, letting the sum $\sum_{a\neq b}^{u,k}$ be over $\{a, b \in \{u, u+1, \dots, k\} : a \neq b\}$, we obtain

$$\begin{aligned}&\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]\\ &\quad= \mathbb{E}_{\boldsymbol{X}^*}\int dP(\{\boldsymbol{x}_1^k\})\exp\sum_{i<j}\Big\{\frac{\lambda}{n}\sum_{a=1}^{k}x_i^a X_i^* x_j^a X_j^* + \frac{\lambda}{2n}\sum_{a\neq b}^{1,k}x_i^a x_i^b x_j^a x_j^b\Big\}.\end{aligned}$$

Because the prior matches the ground-truth distribution $\mathbb{E}_{\boldsymbol{X}^*}\cdots = \int dP(\boldsymbol{x}^*)\cdots$. Then, denoting $\boldsymbol{x}^* = \boldsymbol{x}^0$ and $\int dP(\{\boldsymbol{x}_0^k\})\cdots = \int_{\mathbb{R}^{n(k+1)}}\prod_{a=0}^{k}\prod_{i=1}^{n}dP_X(x_i^a)\cdots$, the replicated partition function can be re-expressed as

$$\begin{aligned}\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] &= \int dP(\{\boldsymbol{x}_0^k\})\exp\frac{\lambda}{2n}\sum_{a\neq b}^{0,k}\sum_{i<j}x_i^a x_i^b x_j^a x_j^b\\ &= \int dP(\{\boldsymbol{x}_0^k\})\exp\frac{\lambda}{4n}\sum_{a\neq b}^{0,k}\Big\{\Big(\sum_{i=1}^{n}x_i^a x_i^b\Big)^2 - \sum_{i=1}^{n}(x_i^a x_i^b)^2\Big\}.\end{aligned} \tag{33}$$

We observe here a direct consequence of the fact that we are considering the problem in the Bayesian optimal setting (i.e., the posterior is known): the ground-truth

plays a perfectly symmetric role as a replica (i.e., a sample from the posterior), thus the notation $\boldsymbol{x}^* = \boldsymbol{x}^0$. Averaging the (random) quenched disorder $\boldsymbol{Y}$ therefore converted a disordered system made of k independent copies/replicas, each dependent on $\boldsymbol{Y}$, into a non-disordered system but where the replicas are now coupled.

From there we use the Hubbard–Stratonovich formula (with $a = n\lambda/4$, $b = (\lambda/2)\sum_i x_i^a x_i^b$) in order to decouple the spins in the "real space" (i.e., we linearize the sums overs i) by introducing coupling Gaussian fields:

$$\exp \frac{\lambda}{4n} \sum_{a\neq b}^{0,k} \Big(\sum_{i=1}^{n} x_i^a x_i^b \Big)^2 = \Big(\frac{n\lambda}{4\pi}\Big)^{\frac{k(k+1)}{2}} \int_{\mathbb{R}^{k(k+1)}} d\boldsymbol{q} \exp \sum_{a\neq b}^{0,k} \Big\{ -\frac{n\lambda q_{ab}^2}{4} + \frac{\lambda q_{ab}}{2} \sum_{i=1}^{n} x_i^a x_i^b \Big\}.$$

Therefore the replicated partition function equals

$$\begin{aligned}\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] &= \Big(\frac{n\lambda}{4\pi}\Big)^{\frac{k(k+1)}{2}} \int d\boldsymbol{q} \exp\Big\{ -\frac{n\lambda}{4} \sum_{a\neq b}^{0,k} q_{ab}^2 \Big\} \\ &\qquad \times \underbrace{\int dP(\{\boldsymbol{x}_0^k\}) \prod_{i=1}^{n} \exp \sum_{a\neq b}^{0,k} \Big\{ \frac{\lambda q_{ab}}{2} x_i^a x_i^b - \frac{\lambda}{4n} (x_i^a x_i^b)^2 \Big\}}_{\left(\int_{\mathbb{R}^{k+1}} dP(\boldsymbol{x}) \exp \sum_{a\neq b}^{0,k} \{ \frac{\lambda q_{ab}}{2} x^a x^b - \Theta(1/n) \} \right)^n} \\ &= \int d\boldsymbol{q} \exp\big\{ -n\mathcal{S}_n(\boldsymbol{q}) \big\},\end{aligned}$$

where the $\Theta(1/n) = \frac{\lambda}{4n}(x_i^a x_i^b)^2$ (recall the prior has finite support so $\frac{\lambda}{4n}(x_i^a x_i^b)^2$ vanishes in the large n limit), and the effective "action" is

$$\begin{aligned}\mathcal{S}_n(\boldsymbol{q}) \equiv &-\frac{k(k+1)}{2n} \ln \frac{n\lambda}{4\pi} + \frac{\lambda}{4} \sum_{a\neq b}^{0,k} q_{ab}^2 \\ &- \ln \int_{\mathbb{R}^{k+1}} dP(\boldsymbol{x}) \exp \sum_{a\neq b}^{0,k} \Big\{ \frac{\lambda q_{ab}}{2} x^a x^b - \Theta(1/n) \Big\}.\end{aligned}$$

We *assume* that the large n and small k limits in (32) commute:

$$-\lim_{n\to+\infty} \lim_{k\to 0_+} \frac{\ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk} = -\lim_{k\to 0_+} \lim_{n\to+\infty} \frac{\ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk}.$$

Then at fixed k (which is still interpreted as an integer at the moment), the large n limit is evaluated by the Laplace/saddle point method (recall the prior is bounded

so the term $\frac{\lambda}{4n}(x^a x^b)^2$ in the action $\mathcal{S}_n(\boldsymbol{q})$ can simply be set to zero)[16]:

$$-\lim_{n\to+\infty}\frac{\ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk} = \min_{\boldsymbol{q}}\frac{\mathcal{S}(\boldsymbol{q})}{k}$$

where $\mathcal{S}(\boldsymbol{q}) = \lim_{n\to+\infty}\mathcal{S}_n(\boldsymbol{q})$. We see that the action is invariant under permutations of lines and columns of the matrix (q_{ab}). This suggests a natural ansatz: the *replica symmetric ansatz* assumes that the saddle point lies on the subset[17]

$$q_{ab} = q \quad \text{for all} \quad a \neq b.$$

The replica symmetric ansatz therefore simplifies the action to

$$\frac{\mathcal{S}(q)}{k} \equiv (k+1)\frac{\lambda q^2}{4} - \frac{1}{k}\ln\int_{\mathbb{R}^{k+1}} dP(\boldsymbol{x})\exp\frac{\lambda q}{2}\Big\{\Big(\sum_{a=0}^{k}x^a\Big)^2 - \sum_{a=0}^{k}(x^a)^2\Big\}.$$

Using once more the Hubbard–Stratonovich formula (with $a = 1/2$ and $b = \sqrt{\lambda q}\sum_{a=0}^{k}x^a$) in order to decouple the spins but this time in the "replica space" (i.e., linearizing the a indices), and letting $Z \sim \mathcal{N}(0,1)$, we get

$$\frac{\mathcal{S}(q)}{k} \equiv (k+1)\frac{\lambda q^2}{4} - \frac{1}{k}\ln\mathbb{E}\int\prod_{a=0}^{k}dP_X(x^a)\exp\Big\{\sqrt{\lambda q}Zx^a - \frac{\lambda q}{2}(x^a)^2\Big\}. \quad (34)$$

The replicas are now decoupled. Let $X^0 = X^* \sim P_X$. Letting $k \to 0_+$,

$$\begin{aligned}
&\frac{1}{k}\ln\mathbb{E}\int\prod_{a=0}^{k}dP_X(x^a)\exp\Big\{\sqrt{\lambda q}Zx^a - \frac{\lambda q}{2}(x^a)^2\Big\}\\
&\quad= \frac{1}{k}\ln\int\frac{dz}{\sqrt{2\pi}}dP_X(x^0)e^{-\frac{z^2}{2}+\sqrt{\lambda q}zx^0-\frac{\lambda q}{2}(x^0)^2}\Big(\int dP_X(x)e^{\sqrt{\lambda q}zx-\frac{\lambda q}{2}x^2}\Big)^k\\
&\quad= \frac{1}{k}\ln\int\frac{dz}{\sqrt{2\pi}}dP_X(x^*)e^{-\frac{1}{2}(z-\sqrt{\lambda q}x^*)^2}\Big(\int dP_X(x)e^{\sqrt{\lambda q}zx-\frac{\lambda q}{2}x^2}\Big)^k\\
&\quad= \frac{1}{k}\ln\mathbb{E}\Big[\Big(\int dP_X(x)e^{\sqrt{\lambda q}Zx+\lambda q x X^*-\frac{\lambda q}{2}x^2}\Big)^k\Big]\\
&\quad= \mathbb{E}\ln\int dP_X(x)\exp\Big\{\sqrt{\lambda q}Zx + \lambda q x X^* - \frac{\lambda q}{2}x^2\Big\} + O(k). \qquad (35)
\end{aligned}$$

[16] In a non-planted problem such as the Sherrington-Kirkpatrick model, the $\min_{\boldsymbol{q}}\mathcal{S}(\boldsymbol{q})$ would be replaced by a $\mathrm{extr}_{\boldsymbol{q}}\mathcal{S}(\boldsymbol{q})$. This is because as $k \to 0_+$ the minima become maxima. But due to the precense of the planted signal, which is equivalent to a 0-th replica by the Nishimori identity, this does not happen. This is easily seen: e.g., in the SK model the term corresponding to the one multiplied by $k+1$ in expression (34) would instead by multiplied by $k-1$. Therefore the sign would be reversed as $k \to 0_+$ which is connected to the swith from minima to maxima.

[17] More complicated ansatz, associated to a *spontaneous breaking* of the overlap matrix permutation symmetry, are called *replica symmetry breaking ansatz*, and were developed by Giorgio Parisi for the study of the Sherrington-Kirkpatrick model.

We used the change of variable $z \to z - \sqrt{\lambda q} x^*$ from third to forth lines, and

$$\frac{1}{k} \ln \mathbb{E}[U^k] = \frac{1}{k} \ln \mathbb{E} \exp(k \ln U) = \frac{1}{k} \ln(1 + k \mathbb{E} \ln U + O(k^2)) = \mathbb{E} \ln U + O(k).$$

The simplified action in the limit $\lim_{k \to 0_+} \min_q \frac{1}{k} \mathcal{S}(q)$, and therefore the conjectured asymptotic free energy, reads

$$\lim_{n \to +\infty} f_n = \min_q \left[\frac{\lambda q^2}{4} - \mathbb{E} \ln \int dP_X(x) \exp\left\{ \sqrt{\lambda q} Z x + \lambda q x X^* - \frac{\lambda q}{2} x^2 \right\} \right]. \quad (36)$$

Adding the missing term $\rho^2 \lambda / 4$, as seen from the link (27) between free energy and mutual information, we (heuristically) recover the expression of Theorem 1 for the mutual information.

2.2.2 "Two-letters" derivation

There exists an alternative derivation by the replica method leading to a more generic (yet equivalent) variational formula in terms of a two-letters potential.

The derivation is the same until (33). This identity shows that the replicated partition function depends on the "microscopic configurations" $\{\boldsymbol{x}^a\}$ only through the "macroscopic" overlap order parameter $Q_{ab} = Q(\boldsymbol{x}^a, \boldsymbol{x}^b) \equiv \frac{1}{n} \boldsymbol{x}^a \cdot \boldsymbol{x}^b$:

$$\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] = \int dP(\{\boldsymbol{x}_0^k\}) \exp \frac{\lambda n}{4} \sum_{a \neq b}^{0,k} Q(\boldsymbol{x}^a, \boldsymbol{x}^b)^2.$$

We dropped the sub-dominant $\frac{\lambda}{4n} \sum_{a \neq b}^{0,k} \sum_{i=1}^{n} (x_i^a x_i^b)^2$ from (33) that plays no role in the large n limit. The expression above suggests to introduce the multiplicity $\exp \Gamma(\boldsymbol{q})$ of the configurations $\{\boldsymbol{x}^a\}$ which have a given overlap matrix $(Q_{ab}) = (q_{ab})$ through the identity:

$$\exp \Gamma(\boldsymbol{q}) \equiv \int dP(\{\boldsymbol{x}_0^k\}) \prod_{a \neq b}^{0,k} \delta\big(q_{ab} - Q(\boldsymbol{x}^a, \boldsymbol{x}^b)\big).$$

So $\Gamma(\boldsymbol{q})$ is interpreted as the entropy associated with the configurations with overlap $\boldsymbol{q}$. We therefore have the identity

$$\begin{aligned} 1 &= \int d\boldsymbol{q} \exp \Gamma(\boldsymbol{q}) \qquad (37) \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^{nk(k+1)}} dP(\{\boldsymbol{x}_0^k\}) \int_{\mathbb{R}^{k(k+1)}} d\boldsymbol{q} \int_{\mathrm{i}\mathbb{R}^{k(k+1)}} d\hat{\boldsymbol{q}} \exp \sum_{a \neq b}^{0,k} \hat{q}_{ab} \big(q_{ab} - Q(\boldsymbol{x}^a, \boldsymbol{x}^b)\big). \end{aligned}$$

Note that the integral over the $\hat{q}_{ab}$'s are on the imaginary axis. We used

$$1 = \frac{1}{2\pi} \int_{\mathbb{R}} dx \int_{\mathrm{i}\mathbb{R}} d\hat{x} \exp\big\{\hat{x}(x-a)\big\} = \int dx \delta(x).$$

This identity is at the origin of the formal Fourier represention of the dirac delta function: $\delta(x) = \frac{1}{2\pi}\int_{\mathrm{i}\mathbb{R}} d\hat{x}\exp(\hat{x}x)$. Note that this can only be used inside an integral over x otherwise the integral $\int_{\mathrm{i}\mathbb{R}} d\hat{x}\exp(\hat{x}x)$ is not convergent[18]. We plug (37) in the replicated partition function:

$$\begin{aligned}
&\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] \\
&= \Big(\frac{1}{2\pi}\Big)^{k(k+1)}\int dP(\{\boldsymbol{x}_0^k\})d\boldsymbol{q}d\hat{\boldsymbol{q}}\exp\sum_{a\neq b}^{0,k}\Big\{\hat{q}_{ab}\Big(\sum_{i=1}^{n}x_i^a x_i^b - nq_{ab}\Big) + \frac{\lambda n q_{ab}^2}{4}\Big\} \\
&= \Big(\frac{1}{2\pi}\Big)^{k(k+1)}\int d\boldsymbol{q}d\hat{\boldsymbol{q}}\Big(\exp\sum_{a\neq b}^{0,k}\Big\{\frac{\lambda q_{ab}^2}{4} - \hat{q}_{ab}q_{ab}\Big\}\int_{\mathbb{R}^{k+1}} dP(\boldsymbol{x})\exp\sum_{a\neq b}^{0,k}\hat{q}_{ab}x^a x^b\Big)^n \\
&= \int d\boldsymbol{q}d\hat{\boldsymbol{q}}\exp\big\{-n\mathcal{S}_n(\boldsymbol{q},\hat{\boldsymbol{q}})\big\} \qquad (38)
\end{aligned}$$

with effective action

$$\mathcal{S}_n(\boldsymbol{q},\hat{\boldsymbol{q}}) \equiv \frac{k(k+1)}{n}\ln 2\pi - \sum_{a\neq b}^{0,k}\Big\{\frac{\lambda q_{ab}^2}{4} - \hat{q}_{ab}q_{ab}\Big\} - \ln\int dP(\boldsymbol{x})\exp\sum_{a\neq b}^{0,k}\hat{q}_{ab}x^a x^b.$$

Note that the decoupling in the real space (over the indices i, j) took place automatically as a consequence of the fixed overlap constraint. As $n \to +\infty$ we evaluate this integral by saddle point assuming a replica symmetric ansatz for the solution:

$$q_{ab} = q \quad \text{and} \quad \hat{q}_{ab} = \frac{\hat{q}}{2} \quad \text{for all} \quad a \neq b.$$

Then the saddle point evaluation gives

$$-\lim_{n\to+\infty}\frac{\ln\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk} = \operatorname*{extr}_{q,\hat{q}}\frac{\mathcal{S}(q,\hat{q})}{k} \qquad (39)$$

where the replica symmetric action is

$$\frac{\mathcal{S}(q,\hat{q})}{k} = (k+1)\Big(\frac{q\hat{q}}{2} - \frac{\lambda q^2}{4}\Big) - \frac{1}{k}\ln\int dP(\boldsymbol{x})\exp\frac{\hat{q}}{2}\Big\{\Big(\sum_{a=0}^{k}x^a\Big)^2 - \sum_{a=0}^{k}(x^a)^2\Big\}.$$

As the integral we estimate by saddle point lives in the complex plane, we need to extremize (not minimize). The extremization (39) is therefore over $q, \hat{q} \in \mathbb{C}$: the effective action is holomorphic so the complex integral (38) is deformed, without changing the integral value, in order to pass through the extremum saddle point.

[18] To cure this issue one can regularize this integral using instead the Poisson kernel $\eta_\epsilon(x)$ in the limit $\epsilon \to 0_+$, where $\eta_\epsilon(x) \equiv \epsilon/(\pi(\epsilon^2 + x^2)) = \frac{1}{2\pi}\int_{\mathrm{i}\mathbb{R}} d\hat{x}\exp(\hat{x}x - |\epsilon\hat{x}|)$

Let $\mathcal{S}(q,\hat{q}) = \lim_{n\to+\infty} \mathcal{S}_n(q,\hat{q})$. In order to decouple the variables in the replica space (indices a, b) we use the Hubbard–Stratonovich transform with $a = 1/2$ and $b = \sqrt{\hat{q}}\sum_{a=0}^{k} x^a$. With $Z \sim \mathcal{N}(0,1)$ it yields

$$\begin{aligned}\frac{\mathcal{S}(q,\hat{q})}{k} &= (k+1)\Big(\frac{q\hat{q}}{2} - \frac{\lambda q^2}{4}\Big) - \frac{1}{k}\ln\int dP(\boldsymbol{x}) \exp\frac{\hat{q}}{2}\Big\{\Big(\sum_{a=0}^{k} x^a\Big)^2 - \sum_{a=0}^{k}(x^a)^2\Big\} \\ &= (k+1)\Big(\frac{q\hat{q}}{2} - \frac{\lambda q^2}{4}\Big) - \frac{1}{k}\ln\mathbb{E}\int \prod_{a=0}^{k} dP(x^a) \exp\Big\{\sqrt{\hat{q}}Zx^a - \frac{\hat{q}}{2}(x^a)^2\Big\}.\end{aligned}$$

Letting $k \to 0_+$ we obtain, by similar manipulations to those leading to (35), that the free energy given by $\lim_{k\to 0_+}\min_{q,\hat{q}} \frac{1}{k}\mathcal{S}(q,\hat{q})$ is (here $X^* \sim P_X$)

$$\lim_{n\to+\infty} f_n = \operatorname*{extr}_{q,\hat{q}}\Big[\frac{q\hat{q}}{2} - \frac{\lambda q^2}{4} - \mathbb{E}\ln\int dP_X(x)\exp\Big\{\sqrt{\hat{q}}Zx + \hat{q}xX^* - \frac{\hat{q}}{2}x^2\Big\}\Big].$$

Let us denote by $f^{(\mathrm{RS})}(q,\hat{q})$ the two-letters replica symmetric potential, i.e., the function above inside the bracket $[\ldots]$ that is being extremized. The stationary conditions for this function are

$$\frac{d}{dq}f^{(\mathrm{RS})}(q,\hat{q}) = 0 \quad \Rightarrow \quad \hat{q} = \lambda q.$$

The second stationary condition gives

$$\begin{aligned}\frac{d}{d\hat{q}}f^{(\mathrm{RS})}(q,\hat{q}) = 0 \quad \Rightarrow \quad q &= 2\,\mathbb{E}\Big\langle \frac{ZX}{2}\frac{1}{\sqrt{\hat{q}}} + XX^* - \frac{X^2}{2}\Big\rangle_{\hat{q}} \\ &= 2\,\mathbb{E}\Big\langle \Big(\frac{X^2}{2} - \frac{XX'}{2}\Big) + XX^* - \frac{X^2}{2}\Big\rangle_{\hat{q}} \\ &\overset{\mathrm{N}}{=} \mathbb{E}\langle XX^*\rangle_{\hat{q}},\end{aligned}$$

where the bracket $\langle - \rangle_{\hat{q}}$ is w.r.t. the measure $\propto dP_X(x)\exp\{\sqrt{\hat{q}}Zx + \hat{q}xX^* - x^2\hat{q}/2\}$. Plugging the first stationnary condition in the two-letters potential, one recovers the single-letter potential that is minimized in (36).

2.3 Why ensembles matter? Concentration of the free energy

In this section we prove concentration of the free energy, which justifies an ensemble (average) analysis of large probabilisic models. Because the free energy (or equivalently the entropy and mutual information) concentrates, computing it for a single large instance of a problem is equivalent to computing it expectation over the problem ensemble. Therefore the locations of phase transitions and the values of the different error metrics become asymtpotically independent of the particular problem realization, because with high probability as $n \to +\infty$ this realization is *typical*, and the ensemble analysis provides the typical behavior.

Proposition 1 (Free energy concentration for the spiked Wigner model). *There exists C a positive constant that may depend on everything but n such that*

$$\mathbb{E}\Big[\Big(-\frac{1}{n}\ln \mathcal{Z}_n(\boldsymbol{Y}) - f_n\Big)^2\Big] \leq \frac{C}{n}\,.$$

The proof will be based on two classical concentration inequalities, namely:

Proposition 2 (Gaussian Poincaré inequality). *Let $\boldsymbol{U} = (U_1, \dots, U_N)$ be a vector of N independent standard normal random variables. Let $g : \mathbb{R}^N \to \mathbb{R}$ be a continuously differentiable function. Then*

$$\mathrm{Var}(g(\boldsymbol{U})) \leq \mathbb{E}\|\nabla g(\boldsymbol{U})\|_2^2\,.$$

Proposition 3 (Efron-Stein inequality). *Let $\mathcal{U} \subset \mathbb{R}$, and a function $g : \mathcal{U}^N \to \mathbb{R}$. Let $\boldsymbol{U} = (U_1, \dots, U_N)$ be a vector of N independent random variables with law P_U that take values in $\mathcal{U}$. Let $\boldsymbol{U}^{(i)}$ a vector which differs from $\boldsymbol{U}$ only by its i-th component, which is replaced by $\tilde{U}_i$ drawn from P_U independently of $\boldsymbol{U}$. Then*

$$\mathrm{Var}(g(\boldsymbol{U})) \leq \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{U}}\mathbb{E}_{\tilde{U}_i}\big[(g(\boldsymbol{U}) - g(\boldsymbol{U}^{(i)}))^2\big]\,.$$

We start by proving the concentration w.r.t. the Gaussian variables:

Lemma 1 (Concentration w.r.t. the Gaussian noise). *We have*

$$\mathbb{E}\Big[\Big(-\frac{1}{n}\ln \mathcal{Z}_n(\boldsymbol{X}^*, \boldsymbol{Z}) + \frac{1}{n}\mathbb{E}_{\boldsymbol{Z}}\ln \mathcal{Z}_n(\boldsymbol{X}^*, \boldsymbol{Z})\Big)^2\Big] \leq \frac{\lambda}{2n}.$$

Proof. The proof is based on Proposition 2. Fix all variables except $\boldsymbol{Z}$. Let $g(\boldsymbol{z}) \equiv -\frac{1}{n}\ln \mathcal{Z}_n(\boldsymbol{x}^*, \boldsymbol{z})$ be the free energy seen as a function of the Gaussian variables only. The free energy gradient w.r.t. these reads

$$\mathbb{E}\|\nabla g(\boldsymbol{Z})\|_2^2 = \mathbb{E}\sum_{i<j}^{n}\Big(\frac{\partial g}{\partial Z_{ij}}\Big)^2.$$

Let us simply denote $\mathcal{H} \equiv \mathcal{H}(\boldsymbol{x}; \boldsymbol{x}^*, \boldsymbol{z})$. We then compute (recall the support of the signal $[-1, 1]$)

$$\mathbb{E}\Big[\Big(\frac{\partial g}{\partial Z_{ij}}\Big)^2\Big] = \frac{1}{n^2}\mathbb{E}\Big[\Big\langle\frac{\partial \mathcal{H}}{\partial Z_{ij}}\Big\rangle^2\Big] = \frac{\lambda}{n^3}\mathbb{E}[\langle X_i X_j\rangle^2] \leq \frac{\lambda}{n^3}.$$

Therefore Proposition 2 directly implies the stated result. □

We now consider the fluctuations related to the signal:

Lemma 2 (Concentration w.r.t. the spike). *We have*

$$\mathbb{E}\Big[\Big(-\frac{1}{n}\mathbb{E}_{\boldsymbol{Z}}\ln \mathcal{Z}_n(\boldsymbol{X}^*, \boldsymbol{Z}) - f_n\Big)^2\Big] \leq \frac{2}{n}.$$

Proof. The proof uses this time Proposition 3. Let $g(\boldsymbol{X}^*) \equiv -\frac{1}{n}\mathbb{E}_{\boldsymbol{Z}} \ln \mathcal{Z}_n(\boldsymbol{X}^*, \boldsymbol{Z})$. Define $\boldsymbol{X}^{(*i)}$ as a vector with same entries as $\boldsymbol{X}^*$ except the i-th one that is replaced by $\tilde{X}_i^*$ drawn independently from P_X. Let us estimate $(g(\boldsymbol{X}) - g(\boldsymbol{X}^{(*i)}))^2$ by interpolation:

$$\begin{aligned}
\mathbb{E}\big[(g(\boldsymbol{X}^*) - g(\boldsymbol{X}^{(*i)}))^2\big] &= \mathbb{E}\Big[\Big(\int_0^1 ds \frac{d}{ds} g(s\boldsymbol{X}^* + (1-s)\boldsymbol{X}^{(*i)})\Big)^2\Big] \\
&= \frac{1}{n^2}\mathbb{E}\Big[\Big(\int_0^1 ds \Big\langle \frac{d}{ds}\mathcal{H}(s\boldsymbol{X}^* + (1-s)\boldsymbol{X}^{(*i)})\Big\rangle\Big)^2\Big] \\
&= \frac{1}{n^2}\mathbb{E}\Big[(X_i^* - \tilde{X}_i^*)^2\Big\langle \frac{1}{n} X_i \sum_{j(\neq i)} X_j^* X_j \Big\rangle^2\Big] \\
&\leq \frac{2}{n^2}.
\end{aligned}$$

Here $\mathcal{H}(s\boldsymbol{X}^* + (1-s)\boldsymbol{X}^{(*i)})$ is the Hamiltonian with $\boldsymbol{X}^*$ replaced by $s\boldsymbol{X}^* + (1-s)\boldsymbol{X}^{(*i)}$. Therefore Proposition 3 implies the claim. □

2.4 Replica symmetry in inference: overlap concentration

In this section with present techniques allowing to prove the concentration of the overlap in a general optimal Bayesian inference setting. This is called *replica symmetry* in physics and is a key result.

The trick is the usual one of statistical mechanics: add a *source term*. This mean that we add in the Hamiltonian a term which derivative will be connected to the object we want to control. E.g., if the Hamiltonian defining some spin system is $\mathcal{H}(\boldsymbol{\sigma})$ and you want to know what is the *magnetization* $m_n \equiv \frac{1}{n}\sum_{i=1}^n \sigma_i$ you can add a control parameter (an external field) conjugate to the magnetization:

$$\mathcal{H}(\boldsymbol{\sigma}) \to \mathcal{H}(\boldsymbol{\sigma}) + h\sum_{i=1}^n \sigma_i.$$

Then a derivative of the free energy w.r.t. the external field around $h = 0$ allows to compute the average magnetization:

$$\lim_{h\to 0} -\frac{d}{dh}\frac{1}{n}\ln \sum_{\boldsymbol{\sigma}} e^{-\mathcal{H}(\boldsymbol{\sigma}) - h\sum_{i=1}^n \sigma_i} = \lim_{h\to 0}\frac{d}{dh} f_n(h) = \lim_{h\to 0}\langle m_n\rangle_h.$$

In general, if there are phase transitions, the result may depend on how is taken the limit. E.g., in the Curie-Weiss model below the critical temperature the magnetization $\lim_{h\to 0}\langle m_n\rangle_h$ will be positive or negative depending wether the limit is taken from above or below zero. Second derivatives then provide information about the *fluctuations* of observables. Here,

$$\frac{1}{n}\frac{d^2}{dh^2} f_n(h) = \frac{1}{n}\frac{d}{dh}\langle m_n\rangle_h = -\big\langle (m - \langle m\rangle_h)^2\big\rangle_h.$$

So the free energy with a source term is a moment generating function of the conjugate quantity to the source: physically, changing a bit a control parameter, i.e., taking a derivative w.r.t. the external field, allows to probe the conjugate physical observable, here the average magnetization.

Consider the most generic inference channel, with data generated as

$$\boldsymbol{y} \sim P_{\rm out}(\cdot \,|\, \boldsymbol{x}^*).$$

The conditional distribution $P_{\rm out}$ is a totally general likelihood, and is assumed to be known in addition of the prior P used to generate the signal $\boldsymbol{x}^*$ (with bounded components in $[-1, 1]$), so that we are in the Bayesian optimal setting.

We want to gain information about the MMSE, or equivalently the overlap $Q \equiv \frac{1}{n}\boldsymbol{x}\cdot\boldsymbol{x}^*$. Naively one could just add a source term proportional to it. But the whole proof will strongly relies on the validity of the Nishimori identity; without it the overlap concentration is not expected to hold in general. Validity of the Nishimori identity is verified as long as the model we consider is an inference problem in the Bayesian optimal setting. *So the source we add must itseld come from an inference/observation channel.* One possibility is to consider having additional infinitesimal *side information* coming from a *decoupled Gaussian channel.* So the *perturbed model* is:

$$\begin{cases} \boldsymbol{y} & \sim P_{\rm out}(\cdot\,|\boldsymbol{x}^*), \\ \tilde{\boldsymbol{y}} & = \sqrt{\epsilon}\,\boldsymbol{x}^* + \tilde{\boldsymbol{z}}, \end{cases}$$

where $\tilde{\boldsymbol{z}}$ is an outcome of $\mathcal{N}(0, \mathrm{I}_n)$. The SNR $\epsilon \in [s_n, 2s_n]$ of the side channel is very small as we let the sequence $s_n \in (0, 1/2]$ vanish with n. Therefore in the large n limit we will recover the original model. The Hamiltonian for this perturbed model becomes

$$\mathcal{H}(\boldsymbol{x}; \boldsymbol{y}, \tilde{\boldsymbol{y}}) = -\ln P_{\rm out}(\boldsymbol{y}|\boldsymbol{x}) + \mathcal{H}_{\rm pert}(\boldsymbol{x}; \tilde{\boldsymbol{y}}),$$

where the pertubation Hamiltonian is

$$\begin{aligned} \mathcal{H}_{\rm pert}(\boldsymbol{x}; \tilde{\boldsymbol{y}}) &\equiv \frac{\epsilon}{2}\|\boldsymbol{x}\|^2 - \sqrt{\epsilon}\,\boldsymbol{x}\cdot\tilde{\boldsymbol{y}} \\ &= \frac{\epsilon}{2}\|\boldsymbol{x}\|^2 - \epsilon\,\boldsymbol{x}\cdot\boldsymbol{x}^* - \sqrt{\epsilon}\,\boldsymbol{x}\cdot\tilde{\boldsymbol{z}}. \end{aligned} \tag{40}$$

We replaced $\tilde{\boldsymbol{y}}$ by its expression. The posterior, partition function and Gibbs-bracket read

$$P(\boldsymbol{x}|\boldsymbol{y}, \tilde{\boldsymbol{y}}) = \frac{e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{y},\tilde{\boldsymbol{y}})}}{\mathcal{Z}(\boldsymbol{y}, \tilde{\boldsymbol{y}})}, \quad \text{with} \quad \mathcal{Z}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = \int dP(\boldsymbol{x}) e^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{y},\tilde{\boldsymbol{y}})},$$

$$\text{and} \quad \langle g(\boldsymbol{X})\rangle_\epsilon \equiv \int dP(\boldsymbol{x}|\boldsymbol{y}, \tilde{\boldsymbol{y}})\, g(\boldsymbol{x}).$$

Let

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{x}^*, \tilde{\boldsymbol{z}}) = \mathcal{L} \equiv \frac{1}{n}\frac{d}{d\epsilon}\mathcal{H}(\boldsymbol{x}; \boldsymbol{y}, \tilde{\boldsymbol{y}}) = \frac{1}{n}\Big(\frac{\|\boldsymbol{x}\|^2}{2} - \boldsymbol{x}\cdot\boldsymbol{x}^* - \frac{\boldsymbol{x}\cdot\tilde{\boldsymbol{z}}}{2\sqrt{\epsilon}}\Big)\,. \tag{41}$$

The overlap fluctuations are upper bounded by those of $\mathcal{L}$ as

$$\mathbb{E}\big\langle (Q - \mathbb{E}\langle Q\rangle_\epsilon)^2\big\rangle_\epsilon \leq 4\,\mathbb{E}\big\langle (\mathcal{L} - \mathbb{E}\langle \mathcal{L}\rangle_\epsilon)^2\big\rangle_\epsilon\,. \tag{42}$$

The point is that as $\mathcal{L}$ is linked to the free energy derivatives we will able to control it, and, because of the Nishimori identity (i.e., that we consider Bayesian optimla inference) we can relate this natural object $\mathcal{L}$ to the one of main interest: the overlap. A detailed derivation can be found in the appendices and involves only elementary algebra using the Nishimori identity and integrations by parts w.r.t. the Gaussian noise $\tilde{Z}_i$. Therefore this identity is totally generic in optimal Bayesian inference. The concentration of the overlap is then a direct consequence of the following result which is general. It applies also away from inference, optimal or not, as this result depends only on the form of the perturbation (40) and strucural properties of the free energy:

Proposition 4 (Total fluctuations of $\mathcal{L}$)**.** *Consider a statistical model described by an Hamiltonian perturbed by $\mathcal{H}_{\rm pert}(\boldsymbol{x}; \tilde{\boldsymbol{y}})$ given by (40) where $\epsilon \in [\underline{\epsilon}, \bar{\epsilon}] \subseteq [s_n, \bar{\epsilon}]$ a bounded interval and where s_n is a positive vanishing sequence, and that $\boldsymbol{x}$ is bounded. There exist constants $C_i > 0$ independent of n s.t.*

$$\int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon\, \mathbb{E}\big\langle (\mathcal{L} - \mathbb{E}\langle \mathcal{L}\rangle_\epsilon)^2\big\rangle_\epsilon \leq C_1\Big(\frac{\bar{\epsilon} - \underline{\epsilon}}{n s_n^2}\Big)^{1/3} + C_2\frac{\ln(\bar{\epsilon}/\underline{\epsilon})}{n} + \frac{C_3}{n}.$$

In particular, choosing $[\underline{\epsilon}, \bar{\epsilon}] = [s_n, 2s_n]$ this implies that there exists a constant $C > 0$ s.t.

$$\int_{s_n}^{2s_n} d\epsilon\, \mathbb{E}\big\langle (\mathcal{L} - \mathbb{E}\langle \mathcal{L}\rangle_\epsilon)^2\big\rangle_\epsilon \leq \frac{C}{(n s_n)^{1/3}}.$$

Note that this results imposes that s_n goes to zero as $n^{-\alpha}$ with $0 < \alpha < 1$. This has a physical meaning: if the perturbation/side-information is too weak the system does not "feel it" so that the perturbation cannot force the overlap to be self-averaging. The average over a small window of ϵ is *not* an artefact of the proof. Indeed, there might be a (zero-measure) of λ values (and other parameters of the problem such as ρ etc) where, in the $n \to +\infty$ limit, there are *phase transitions*. These precisely manifest by non self-averaging of the physical observables such as the overlap. But averaging over a vanishing window of ϵ, which importantly is independent of the other control parameters like λ, allows to "smoothen" the overlap fluctuations, effectively cancelling the dramatic effect of possible phase transitions.

The proof of this proposition is broken in two parts, using again the decomposition

$$\mathbb{E}\big\langle (\mathcal{L} - \mathbb{E}\langle \mathcal{L}\rangle_\epsilon)^2\big\rangle_\epsilon = \mathbb{E}\big\langle (\mathcal{L} - \langle \mathcal{L}\rangle_\epsilon)^2\big\rangle_\epsilon + \mathbb{E}\big[(\langle \mathcal{L}\rangle_\epsilon - \mathbb{E}\langle \mathcal{L}\rangle_\epsilon)^2\big]\,.$$

Thus it suffices to prove the two following lemmas. The first lemma expresses concentration w.r.t. the posterior distribution (or "thermal fluctuations") and is a direct consequence of concavity properties of the average free energy and the Nishimori identity. The second fluctuations are w.r.t. the quenched disorder.

Let the free energy of this generic pertubed model be, as usual, defined as minus the log-partiton function:

$$F_{n,\epsilon}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \equiv -\frac{1}{n} \ln \mathcal{Z}(\boldsymbol{y}, \tilde{\boldsymbol{y}}), \qquad f_{n,\epsilon} \equiv \mathbb{E}\, F_{n,\epsilon}(\boldsymbol{Y}, \tilde{\boldsymbol{Y}}).$$

We have the following identities: for any given realisation of the quenched variables

$$\frac{dF_{n,\epsilon}(\boldsymbol{y}, \tilde{\boldsymbol{y}})}{d\epsilon} = \langle \mathcal{L} \rangle_\epsilon \,, \tag{43}$$

$$\frac{1}{n} \frac{d^2 F_{n,\epsilon}(\boldsymbol{y}, \tilde{\boldsymbol{y}})}{d\epsilon^2} = -\Big\langle (\mathcal{L} - \langle \mathcal{L} \rangle_\epsilon)^2 \Big\rangle_\epsilon + \frac{1}{4n^2\epsilon^{3/2}} \langle \boldsymbol{X} \rangle_\epsilon \cdot \tilde{\boldsymbol{z}} \,. \tag{44}$$

Averaging (43) and (44), using a Gaussian integration by parts w.r.t. $\tilde{Z}_i$ and the Nishimori identity $\mathbb{E}\langle X_i X_i^* \rangle_\epsilon = \mathbb{E}[\langle X_i \rangle_\epsilon^2]$ we find

$$\frac{df_{n,\epsilon}}{d\epsilon} = \mathbb{E}\langle \mathcal{L} \rangle_\epsilon = -\frac{1}{2n} \mathbb{E} \|\langle \boldsymbol{X} \rangle_\epsilon\|^2 = -\frac{1}{2} \mathbb{E} \langle Q(\boldsymbol{X}, \boldsymbol{X}^*) \rangle_\epsilon \,, \tag{45}$$

$$\frac{1}{n} \frac{d^2 f_{n,\epsilon}}{d\epsilon^2} = -\mathbb{E}\Big\langle (\mathcal{L} - \langle \mathcal{L} \rangle_\epsilon)^2 \Big\rangle_\epsilon + \frac{1}{4n^2\epsilon} \mathbb{E}\Big\langle \|\boldsymbol{X} - \langle \boldsymbol{X} \rangle_\epsilon\|^2 \Big\rangle_\epsilon \,. \tag{46}$$

The first identity is similar to the I-MMSE formula, except that as we worked with an Hamiltonian where the x-independent terms have been simplified, it is the overlap instead of the MMSE that pops out when deriving the free energy w.r.t. the SNR.

Lemma 3 (Thermal fluctuations of $\mathcal{L}$)**.** *Consider a statistical model described by an Hamiltonian perturbed by $\mathcal{H}_{\rm pert}(\boldsymbol{x}; \tilde{\boldsymbol{y}})$ given by (40) where $\epsilon \in [\underline{\epsilon}, \bar{\epsilon}]$ a bounded interval, and that the prior distribution P has finite second moment ρ. We have*

$$\int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon\, \mathbb{E}\Big\langle (\mathcal{L} - \langle \mathcal{L} \rangle_\epsilon)^2 \Big\rangle_\epsilon \le \frac{\rho}{n} \Big(1 + \frac{\ln(\bar{\epsilon}/\underline{\epsilon})}{4} \Big) \,.$$

Proof. From (46)

$$\begin{aligned} \mathbb{E}\Big\langle (\mathcal{L} - \langle \mathcal{L} \rangle_\epsilon)^2 \Big\rangle_\epsilon &= -\frac{1}{n} \frac{d^2 f_{n,\epsilon}}{d\epsilon^2} + \frac{1}{4n^2\epsilon} \mathbb{E}\Big\langle \|\boldsymbol{X} - \langle \boldsymbol{X} \rangle_\epsilon\|^2 \Big\rangle_\epsilon \\ &\le -\frac{1}{n} \frac{d^2 f_{n,\epsilon}}{d\epsilon^2} + \frac{\rho}{4n\epsilon} \,, \end{aligned}$$

where we used $n\mathrm{MMSE} = \mathbb{E}\Big\langle \|\boldsymbol{X} - \langle \boldsymbol{X} \rangle_\epsilon\|^2 \Big\rangle_\epsilon \le \mathbb{E}\Big\langle \|\boldsymbol{X}\|^2 \Big\rangle_\epsilon \overset{\rm N}{=} \mathbb{E}\|\boldsymbol{X}^*\|^2 \equiv n\rho$. We integrate this inequality over ϵ:

$$\begin{aligned} \int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon\, \mathbb{E}\Big\langle (\mathcal{L} - \langle \mathcal{L} \rangle_\epsilon)^2 \Big\rangle_\epsilon &= -\frac{1}{n} \int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon\, \frac{d^2 f_{n,\epsilon}}{d\epsilon^2} + \frac{\rho}{4n} \int_{\underline{\epsilon}}^{\bar{\epsilon}} \frac{d\epsilon}{\epsilon} \\ &= \frac{1}{n} \Big(\frac{df_{n,\epsilon}}{d\epsilon}(\epsilon = \underline{\epsilon}) - \frac{df_{n,\epsilon}}{dR_\epsilon}(\epsilon = \bar{\epsilon}) \Big) + \frac{\rho}{4n} \ln(\bar{\epsilon}/\underline{\epsilon}) \,. \end{aligned}$$

From (45) we have $|df_{n,\epsilon}/d\epsilon| = |\mathbb{E}\langle Q\rangle_\epsilon/2| \leq \rho/2$ so the first term is certainly smaller in absolute value than ρ/n. This concludes the proof. □

The second lemma expresses the concentration w.r.t. the quenched disorder variables and is a consequence of the ϵ-concavity and concentration of the free energy onto its average (w.r.t. the quenched variables) This idea is to relate the quenched fluctuations of $\mathcal{L}$ to the difference of ϵ-derivatives of the free energy and its expectation. Now, because the free energy is (almost) concave in ϵ, and it concentrates, then its derivatives should concentrate too, which gives the result.

Lemma 4 (Quenched fluctuations of $\mathcal{L}$)**.** *Consider a statistical model described by an Hamiltonian perturbed by $\mathcal{H}_{\text{pert}}(\boldsymbol{x};\tilde{\boldsymbol{y}})$ given by (40) where $\epsilon \in [\underline{\epsilon},\bar{\epsilon}] \subseteq [s_n,\bar{\epsilon}]$ a bounded interval and where s_n is a positive vanishing sequence, and that $\boldsymbol{x}$ is bounded. There exist constants $C_i > 0$ independent of n s.t.*

$$\int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon\, \mathbb{E}\big[(\langle\mathcal{L}\rangle_\epsilon - \mathbb{E}\langle\mathcal{L}\rangle_\epsilon)^2\big] \leq C_1\Big(\frac{\bar{\epsilon}-\underline{\epsilon}}{n s_n^2}\Big)^{1/3} + C_2\frac{\ln(\bar{\epsilon}/\underline{\epsilon})}{n}.$$

Proof. Consider the following functions of ϵ:

$$\begin{aligned}\tilde{F}(\epsilon) &\equiv F_{n,\epsilon} + \frac{\sqrt{\epsilon}}{n}\sum_{i=1}^n |\tilde{z}_i|\,,\\ \tilde{f}(\epsilon) &\equiv \mathbb{E}\,\tilde{F}(\epsilon) = f_{n,\epsilon} + \frac{\sqrt{\epsilon}}{n}\sum_{i=1}^n \mathbb{E}\,|\tilde{Z}_i|\,.\end{aligned} \tag{47}$$

Because of (44) we see that the second derivative of $\tilde{F}(\epsilon)$ is negative (recall the signal components are bounded by 1) so that it is concave. Note $F_{n,\epsilon}$ itself is not necessarily concave in ϵ, although $f_{n,\epsilon}$ is. Evidently $\tilde{f}(\epsilon)$ is concave too. Concavity then allows to use the following lemma:

Lemma 5 (A bound for concave functions)**.** *Let $G(x)$ and $g(x)$ be concave functions. Let $\delta > 0$ and define $C_\delta^-(x) \equiv g'(x-\delta) - g'(x) \geq 0$ and $C_\delta^+(x) \equiv g'(x) - g'(x+\delta) \geq 0$. Then*

$$|G'(x) - g'(x)| \leq \delta^{-1}\sum_{u\in\{x-\delta,\,x,\,x+\delta\}} |G(u) - g(u)| + C_\delta^+(x) + C_\delta^-(x)\,.$$

First, from (47) we have

$$\tilde{F}(\epsilon) - \tilde{f}(\epsilon) = F_n(\epsilon) - f_n(\epsilon) + \sqrt{\epsilon}A_n \tag{48}$$

with $A_n \equiv \frac{1}{n}\sum_{i=1}^n(|\tilde{z}_i| - \mathbb{E}\,|\tilde{Z}_i|)$. Second, from (43), (45) we obtain for the ϵ-derivatives

$$\tilde{F}'(\epsilon) - \tilde{f}'(\epsilon) = \langle\mathcal{L}\rangle_\epsilon - \mathbb{E}\langle\mathcal{L}\rangle_{t,\epsilon} + \frac{A_n}{2\sqrt{\epsilon}}\,. \tag{49}$$

From (48) and (49) it is then easy to show that Lemma 5 implies

$$|\langle \mathcal{L} \rangle_\epsilon - \mathbb{E}\langle \mathcal{L} \rangle_\epsilon| \leq \delta^{-1} \sum_{u \in \{\epsilon-\delta,\, \epsilon,\, \epsilon+\delta\}} \big(|F_n(u) - f_n(u)| + |A_n|\sqrt{u}\big) + C_\delta^+(\epsilon) + C_\delta^-(\epsilon) + \frac{|A_n|}{2\sqrt{\epsilon}} \tag{50}$$

where $C_\delta^-(\epsilon) \equiv \tilde{f}'(\epsilon - \delta) - \tilde{f}'(\epsilon) \geq 0$ and $C_\delta^+(\epsilon) \equiv \tilde{f}'(\epsilon) - \tilde{f}'(\epsilon + \delta) \geq 0$. Note that δ will be chosen later on strictly smaller than $\underline{\epsilon}$ so that $\epsilon - \delta \geq \underline{\epsilon} - \delta$ remains positive. Remark that by independence of the noise variables $\mathbb{E}[A_n^2] = (1 - 2/\pi)/n < 1/n$. We square the identity (50) and take its expectation. Then using $(\sum_{i=1}^p v_i)^2 \leq p \sum_{i=1}^p v_i^2$, and that $\epsilon \leq \bar{\epsilon}$, as well as the free energy concentration

$$\frac{1}{9}\mathbb{E}\big[(\langle \mathcal{L} \rangle_\epsilon - \mathbb{E}\langle \mathcal{L} \rangle_\epsilon)^2\big] \leq \frac{3}{n\delta^2}(C + \bar{\epsilon} + \delta) + C_\delta^+(\epsilon)^2 + C_\delta^-(\epsilon)^2 + \frac{1}{4n\epsilon}\,. \tag{51}$$

Recall $|C_\delta^\pm(\epsilon)| = |\tilde{f}'(\epsilon \pm \delta) - \tilde{f}'(\epsilon)|$. We have

$$|\tilde{f}'(\epsilon)| \leq \frac{1}{2}\Big(\rho + \frac{1}{\sqrt{\epsilon}}\Big). \tag{52}$$

Therefore, as $\epsilon \geq \underline{\epsilon}$,

$$|C_\delta^\pm(\epsilon)| \leq \rho + \frac{1}{\sqrt{\epsilon - \delta}} \leq \rho + \frac{1}{\sqrt{\underline{\epsilon} - \delta}}.$$

We reach

$$\begin{aligned}
&\int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon \left\{C_\delta^+(\epsilon)^2 + C_\delta^-(\epsilon)^2\right\} \\
&\qquad \leq \Big(\rho + \frac{1}{\sqrt{\underline{\epsilon} - \delta}}\Big) \int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon \left\{C_\delta^+(\epsilon) + C_\delta^-(\epsilon)\right\} \\
&= \Big(\rho + \frac{1}{\sqrt{\underline{\epsilon} - \delta}}\Big)\Big[\Big(\tilde{f}(\underline{\epsilon} + \delta) - \tilde{f}(\underline{\epsilon} - \delta)\Big) + \Big(\tilde{f}(\bar{\epsilon} - \delta) - \tilde{f}(\bar{\epsilon} + \delta)\Big)\Big].
\end{aligned}$$

The mean value theorem and (52) imply $|\tilde{f}(\epsilon - \delta) - \tilde{f}(\epsilon + \delta)| \leq \delta(\rho + \frac{1}{\sqrt{\underline{\epsilon} - \delta}})$. Therefore

$$\int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon \left\{C_\delta^+(R_\epsilon)^2 + C_\delta^-(R_\epsilon)^2\right\} \leq 2\delta\Big(\rho + \frac{1}{\sqrt{\underline{\epsilon} - \delta}}\Big)^2.$$

Set $\delta = \delta_n \ll s_n \leq \underline{\epsilon}$. Thus, integrating (51) over $\epsilon \in [\underline{\epsilon}, \bar{\epsilon}]$ yields

$$\begin{aligned}
&\int_{\underline{\epsilon}}^{\bar{\epsilon}} d\epsilon\, \mathbb{E}\big[(\langle \mathcal{L} \rangle_\epsilon - \mathbb{E}\langle \mathcal{L} \rangle_{t,\epsilon})^2\big] \\
&\qquad \leq \frac{27}{n\delta_n^2}(\bar{\epsilon} - \underline{\epsilon})(C + \bar{\epsilon} + \delta_n) + 18\delta_n\Big(\rho + \frac{1}{\sqrt{\underline{\epsilon} - \delta_n}}\Big)^2 + \frac{9\ln(\bar{\epsilon}/\underline{\epsilon})}{4n} \\
&\qquad \leq \frac{C(\bar{\epsilon} - \underline{\epsilon})}{n\delta_n^2} + \frac{C\delta_n}{s_n} + \frac{C\ln(\bar{\epsilon}/\underline{\epsilon})}{n},
\end{aligned}$$

where the constant C is generic, and may change from place to place. Finally we optimize the bound choosing $\delta_n^3 = s_n(\bar{\epsilon} - \underline{\epsilon})/n$. □

2.5 Rigorous approach: the (adaptive) interpolation method

Before entering the proof let us give the generic roadmap of the adaptive interpolation method, and emphasize the main differences with the canonical Guerra-Toninelli interpolation method [9–11].

The aim of the method is to prove a variational formula for the thermodynamic limit of the free energy of some complex statistical model of interacting variables/spins. This variational formula corresponds to the extremization of a proper potential.

$i)$ The first step consists in defining an "interpolating model" parametrized by "time" $t \in [0,1]$. Its associated t-dependent mutual information $i(t)$ must interpolate between the one of the model of interest at, say, $t = 0$, and the one of a properly chosen "decoupled mean-field model" at $t = 1$ where the variables do not interact anymore and with a tractable mutual information (tractable because the system is decoupled) which constructs part of the potential. The basic idea is therefore similar to the canonical interpolation method except that usually the interpolation path depends "trivially" on t while in the adaptive interpolation method the interpolation path is generic, and is parametrized by an *interpolation function* that allows for much more flexibility.

$ii)$ In the second step we want to "compare" the two boundary values using $i(0) = i(1) - \int_0^1 dt\, i'(t)$, where $i(0) \approx \frac{1}{n} I(\boldsymbol{X}^*; \boldsymbol{Y})$ is what we want to compute while $i(1)$ is a piece of the potential. We therefore need to compute the t-derivative $i'(t)$. When $i'(t)$ is then plugged in the previous relation this gives the so-called sum rule, which links the mutual information of interest and the potential (or part of it).

$iii)$ The third step consists in simplifying the obtained sum rule thanks to the concentration of the identified order parameter of the problem (the overlap with the planted solution/signal in Bayesian inference problems). Self-averaging of the order parameter, refered to as *replica symmetry* [12], has to be proven for all $t \in [0,1]$. It requires a proper "perturbation" of the model with a strenght controlled by a perturbation parameter ϵ. Perturbing the system allows to "avoid" possible isolated phase transitions points where concentration does not occur. This step is model-dependent as such results can be proven only under specific settings. For ferromagnetic models (such as the Curie-Weiss model) at any temperature, Bayesian inference (in the so-called "Bayesian optimal setting"), or generic disordered spin models at high temperature this is doable. In the first case thanks to the ferromagnetic nature of the model (see [13] for a proof that ferromagnetism implies replica symmetry in full generality), in the second case thanks to the Nishimori identity implying plethora of sub-identites for the correlation functions of the model, and in the last case by concentration techniques [14, 15]. But away from these settings, e.g., in combinatorial optimization, in generic dis-

ordered spin models at low temperature, or in non-optimal Bayesian inference, this is usually not possible as *replica symmetry breaking* may occur [12, 14, 15] and prevents the order parameter to concentrate[19].

iv) In a fourth step, once the sum rule has been simplified thanks to the order parameter concentration, the flexibility allowed by the choice of the interpolation functions (i.e., the choice of the interpolation path) is exploited in order to obtain two matching bounds for the mutual information. One bound is simply obtained by choosing a "trivial" interpolation path. The other one requires a smarter choice: it appears that, given the decoupled model towards which we interpolate, there is a unique choice of the interpolation functions allowing to obtain the converse bound. This choice corresponds to the solution of a first order differential equation over the interpolation functions (in which the perturbation parameter ϵ will play the role of the initial condition). The interpolation functions have therefore been *adapted* in order to finish the proof, thus the name of the method.

2.5.1 The interpolating model

Let $\epsilon \in [s_n, 2s_n]$, for some sequence $(s_n) \in (0, 1/2)^{\mathbb{N}}$ that tends to 0_+ as $s_n = (1/2)n^{-\alpha}$ for $\alpha > 0$. Let $q_n : [0,1] \times [s_n, 2s_n] \mapsto [0, \rho]$ and set the *interpolating function*

$$R_n(t,\epsilon) \equiv \epsilon + \lambda \int_0^t dt'\, q_n(t',\epsilon). \tag{53}$$

Consider the following interpolating $(n, t, R_n(t,\epsilon))$-dependent estimation model, where $t \in [0,1]$ is the interpolation parameter, with accessible data $\boldsymbol{y} = \boldsymbol{y}(t)$ and $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{y}}(t,\epsilon)$ obtained through

$$\begin{cases} y_{ij} &= \sqrt{(1-t)\frac{\lambda}{n}}\, x_i^* x_j^* + z_{ij}\,, \qquad 1 \le i < j \le n\,, \\ \tilde{\boldsymbol{y}} &= \sqrt{R_n(t,\epsilon)}\, \boldsymbol{x}^* + \tilde{\boldsymbol{z}}\,, \end{cases}$$

where all the z's are all i.i.d. outcomes of a $\mathcal{N}(0,1)$ random variable, with $z_{ij} = z_{ji}$. The posterior associated with this model reads

$$P(\boldsymbol{x}|t, R_n(t,\epsilon), \boldsymbol{y}, \tilde{\boldsymbol{y}}) = \frac{P(\boldsymbol{x}) e^{-\mathcal{H}(\boldsymbol{x};t,R_n(t,\epsilon),\boldsymbol{y},\tilde{\boldsymbol{y}})}}{\mathcal{Z}_n(t, R_n(t,\epsilon), \boldsymbol{y}, \tilde{\boldsymbol{y}})} \tag{54}$$

[19] At the moment the adaptive interpolation method is specifically designed for replica symmetric models but it is an interesting direction to see wether it can tackle more complicated models where replica symmetry breaking occurs.

with factorized prior $P(\boldsymbol{x}) = \prod_{i=1}^n P(x_i)$, and normalization (or partition function) and interpolating Hamiltonian given by

$$\begin{aligned}
\mathcal{Z}_n(t, R_n(t,\epsilon), \boldsymbol{y}, \tilde{\boldsymbol{y}}) &\equiv \int dP(\boldsymbol{x}) e^{-\mathcal{H}(\boldsymbol{x}; t, R_n(t,\epsilon), \boldsymbol{y}, \tilde{\boldsymbol{y}})} \\
\mathcal{H}_n(\boldsymbol{x}; t, R_n(t,\epsilon), \boldsymbol{y}, \tilde{\boldsymbol{y}}) &\equiv \sum_{i<j}^n \Big((1-t)\frac{\lambda}{n}\frac{x_i^2 x_j^2}{2} - \sqrt{(1-t)\frac{\lambda}{n}} x_i x_j y_{ij} \Big) \\
&\quad + \sum_{i=1}^n \Big(R_n(t,\epsilon)\frac{x_i^2}{2} - \sqrt{R_n(t,\epsilon)} x_i \tilde{y}_i \Big).
\end{aligned}$$

We also define the mutual information for the interpolating model:

$$i_n(t,\epsilon) \equiv \frac{1}{n} I\big(\boldsymbol{X}^*; \boldsymbol{Y}(t), \tilde{\boldsymbol{Y}}(t,\epsilon)\big).$$

The $(n, t, R_n(t,\epsilon))$-dependent Gibbs-bracket (that we simply denote $\langle - \rangle_t$ for the sake of readibility) is defined as usual for functions $A(\boldsymbol{x}) = A$:

$$\langle A \rangle_t = \langle A \rangle_{n,t,R_n(t,\epsilon)} \equiv \int dP(\boldsymbol{x}|t, R_n(t,\epsilon), \boldsymbol{Y}, \tilde{\boldsymbol{Y}})\, A(\boldsymbol{x})\,.$$

By design of the interpolating model we have:

Lemma 6 (Boundary values)**.** *The mutual information for the interpolating model verifies*

$$\begin{cases}
i_n(0,\epsilon) &= \frac{1}{n} I(\boldsymbol{X}^*; \boldsymbol{Y}(0)) + \frac{1}{n} I(\boldsymbol{X}^*; \tilde{\boldsymbol{Y}}(0,\epsilon) | \boldsymbol{Y}(0)) \\
&= \frac{1}{n} I(\boldsymbol{X}^*; \boldsymbol{Y}) + O(s_n)\,, \\
i_n(1,\epsilon) &= \frac{1}{n} I(\boldsymbol{X}^*; \tilde{\boldsymbol{Y}}(1,\epsilon)) + \frac{1}{n} I(\boldsymbol{X}^*; \boldsymbol{Y}(1) | \tilde{\boldsymbol{Y}}(1,\epsilon)) \\
&= I(X^*; \{\lambda \textstyle\int_0^1 dt\, q_n(t,\epsilon)\}^{1/2} X^* + Z) + O(s_n)\,.
\end{cases}$$

where $I(X^*; \{\lambda \int_0^1 dt\, q_n(t,\epsilon)\}^{1/2} X^* + Z)$ is the mutual information for a scalar Gaussian channel (here $X^* \sim P$, $Z \sim \mathcal{N}(0,1)$):

$$y = \sqrt{\lambda \textstyle\int_0^1 dt\, q_n(t,\epsilon)}\, x^* + z.$$

Proof. The first equality follows from the chain rule for mutual information. The second one relies on $I(\boldsymbol{X}^*; \boldsymbol{Y}(0)) = I(\boldsymbol{X}^*; \boldsymbol{Y})$ which is obvious, as well as

$$\frac{1}{n} I(\boldsymbol{X}^*; \tilde{\boldsymbol{Y}}(0,\epsilon) | \boldsymbol{Y}(0)) = O(s_n)\,. \tag{55}$$

This claim simply follows from the I-MMSE relation and $R_n(0,\epsilon) = \epsilon$:

$$\frac{d}{d\epsilon}\frac{1}{n} I(\boldsymbol{X}^*; \tilde{\boldsymbol{Y}}(0,\epsilon) | \boldsymbol{Y}(0)) = \frac{1}{2n} \mathrm{MMSE}(\boldsymbol{X}^* | \tilde{\boldsymbol{Y}}(0,\epsilon), \boldsymbol{Y}(0)) \le \frac{\rho}{2}\,. \tag{56}$$

This last inequality is true because $\mathrm{MMSE}(\boldsymbol{X}^*|\tilde{\boldsymbol{Y}}(0,\epsilon),\boldsymbol{Y}(0)) \le \mathbb{E}\|\boldsymbol{X}^*\|^2 = n\rho$, as the components of $\boldsymbol{X}^*$ as i.i.d. from P. Therefore $\frac{1}{n}I(\boldsymbol{X}^*;\tilde{\boldsymbol{Y}}(0,\epsilon)|\boldsymbol{Y}(0))$ is $\frac{\rho}{2}$-Lipschitz in $\epsilon \in [s_n, 2s_n]$. Moreover $I(\boldsymbol{X}^*;\tilde{\boldsymbol{Y}}(0,0)|\boldsymbol{Y}(0)) = 0$. This implies the claim (55).

The third equality follows again from the chain rule for mutual information. The last equality uses $I(\boldsymbol{X}^*;\boldsymbol{Y}(1)|\tilde{\boldsymbol{Y}}(1,\epsilon)) = 0$ as $\boldsymbol{Y}(1)$ does not depend on $\boldsymbol{X}$. Moreover, by decoupling,

$$\begin{aligned}\frac{1}{n}I(\boldsymbol{X}^*;\tilde{\boldsymbol{Y}}(1,\epsilon)) &= I(X^*;\sqrt{R_n(1,\epsilon)}X^* + Z)\\ &= I(X^*;\{\lambda\textstyle\int_0^1 dt\, q_n(t,\epsilon)\}^{1/2}X^* + Z) + O(s_n)\,.\end{aligned}$$

The last step follows from $I(X^*;\sqrt{\gamma}X^* + Z)$ being a $\frac{\rho}{2}$-Lipschitz function of γ, again shown by the I-MMSE relation as for (56). □

2.5.2 Fundamental sum rule

The core identity of our proof is:

Proposition 5 (Sum rule). *The mutual information verifies the following sum rule:*

$$\frac{1}{n}I(\boldsymbol{X};\boldsymbol{Y}) = i_n^{(\mathrm{RS})}\big(\textstyle\int_0^1 dt\, q_n(t,\epsilon);\lambda,\rho\big) + \frac{\lambda}{4}\big(\mathcal{R}_1 - \mathcal{R}_2 - \mathcal{R}_3\big) + O(s_n) \quad (57)$$

with non-negative (n, R_n, ϵ)-dependent "remainders"

$$\begin{cases}\mathcal{R}_1 &\equiv \int_0^1 dt\,\big(q_n(t,\epsilon) - \int_0^1 ds\, q_n(s,\epsilon)\big)^2\,,\\ \mathcal{R}_2 &\equiv \int_0^1 dt\,\mathbb{E}\big\langle\big(Q - \mathbb{E}\langle Q\rangle_t\big)^2\big\rangle_t\,,\\ \mathcal{R}_3 &\equiv \int_0^1 dt\,\big(\mathbb{E}\langle Q\rangle_t - q_n(t,\epsilon)\big)^2\,.\end{cases}$$

Proof. We compare the boundaries using the fundamental theorem of calculus

$$i_n(0,\epsilon) = i_n(1,\epsilon) - \int_0^1 dt\frac{d}{dt}i_n(t,\epsilon)\,.$$

The t-derivative of the interpolating mutual information is simply computed combining the I-MMSE relation with the chain rule for derivatives:

$$\begin{aligned}&\frac{d}{dt}i_n(t,\epsilon)\\ &\quad= -\frac{\lambda}{2}\frac{1}{n^2}\sum_{i>j}\mathbb{E}\big[(X_i^*X_j^* - \langle X_jX_j\rangle_t)^2\big] + \frac{\lambda q_n(t,\epsilon)}{2}\frac{1}{n}\mathbb{E}\|\boldsymbol{X}^* - \langle\boldsymbol{X}\rangle_t\|_2^2\\ &\quad= -\frac{\lambda}{4}\frac{1}{n^2}\mathbb{E}\|\boldsymbol{X}^*\otimes\boldsymbol{X}^* - \langle\boldsymbol{X}\otimes\boldsymbol{X}\rangle_t\|_{\mathrm{F}}^2 + \frac{\lambda q_n(t,\epsilon)}{2}\big(\rho - \mathbb{E}\langle Q\rangle_t\big) + O(1/n).\end{aligned}$$

The $O(1/n)$ term comes form completing the diagonal in the sum $\sum_{i>j}$ in order to construct the matrix-MMSE. The second step used the following identity that we have shown previously in (6) based on the Nishimori identity:

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{X}^* - \langle \boldsymbol{X}\rangle_t\|_2^2 \overset{\text{N}}{=} \rho - \mathbb{E}\langle Q\rangle_t \qquad \text{where} \qquad Q \equiv \frac{1}{n}\boldsymbol{X}^*\cdot\boldsymbol{X}.$$

By similar manipulations we obtain for the matrix-MMSE

$$\frac{1}{n^2}\mathbb{E}\|\boldsymbol{X}^*\otimes\boldsymbol{X}^* - \langle \boldsymbol{X}\otimes\boldsymbol{X}\rangle_t\|_{\text{F}}^2 \overset{\text{N}}{=} \rho^2 - \mathbb{E}\langle Q^2\rangle_t.$$

Combining everything with the boundary values, and once replaced into the fundamental theorem of calculus we deduce

$$\begin{aligned}\frac{1}{n}I(\boldsymbol{X}^*;\boldsymbol{Y}) = I_n\big(X^*;\{\lambda\textstyle\int_0^1 dt\, q_n(t,\epsilon)\}^{1/2}X^* + Z\big)\\ + \frac{\lambda}{4}\int_0^1 dt\,\big\{\rho^2 - \mathbb{E}\langle Q^2\rangle_t - 2q_n(t,\epsilon)(\rho - \mathbb{E}\langle Q\rangle_t)\big\} + O(s_n),\end{aligned}$$

where the constants in $O(s_n)$ are uniform in ϵ, t, n. Let us re-arrange so that the replica symmetric potential appears:

$$\begin{aligned}\frac{1}{n}I(\boldsymbol{X}^*;\boldsymbol{Y}) = i_n^{(\text{RS})}\big(\textstyle\int_0^1 dt\, q_n(t,\epsilon);\lambda,\rho\big) - \frac{\lambda}{4}\big\{\int_0^1 dt\, q_n(t,\epsilon) - \rho\big\}^2\\ + \frac{\lambda}{4}\int_0^1 dt\big\{\rho^2 - \mathbb{E}\langle Q^2\rangle_t - 2q_n(t,\epsilon)(\rho - \mathbb{E}\langle Q\rangle_t)\big\} + O(s_n)\end{aligned}$$

which, after a line of basic algebra, finally simplifies to the claimed sum rule. □

2.5.3 Upper bound: "trivial" interpolation path

The replica symmetric formula for the mutual information follows directly from the two bounds proven below.

Proposition 6 (Upper bound)**.** *We have*

$$\limsup_{n\to+\infty}\frac{1}{n}I(\boldsymbol{X}^*;\boldsymbol{Y}) \le \inf_{q\in[0,\rho]} i_n^{(\text{RS})}(q;\lambda,\rho)\,.$$

Proof. Fix for all $t\in[0,1]$:

$$q_n(t,\epsilon) = \operatorname*{argmin}_{q\in[0,\rho]} i_n^{(\text{RS})}(q;\lambda,\rho).$$

The interpolation is therefore a simple linear (in time) path. The "interpolation path variance" $\mathcal{R}_1$ cancels. The two other remainders $\mathcal{R}_2$ and $\mathcal{R}_3$ being non-negative we reach the result. □

2.5.4 Lower bound: the adaptive interpolation path

We start with a definition: we say that the map $\epsilon \mapsto R_n(t,\epsilon)$ is *regular* if it is a $\mathcal{C}^1$ diffeomorphism whose Jacobian is greater or equal to one for all $t \in [0,1]$.

Proposition 7 (Lower bound). *We have*

$$\liminf_{n\to+\infty} \frac{1}{n} I(\boldsymbol{X}^*;\boldsymbol{Y}) \geq \inf_{q\in[0,\rho]} i_n^{(\mathrm{RS})}(q;\lambda,\rho). \tag{58}$$

Proof. In order to control $\mathcal{R}_2$ we need to prove the overlap concentration for the interpolating model. Looking at section 2.4 we see that in the interpolating model $R_n(t,\epsilon)$ plays the role of ϵ. So integrating the overlap fluctuation for the interpolating model $\mathcal{R}_2$ over $R_n(t,\epsilon) \in [s_n, 2s_n]$ would allow to show that it is small. But we need $R_n(t,\epsilon)$ to be free in order to choose it smartly later on in the proof. The perturbation parameter over which we integrate really needs to be ϵ. First note that (42) generalizes directly because it only depends on the validity of Nishimori's identity.

Now let $\underline{R} \equiv R_n(t,s_n)$, $\bar{R} \equiv R_n(t,2s_n)$. From the definition (53) of $R_n(t,\epsilon)$ we have $[\underline{R},\bar{R}] \subseteq [s_n, 2s_n+\lambda\rho]$. Using that the map $\epsilon \mapsto R_n(t,\epsilon)$ is regular, Proposition 4 combined with Fubini's theorem for interverting the t and $R_n(t,\epsilon)$ integrals implies

$$\begin{aligned} &C_1\Big(\frac{s_n+\lambda\rho}{ns_n^2}\Big)^{1/3} + \frac{C_2}{n}\ln\frac{2s_n+\lambda\rho}{s_n} + \frac{C_3}{n} \\ &\qquad \geq \int_{\underline{R}}^{\bar{R}} dR_n(t,\epsilon)\,\mathcal{R}_2 = \int_{s_n}^{2s_n} d\epsilon \frac{dR_n(t,\epsilon)}{d\epsilon}\mathcal{R}_2 \geq \int_{s_n}^{2s_n} d\epsilon\,\mathcal{R}_2. \end{aligned}$$

Choosing a proper rate of convergence to 0_+ of s_n the dominating term on the left-hand side is the first one. We then have that for some constant $C \geq 0$

$$\frac{1}{s_n}\int_{s_n}^{2s_n} d\epsilon\,\mathcal{R}_2 = \frac{1}{s_n}\int_{s_n}^{2s_n} d\epsilon \int_0^1 dt\,\mathbb{E}\big\langle (Q-\mathbb{E}\langle Q\rangle_t)^2\big\rangle_t \leq \frac{C}{(ns_n^5)^{1/3}}\,.$$

This bound is uniform in ϵ. Using this Q-concentration result under the regularity assumption for the map $\epsilon \mapsto R_n(t,\epsilon)$ as well as the fact that $\mathcal{R}_1$ is non-negative, the sum rule (57), when averaged over the perturbation, simplifies to

$$\begin{aligned} \frac{1}{n}I(\boldsymbol{X};\boldsymbol{Y}) \geq \frac{1}{s_n}\int_{s_n}^{2s_n} d\epsilon\, i_n^{(\mathrm{RS})}\big(\textstyle\int_0^1 dt\, q_n(t,\epsilon);\lambda,\rho\big) \\ - \frac{\lambda}{4}\frac{1}{s_n}\int_{s_n}^{2s_n} d\epsilon \int_0^1 dt\,\big(\mathbb{E}\langle Q\rangle_t - q_n(t,\epsilon)\big)^2 + O\Big(\frac{1}{(ns_n^5)^{1/3}}\Big)\,. \end{aligned}$$

We used that $I(\boldsymbol{X};\boldsymbol{Y})$ is independent of ϵ. At this stage it is natural to choose $q_n(t,\epsilon)$ to be the solution of

$$q_n(t,\epsilon) = \mathbb{E}\langle Q\rangle_{n,t,R_n(t,\epsilon)}\,. \tag{59}$$

Setting $G_n(t, R_n(t,\epsilon)) \equiv \mathbb{E}\langle Q\rangle_{n,t,R_n(t,\epsilon)}$, we recognize a first order ordinary differential equation

$$\frac{d}{dt}R_n(t,\epsilon) = G_n(t, R_n(t,\epsilon)) \quad \text{with initial condition} \quad R_n(0,\epsilon) = \epsilon\,. \tag{60}$$

So the perturbatio parameter ϵ actually plays the role of initial condition of the ODE naturally appearing. As $G_n(t, R_n(t,\epsilon))$ is $\mathcal{C}^1$ with bounded derivative w.r.t. its second argument the Cauchy-Lipschitz theorem implies that (60) admits a unique global solution

$$R_n^*(t,\epsilon) = \epsilon + \int_0^t ds\, q_n^*(s,\epsilon),$$

where $q_n^* : [0,1] \times [s_n, 2s_n] \mapsto [0,\rho]$ because $\mathbb{E}\langle Q\rangle_{n,t,\epsilon} \in [0,\rho]$ (as seen from (6)). Under the choice R_n^* let us check the regularity assumption that we assumed until now. By Liouville's formula the flow $\epsilon \mapsto R_n^*(t,\epsilon)$ satisfies

$$\frac{d}{d\epsilon}R_n^*(t,\epsilon) = \exp\int_0^t ds\, \frac{d}{dR}G_n(s,R)\Big|_{R=R_n^*(s,\epsilon)}\,.$$

Using repeatedly the Nishimori identity one obtains

$$\frac{d}{dR}G_n(s,R) = \frac{1}{n}\sum_{i,j=1}^n \mathbb{E}\big[(\langle x_i x_j\rangle_{n,s,R} - \langle x_i\rangle_{n,s,R}\langle x_j\rangle_{n,s,R})^2\big] \geq 0$$

so that the flow has a Jacobian ≥ 1 and is a diffeomophism. Thus $\epsilon \mapsto R_n^*(t,\epsilon)$ is regular. This computation does not present any difficulty and can be found in section 6 of [16]. With the choice R_n^*, i.e., by *adapting* the interpolation path, we have then cancelled the remainder $\mathcal{R}_3$. This yields

$$\begin{aligned}\frac{1}{n}I(\boldsymbol{X};\boldsymbol{Y}) &\geq \frac{1}{s_n}\int_{s_n}^{2s_n} d\epsilon\, i_n^{(\mathrm{RS})}\big(\textstyle\int_0^1 dt\, q_n^*(t,\epsilon);\lambda,\rho\big) + O\Big(\frac{1}{(ns_n^5)^{1/3}}\Big)\\ &\geq \inf_{q\in[0,\rho]} i_n^{(\mathrm{RS})}(q;\lambda,\rho) + O\Big(\frac{1}{(ns_n^5)^{1/3}}\Big).\end{aligned}$$

Taking the $\liminf_{n\to+\infty}$ and choosing $s_n = \Theta(n^{-\alpha})$ with $\alpha \in (0,1/5)$ yields the desired result. □

2.6 A detour in physics: the cavity method for the Curie-Weiss model

The Curie-Weiss model. The Curie-Weiss model is defined by the following Hamiltonian for binary spins $\boldsymbol{\sigma} \in \{-1,1\}^n$ living on a complete graph:

$$\mathcal{H}_n(\boldsymbol{\sigma}) \equiv -\frac{J}{n}\sum_{i<j}^n \sigma_i\sigma_j - h\sum_{i=1}^n \sigma_i\,. \tag{61}$$

The Gibbs-Boltzmann measure that describes it random behavior is, as usual,

$$P(\boldsymbol{\sigma}) = \frac{e^{-\mathcal{H}_n(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} e^{-\mathcal{H}_n(\boldsymbol{\sigma})}}.$$

The temperature has been absorbed in J and h. Its free energy density is defined as (considering the temperature to be equal to one)

$$f_n = \frac{1}{n}F_n \equiv -\frac{1}{n}\ln \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} e^{-\mathcal{H}(\boldsymbol{\sigma})}\,.$$

We will (partially) prove that its free energy density is given by the following variational formulas

$$\lim_{n\to\infty} f_n = \inf_{m\in[-1,1]} f^{(\mathrm{RS})}(m),$$
$$\text{with potential} \quad f^{(\mathrm{RS})}(m) \equiv \frac{Jm^2}{2} - \ln\big(2\cosh[Jm+h]\big). \tag{62}$$

The cavity method: Aizenman-Sims-Starr bound. The cavity method allows to prove a bound for the variational expression (62). It works along the following lines. The free energy can be constructed by adding one spin after the other, and therefore is equal to the following telescopic sum:

$$f_n = \frac{1}{n}\sum_{k=0}^{n-1}(F_{k+1}-F_k) \quad \Rightarrow \quad \liminf_{n\to\infty} f_n \geq \liminf_{n\to\infty}\,(F_{n+1}-F_n)\,. \tag{63}$$

We therefore aim at computing $F_{n+1}-F_n$, i.e., the free energy cost of adding one spin to the system. Lets then add one spin $\tilde{\sigma}\equiv\sigma_{n+1}$ to the system and "isolate" its contribution in the Hamiltonian (61):

$$\mathcal{H}_{n+1}(\boldsymbol{\sigma},\tilde{\sigma}) \equiv \tilde{\mathcal{H}}_n(\boldsymbol{\sigma}) - \tilde{\sigma}(\tilde{J}_n m_n + h),$$

where the magnetization and rescaled interaction strenght are

$$m_n \equiv \frac{1}{n}\sum_{i=1}^{n}\sigma_i, \qquad \tilde{J}_n \equiv J\frac{n}{n+1}.$$

Moreover $\tilde{\mathcal{H}}_n(\boldsymbol{\sigma})$ is the Hamiltonian of the n spins system with rescaled interaction:

$$\tilde{\mathcal{H}}_n(\boldsymbol{\sigma}) \equiv -\frac{\tilde{J}_n}{n}\sum_{i<j}^{n}\sigma_i\sigma_j - h\sum_{i=1}^{n}\sigma_i.$$

The free energy of the n-spins model with rescaled interaction is

$$\tilde{F}_n \equiv -\ln \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} e^{-\tilde{\mathcal{H}}_n(\boldsymbol{\sigma})}\,.$$

Let the Gibbs-bracket associated with a generic Hamiltonian $\mathcal{H}(x)$ be

$$\langle A\rangle_{\mathcal{H}} \equiv \frac{\sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} A(\boldsymbol{\sigma}) e^{-\mathcal{H}(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} e^{-\mathcal{H}(\boldsymbol{\sigma})}}.$$

We can then write the free energy variation as

$$\begin{aligned} F_{n+1} - F_n &= (F_{n+1} - \tilde{F}_n) - (F_n - \tilde{F}_n) \\ &= -\ln \Big\langle \sum_{\tilde{\sigma}=\pm 1} e^{\tilde{\mathcal{H}}_n(\boldsymbol{\sigma}) - \mathcal{H}_{n+1}(\boldsymbol{\sigma},\tilde{\sigma})} \Big\rangle_{\tilde{\mathcal{H}}_n} + \ln \Big\langle e^{\tilde{\mathcal{H}}_n(\boldsymbol{\sigma}) - \mathcal{H}_n(\boldsymbol{\sigma})} \Big\rangle_{\tilde{\mathcal{H}}_n} \\ &= -\ln \Big\langle \sum_{\tilde{\sigma}=\pm 1} e^{\tilde{\sigma}(\tilde{J}_n m_n + h)} \Big\rangle_{\tilde{\mathcal{H}}_n} + \ln \Big\langle e^{\frac{\tilde{J}_n}{n^2}\sum_{i<j}^n \sigma_i\sigma_j} \Big\rangle_{\tilde{\mathcal{H}}_n}. \end{aligned}$$

Now we use the concentration of the magnetization under the measure $\langle - \rangle_{\tilde{\mathcal{H}}_n}$ (this step is simple here because there is no quenched disorder; in disordered models additional steps are needed). This allows to replace (away from possible phase transition points, i.e., for almost all (J,h))

$$m_n \approx \tilde{m}_n \equiv \langle m_n\rangle_{\tilde{\mathcal{H}}_n} \in [-1,1]\,.$$

This yields, using $\frac{1}{n^2}\sum_{i<j}^n \sigma_i\sigma_j = \frac{1}{2}m_n^2 + O(1/n) = \frac{1}{2}\tilde{m}_n^2 + o_n(1)$,

$$\begin{aligned} F_{n+1} - F_n &= -\ln \sum_{\tilde{\sigma}=\pm 1} e^{\tilde{\sigma}(\tilde{J}_n\tilde{m}_n + h)} + \frac{\tilde{J}_n\tilde{m}_n^2}{2} + o_n(1) \\ &= -\ln\big(2\cosh[J\tilde{m}_n + h]\big) + \frac{J\tilde{m}_n^2}{2} + o_n(1) \end{aligned} \tag{64}$$

as $\tilde{J}_n \to J$. Therefore we reach the desired bound

$$\liminf_{n\to\infty} f_n \geq \liminf_{n\to\infty} (F_{n+1} - F_n) \geq \inf_{m\in[-1,1]} f^{(\mathrm{RS})}(m)\,. \tag{65}$$

Note that we dit *not* assume existence of the thermodynamic limit $\lim_{n\to\infty} f_n$. In general settings this can be hard to prove. It took decades to find how to do for the mean-field spin glass (the Sherrington-Kirckpatrick model), and the proof gave rise to the *interpolation method*. When applying the cavity method, physicists usually assume the existence of the thermodynamic limit, and that it is given by the limit of free energy difference: $f = \lim_{n\to+\infty}(F_{n+1} - F_n)$.

Note that we could have instead started from the identity $\limsup_{n\to\infty} f_n \leq \limsup_{n\to\infty}(F_{n+1} - F_n)$ instead of (63). But we would have been stuck at the end of the proof, because this inequality does not go in the same direction as the one used when going from (64) to (65), where the latter inequality's direction is constrained by the fact that we want to show a variational formula witn an infimum, not a supremum. So the form of the variational formula fixes the initial bound to start from. So generally this method allows to rigorously prove a single-sided bound.

3 Algorithmic limits

Until now we have focused on static, ensemble (i.e., typical), properties. Mainly locating the information theoretic limits/phase transitions and the value of the optimal error we can aim for: we established the absolute fundamental limits of inference, independently of any algorithm. It is then natural to ask ourselves: for a given *instance* of the problem, can we approach these limits with efficient algorithms? The answer depends on where we are in the phase diagram (i.e., the parameters values such as the SNR). Let us study one particular algorithm that is special for a number of reasons that we will see: the approximate message-passing algorithm (AMP). This algorithm is closely connected to the so-called Thouless-Anderson-Palmer equations (TAP) from statistical physics, which is a key tool in the study of spin glasses.

3.1 Message-passing

Let us derive the AMP algorithm for the special case of the planted SK model, i.e., the components of the signal $\boldsymbol{X}^*$ are drawn i.i.d. from the Rademacher distribution (i.e., uniformly in $\{-1, 1\}$). The model under study is therefore given by the Hamiltonian (28). This will lead to great simplifications but, yet, the derivation presented here contains all key ingredients to then derive AMP algorithm in more generic settings, including other problems than probabilistic PCA. We defer to the appendices the derivation of AMP for the spiked Wigner model with generic prior.

Single instance cavity equations: belief propagation. The AMP algorithm is actually a simplified version of a more primitive algorithm called belief propagation (BP), or the sum-product algorithm. This algorithm has tons of applications in combinatorial optimization, error-correcting codes, machine learning etc. It is based on recursive equations over "messages" (also called "beliefs") flowing on the edges of the factor graph associated to the model of interest. A factor graph graphically represents a factorized probability distribution of the form

$$P(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \prod_{a=1}^{m} \psi_a(\boldsymbol{x}_a; \theta_a).$$

Here $\psi_a(\boldsymbol{x}_a; \theta_a)$ is a factor/compatibility function (a generic positively valued function) depending on a subset $\boldsymbol{x}_a = (x_{a_1}, x_{a_2}, \dots)$ of the variables $\boldsymbol{x}$, and possible parameter(s) θ_a. If we use indices i, j for variable nodes and a, b for factor

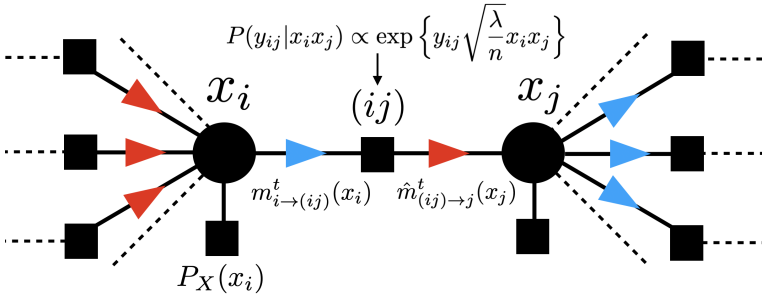


Figure 10: Visualization of the messages flowing on a piece of the factor graph for the spiked Wigner model with Rademacher variables. Factors are represented by squares, variable nodes by disks. The hanging factors of conectivity one represent the prior $\psi_i(x_i) = P_X(x_i) \propto \delta_{x_i,1} + \delta_{x_i,-1}$. From left to right: the variable node i receives factor-to-node messages (red) from its connected factors *except* the j-th one, i.e., from $n-2$ of them. It combines them and spit out a node-to-factor message $m^t_{i\to(ij)}(x_i)$ (in blue) to factor (ij) representing the compatibility function $\psi_{(ij)}(x_i, x_j; y_{ij}) \propto P(y_{ij}|x_i x_j) \propto \exp\{y_{ij}\sqrt{\lambda/n}x_i x_j\}$. The factor (ij) then computes the factor-to-node message $\hat{m}^t_{(ij)\to j}(x_j)$ (red) according to the BP rule and send the result to its next neighbor j. And so forth.

nodes the BP recursions for the messages are:

$$\text{Belief propagation: for } i = 1, \dots, n \text{ and } a = 1, \dots, m:$$

$$\hat{m}^t_{a\to i}(x_i) = \frac{1}{\hat{\mathcal{Z}}^t_{a\to i}} \int \psi_a(\boldsymbol{x}_a; \theta_a) \prod_{j\in\partial_a\setminus i} m^t_{j\to a}(x_j) dx_j,$$

$$m^{t+1}_{i\to a}(x_i) = \frac{1}{\mathcal{Z}^{t+1}_{i\to a}} \prod_{b\in\partial_i\setminus a} \hat{m}^t_{b\to i}(x_i).$$

The notation $j \in \partial_a\setminus i$ means all variable nodes belonging to the neighbor of (i.e., that share an edge with) the factor node a, except the i-th one. Similarly $b \in \partial_i\setminus a$ is the set of factor nodes connected to variable node i except the a-th one. The "factor-to-node message" $\hat{m}^t_{a\to i}(x_i)$ represents the current belief (marginal probability) of variable node i taking value x_i at iteration t in a modified factor graph, called *cavity graph*, where the node i is connected to factor a only (i.e., $|\partial_i| - 1$ edges of the original factor graph have been removed, the others remain unchanged). Instead, the "node-to-factor message" $m^t_{i\to a}(x_i)$ is the belief of node i taking value x_i in a cavity graph where node i is connected to all its neighbors in the original graph except factor a (a single edge has been removed). The $\mathcal{Z}$'s are the normalization constants. In writing the update rules, we are assuming that the update is done in parallel at all the variable nodes, then in parallel at all function nodes and so on. Clearly, in this case, the iteration number must be incremented either at variable nodes or at factor nodes, but not necessarily at both. After convergence the true marginals are approximated by $m^\infty_i(x_i)$ where the BP

marginal $m_i^t(x_i)$ combines all factor-to-nodes messages (we use the symbol $\propto$ to mean "up to a normalization constant"):

$$m_i^t(x_i) \propto \prod_{a \in \partial_i} \hat{m}_{a\to i}^t(x_i).$$

These can then be used for inference by computing, e.g., their mean (for MMSE estimation) or their argmax (for MAP estimation). Fot t larger than the maximum distance $d_{\max}$ between any two nodes in the graph, the BP marginals are *exact* if the factor graph is a tree (has no loops). This is quite easy to see; to convince yourself you can write the BP equations for a small tree and update the BP equations for a number of steps $t > d_{\max}$. Runned on a graph with loops BP (then called loopy-BP) yields an approximation to the marginals (if it converges), and often comes with few or no guarantees. But we will see soon that sometimes the situation is more favorable.

For the particular case of the planted SK model the posterior is factorized as

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{\mathcal{Z}(\boldsymbol{y})} \prod_{i=1}^{n} \left(\delta_{x_i,1} + \delta_{x_i,-1}\right) \prod_{i<j} \exp\left\{y_{ij}\sqrt{\frac{\lambda}{n}} x_i x_j\right\}.$$

Because the factors $\psi_{(ij)}(x_i, x_j; y_{ij}) = \exp\{y_{ij}\sqrt{\lambda/n} x_i x_i\}$ have connectivity 2 we can index them by $(ij) = (ji)$. The BP equations are then:

Belief propagation for the planted SK model: for $i, j = 1, \dots, n$:

$$\hat{m}_{(ij)\to i}^t(x_i) \propto \int dx_j m_{j\to(ij)}^t(x_j) \exp\left\{y_{ij}\sqrt{\frac{\lambda}{n}} x_i x_j\right\},$$
$$m_{i\to(ij)}^{t+1}(x_i) \propto \left(\delta_{x_i,1} + \delta_{x_i,-1}\right) \prod_{k(\neq i,j)} \hat{m}_{(ki)\to i}^t(x_i).$$

Unfortunately BP is in this case totally inpractical, because there are $\Theta(n^2)$ messages flowing on the edges of the densely connected factor graph, and the variables x_i may in general be real. So computationally and memory-wise BP is not an efficient algorithm for inference on dense graphs, and must be used for sparse graphs instead. As we will see there is a way to overcome these difficulties.

From locally tree-like to dense graphs: approximate message-passing. We said that BP is exact on trees. It can also be used for *sparse* factor graphs (i.e., for graphs where the average connectivity of the variable and factor node does not scale with the number of variables n when it increases). In sparse graphs loops are typically of size $\Theta(\ln n)$. Therefore the shortest path between variables in cavity graphs –where direct connecting edges are removed– are typically scaling as $\Theta(\ln n)$, so locally (i.e., at any finite distance around a root node when $n \to +\infty$) the graph is a tree with high probability. This may imply low correlations between well separated variables and is the reason why BP may still work, at least

until some threshold. But what about dense factor graphs? E.g., in the one of the spiked Wigner model variables nodes have connectivity $n-1$. This is very far from tree-like graphs... Nevertheless message-passing techniques, in the form of AMP, still work. The reason is that, like in tree-like graphs, correlations between any two nodes in the graph are typically small. Even if they are connected by an edge, they are connected to so many other ones that the influence of one node onto its neighbor is vanishingly small compared to the global influence of all the other ones. This is the reason why sparse and dense graphs actually both fall in the class of mean-field models.

Many of the equalities below are valid at leading order. The derivation of the AMP algorithm then starts from the leap of faith that BP should work on densely connected graphs. The validity of this assumption will then be rigorously justified through the *state evolution analysis*. Let us define the *cavity means* as the expectation of the messages (we do not write anymore the prior explicitly but instead simply replace integrals by sums over the binary values of the variables):

$$b^t_{i\to(ij)} \equiv \sum_{x=\pm 1} x\, m^t_{i\to(ij)}(x).$$

We expand the factor-to-node message in $1/\sqrt{n}$:

$$\begin{aligned}\hat{m}^t_{(ij)\to i}(x_i) &\propto \sum_{x=\pm 1} m^t_{j\to(ij)}(x)\Big\{1 + y_{ij}\sqrt{\frac{\lambda}{n}}x_i x_j + o(1/n)\Big\}\\ &= 1 + x_i y_{ij}\sqrt{\frac{\lambda}{n}} b^t_{j\to(ij)} + o(1/n)\\ &\approx \exp\{x_i \hat{B}^t_{(ij)\to i}\} \qquad \text{where} \qquad \hat{B}^t_{(ij)\to i} \equiv \sqrt{\frac{\lambda}{n}} y_{ij} b^t_{j\to(ij)}.\end{aligned}$$

At the first step the next x_i-dependent terms were $o(1/n)$ so they could be dropped safely. $\hat{B}^t_{(ij)\to i}$ is called *cavity field*, and is an effective field felt by the node i in a cavity graph where it is only connected to (ij). We plug this in the second BP equation for the node-to-factor message:

$$m^{t+1}_{i\to(ij)}(x_i) \propto \exp\{x_i B^t_{i\to(ij)}\} \qquad \text{where} \qquad B^t_{i\to(ij)} \equiv \sum_{k(\neq i,j)} \hat{B}^t_{(ki)\to i}.$$

$B^t_{i\to(ij)}$ is a cavity field felt by the variable in the cavity graph where it is connected to all but the (ij) factor. Computing the mean $b^{t+1}_{i\to(ij)}$ of $m^{t+1}_{i\to(ij)}(x_i)$ we obtain closed equations on the cavity means:

Relaxed belief propagation: for $i, j = 1, \dots, n$:

$$b^{t+1}_{i\to(ij)} = \eta(B^t_{i\to(ij)}) = \tanh B^t_{i\to(ij)}, \quad B^t_{i\to(ij)} = \sqrt{\frac{\lambda}{n}} \sum_{k(\neq i,j)} y_{ki} b^t_{k\to(ik)}. \quad (66)$$

The initial value for the the cavity means are random numbers of the order $\epsilon \ll 1$. This allows to lower initial bias while breaking the possible all zeros fixed point. The function η is called *denoiser* in the AMP terminology. The resulting algorithm is called *relaxed BP*, as the message passing is over simple parameters instead of distributions.

In relaxed BP cavity means are flowing on the factor graph edges, so there are $\Theta(n^2)$ which is computationally and memory costly. The next step is to re-express the equations in terms of *marginal means*

$$b_i^t \equiv \sum_{x=\pm 1} x\, m_i^t(x).$$

The AMP marginal means approximate the true means/magnetizations of the variables. We obtain similarly as before

$$b_i^{t+1} = \eta(B_i^t), \qquad B_i^t = \sqrt{\frac{\lambda}{n}} \sum_{k(\neq i)} y_{ki} b_{k\to(ik)}^t. \tag{67}$$

We expand the cavity mean

$$\begin{aligned} b_{k\to(ik)}^t &= \eta\Big(\sqrt{\frac{\lambda}{n}} \sum_{\ell(\neq k)} y_{\ell k} b_{\ell\to(\ell k)}^{t-1} - \sqrt{\frac{\lambda}{n}} y_{ik} b_{i\to(ik)}^{t-1}\Big) \\ &= \eta(B_k^{t-1}) - \sqrt{\frac{\lambda}{n}} y_{ik} b_{i\to(ik)}^{t-1} \eta'(B_k^{t-1}) + O(1/n) \\ &= b_k^t - \sqrt{\frac{\lambda}{n}} y_{ik} b_{i\to(ik)}^{t-1} (1 - (b_k^t)^2) + O(1/n) \\ &= b_k^t - \sqrt{\frac{\lambda}{n}} y_{ik} b_i^{t-1} (1 - (b_k^t)^2) + O(1/n). \end{aligned} \tag{68}$$

We used $\eta'(B_k^{t-1}) = 1 - \eta(B_k^{t-1})^2 = 1 - (b_k^t)^2$, as well as (68) yielding

$$b_{i\to(ik)}^{t-1} = b_i^{t-1} + O(1/\sqrt{n}) \tag{69}$$

that we used in the last step. Plugging this back in (67) we close the equations on the marginal means and get at leading order:

Approximate message-passing: for $i = 1, \dots, n$:

$$b_i^{t+1} = \eta(C_i^t), \qquad C_i^t = \sqrt{\frac{\lambda}{n}} \sum_{k(\neq i)} y_{ki} b_k^t - \frac{\lambda b_i^{t-1}}{n} \sum_{k(\neq i)} y_{ki}^2 (1 - (b_k^t)^2). \tag{70}$$

Again the initialization is random times a small constant $\epsilon \ll 1$. At each step the rank-one spike is estimated as

$$\boldsymbol{b}^{t+1} \otimes \boldsymbol{b}^{t+1} = (b_i^{t+1} b_j^{t+1}).$$

The AMP algorithm, that is equivalent to BP at leading order, iterates only n quantities instead of $\Theta(n^2)$ which is computationally and memory-wise much more efficient. When the signal components are drawn from a generic prior P_X rather than Rademacher the messages cannot be parametrized anymore by their mean only. The messages variances has to be tracked as well. We derive this more general AMP as well as its asymptotic analysis in the appendix.

The fixed point equations associated with AMP in the case of y_{ij} being outcomes of a standard Gaussian are the so-called Thouless-Anderson-Palmer self-consistency equations for the magnetizations of the SK model (without external field; just add the external field value h to the argument of the $\tanh$ below if it is present):

$$\text{Thouless-Anderson-Palmer equations: for } i = 1, \dots, n\text{:}$$
$$b_i = \tanh\Big(\frac{1}{\sqrt{n}} \sum_{k(\neq i)} y_{ki} b_k - \frac{b_i}{n} \sum_{k(\neq i)} y_{ki}^2 (1 - b_k^2)\Big).$$

Note that the time indices combinaison is non-trivial in AMP. Any other combinaison will lead to convergence issues. One cannot simply start from the TAP equations and index by $t+1$ on the left of the equality and t on the right. The term

$$-\frac{\lambda b_i^{t-1}}{n} \sum_{k(\neq i)} y_{ki}^2 (1 - (b_k^t)^2)$$

in AMP (or the similar one without time indices in the TAP equations) is called *Onsager reaction term*. This term with the proper time indices is absolutely crucial for the convergence of AMP and makes all the diffeence between AMP (or TAP) and the "naive mean-field" algorithm/equations:

$$\text{Naive mean-field: for } i = 1, \dots, n\text{:}$$
$$b_i^{t+1} = \eta\Big(\sqrt{\frac{\lambda}{n}} \sum_{k(\neq i)} y_{ki} b_k^t\Big), \qquad b_i = \eta\Big(\sqrt{\frac{\lambda}{n}} \sum_{k(\neq i)} y_{ki} b_k\Big).$$

3.2 State evolution, and optimality of AMP

We will now heuristically show that due to the presence of the Onsager reaction term, the fields (C_i^t) in AMP asymptotically behave as independent Gaussian random variables. This will permit us to compute their distribution and as a consequence to track AMP's performance using the *state evolution* analysis. In order to understand the mechanism behind state evolution, it is easier to first gain some intuition on what happens on locally tree-like graphs for which BP has originally been formulated.

Asymptotic analysis of belief propagation: density evolution. We consider that the graph is an instance of an ensemble of large locally tree-like graphs

with loops of typical size at least $\Theta(\ln n)$ with high probability as $n \to +\infty$ (this is the case, e.g., for sparse regular or Erdöes-Renyi random graphs). We analyze the plain belief propagation algorithm. *Density evolution* is then a statistical analysis of the messages distributions. It assumes:

- The number of iterations t is fixed while the number of variables $n \to +\infty$. Therefore, as a consequence of the locally tree-like structure the t-radius neighborhood of a root node drawn at random is a tree with probability 1.
- The BP messages are initialized independently.

Under these assumptions you can convince yourself that *two messages selected randomly in the graph that are iterated through the BP algorithm are independent with probability one.* Indeed, a message (of any of the two types) associated with a root node can only depend after t iterations on the messages associated with nodes that are at a distance lower than $t+1$ along the directed tree starting at this root. Because $t/n \to 0$ the claim follows. This implies the following distributional equations, coined *density evolution* (as they track the density of messages in the limit $n \to +\infty$ and $t \leq T$ finite):

$$\text{Density evolution: for } t = 1, \dots, T\text{:}$$
$$\begin{cases} \hat{m}^t & \stackrel{\mathrm{d}}{=} \hat{\Psi}_{f\to n}\big((m_j^t)_{j=1}^{\alpha_f - 1}, \psi\big), \\ m^{t+1} & \stackrel{\mathrm{d}}{=} \Psi_{n\to f}\big((\hat{m}_b^t)_{b=1}^{\alpha_v - 1}\big). \end{cases}$$

Here the (m_j^t) are i.i.d. copies of the random message/distribution m^t, and $(\hat{m}_b^t)$ i.i.d. copies of $\hat{m}^t$. The functional $\hat{\Psi}_{f\to n}$ represents the first BP update rule for factor-to-nodes messages, $\Psi_{n\to f}$ the BP update rule for the node-to-factor messages. ψ is a random factor/compatibility function distributed as the ones in the problem under study. E.g., in the planted SK model a random factor takes the form $\exp\{(\lambda/n)X_1^* X_2^* x_1 x_2 + Z\sqrt{\lambda/n} x_1 x_2\}$ with X_1^*, X_2^* drawn independently from P_X and Z a standard Gaussian, x_1 and x_2 the arguments of the random messages m_1^t and m_2^t entering in $\hat{\Psi}_{f\to n}$. The random variables α_f and α_v have distributions corresponding respectively to the ones of the factor nodes connectivity, and variable nodes connectivity. These take care of the graph randomness. For the spiked Wigner model they are deterministically equal to $\alpha_f = 2$ and $\alpha_v = n - 2$; in sparse random graphs they are generally random variables with, e.g., a Poisson distribution in the case of sparse Erdöes-Renyi graphs.

The density evolution distributional equations can be solved by *population dynamics*, where large populations of messages are kept in memory and randomly updated through the density evolution rule. The empirical expectation of the messages over the population approximates the true message densities. If a given graph instance drawn from the ensemble under study is very large (and therefore typical with high propability) compared to the number of BP updates, density evolution should accurately predict the empirical distribution of BP messages for

this particular instance by the law of large numbers. See [1, 3, 17] for knowing more.

Asymptotic analysis of approximate message-passing: state evolution. We will actually analyze the relaxed BP algorithm (66) which is equivalent to AMP at leading order, in the sense that its approximation (67) for the marginal means are asymptotically the same as the AMP estimates.

For the state evolution asymptotic analysis of relaxed BP/AMP over dense graphs, we assume similarly as in the density evolution analysis that $t/n \to 0$ in the thermodynamic limit $n \to +\infty$ and that the cavity means are initialized independently. But why would this imply independence of the cavity means, as in a dense graph any two nodes are essentially at distance 1? In order to heuristically estimate the dependence between two given cavity means $b^t_{i\to(ij)}$ and $b^t_{k\to(k\ell)}$ we approximately count the number of directed paths of lenght at most t that are starting from node i and passing at least once through node k. It is only through such paths that information may flow, and therefore create correlations, between the two cavity means. We assume that the denoiser η is Lipschitz with finite Lipschitz constant and that all paths have statistically the same weight/influence; the weights are related to the observations which are statistically equivalent. Let us use a probabilistic argument rather than combinatorial. The probability that a uniformly sampled path of lenght t with initial node i avoid node k is $(\frac{n-2}{n-1})^t \approx \exp\{-t/(n-1)\}$ when $n \gg 1$. So the probability that it crosses node k at least once is approximately $1 - \exp\{-t/(n-1)\}$ which tends to t/n if $n \gg t$. So if we were excluding all these computation paths in the iterations of the relaxed BP iterations in order to make the two cavity means strictly independent, because η is Lipschitz this would have an impact of $O(t/n)$ on the value of $b^t_{k\to(k\ell)}$. Therefore different cavity means are asymptotically independent as $n \to +\infty$ and t fixed. By the same argument a cavity mean $b^t_{i\to(ij)}$ at any fixed time t is asymptotically independent of the observation $y_{k\ell}$ (and therefore of $z_{k\ell}$) for all $(k\ell) \neq (ij)$. This is because information about $y_{k\ell}$ can only reach $b^t_{i\to(ij)}$ through this vanishing fraction of correlating paths. Recall also that the cavity mean $b^t_{i\to(ij)}$ is a marginal mean in a cavity graph where factor (ij) is not present, so $b^t_{i\to(ij)}$ is also supposed independent of y_{ij} (and therefore of z_{ij}). It might seem paradoxal that the cavity means are overall independent of all data, but all this reasonning is only true at leading order. Dependencies appear only at lower order which simplifies a lot the analyis. Note that if t is comparable to n or larger all the argument collapses and cavity means become strongly dependent, thus the assumption of the state evolution analysis.

Let us now use the cavity means independence assumption to derive the state evolution recursion. The observations (y_{ik}) are independent conditionally on the signal $\boldsymbol{x}^*$ because the noise is independent for each observations. At fixed $\boldsymbol{x}^*$ the BP cavity field $B^t_{i\to(ij)}$ is therefore a sum of (asymptotically) independent terms. By the central limit theorem it tends to a Gaussian random variable. We compute

its conditional mean (all that at leading order):

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x}^*} B^t_{i\to(ij)} &= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x}^*}\sqrt{\frac{\lambda}{n}}\sum_{k(\neq i,j)} Y_{ki} b^t_{k\to(ik)} \\ &= \sum_{k(\neq i,j)} \mathbb{E}_{\boldsymbol{Z}}\Big[\frac{\lambda}{n} x^*_i x^*_k b^t_{k\to(ik)} + \sqrt{\frac{\lambda}{n}} Z_{ik} b^t_{k\to(ik)}\Big] \\ &= \lambda x^*_i \frac{1}{n}\sum_k x^*_k b^t_k + O(1/\sqrt{n}) \\ &= \lambda x^*_i Q(\boldsymbol{x}^*, \boldsymbol{b}^t) + O(1/\sqrt{n}).\end{aligned}$$

Recall that at leading order cavity means are independent between themselves and of the data. We used for the third step (69) that reads $b^t_{k\to(ik)} = b^t_k + O(1/\sqrt{n})$. The independence assumption of the cavity means implies that the cavity fields

$$B^t_{i\to(ij)} = \eta^{-1}(b^{t+1}_{i\to(ij)})$$

are also pairwise independent (one can also explicitely compute the covariance $\mathrm{Cov}_{\boldsymbol{Z}}(B^t_{i\to(ij)}, B^t_{k\to(k\ell)})$ and see it vanishes). We therefore need only to compute their variances, which is simply the sum of variances of each individual terms in the sum defining $B^t_{i\to(ij)}$ by independence:

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{Z}}\, B^t_{i\to(ij)} &= \sum_{k(\neq i,j)} \mathrm{Var}_{\boldsymbol{Z}}\Big[\frac{\lambda}{n} x^*_i x^*_k b^t_{k\to(ik)} + \sqrt{\frac{\lambda}{n}} Z_{ik} b^t_{k\to(ik)}\Big] \\ &= \frac{\lambda}{n}\sum_{k(\neq i,j)} (b^t_{k\to(ik)})^2 \\ &= \frac{\lambda}{n}\sum_{k=1}^n (b^t_k)^2 + O(1/n) \\ &= \lambda Q(\boldsymbol{b}^t, \boldsymbol{b}^t) + O(1/n).\end{aligned}$$

We conclude that at leading order the cavity fields, and therefore the fields (B^t_i) in (67) and (C^t_i) in AMP (70), are independent Gaussian variables:

$$B^t_{i\to(ij)} \approx B^t_i \approx C^t_i \sim \mathcal{N}\big(\lambda Q(\boldsymbol{x}^*, \boldsymbol{b}^t) x^*_i, \lambda Q(\boldsymbol{b}^t, \boldsymbol{b}^t)\big),$$

so the AMP marginal means are asymptotically independent and distributed as

$$b^{t+1}_i \sim \eta\big(\lambda Q(\boldsymbol{x}^*, \boldsymbol{b}^t) x^*_i + \sqrt{\lambda Q(\boldsymbol{b}^t, \boldsymbol{b}^t)} Z_i\big)$$

with (Z_i) are i.i.d. standard Gaussian random variables. This implies

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n b^{t+1}_i x^*_i \equiv Q(\boldsymbol{x}^*, \boldsymbol{b}^{t+1}) &= \frac{1}{n}\sum_{i=1}^n \eta\big(\lambda Q(\boldsymbol{x}^*, \boldsymbol{b}^t) x^*_i + \sqrt{\lambda Q(\boldsymbol{b}^t, \boldsymbol{b}^t)} Z_i\big) x^*_i, \\ \frac{1}{n}\sum_{i=1}^n (b^{t+1}_i)^2 \equiv Q(\boldsymbol{b}^{t+1}, \boldsymbol{b}^{t+1}) &= \frac{1}{n}\sum_{i=1}^n \eta\big(\lambda Q(\boldsymbol{x}^*, \boldsymbol{b}^t) x^*_i + \sqrt{\lambda Q(\boldsymbol{b}^t, \boldsymbol{b}^t)} Z_i\big)^2.\end{aligned}$$

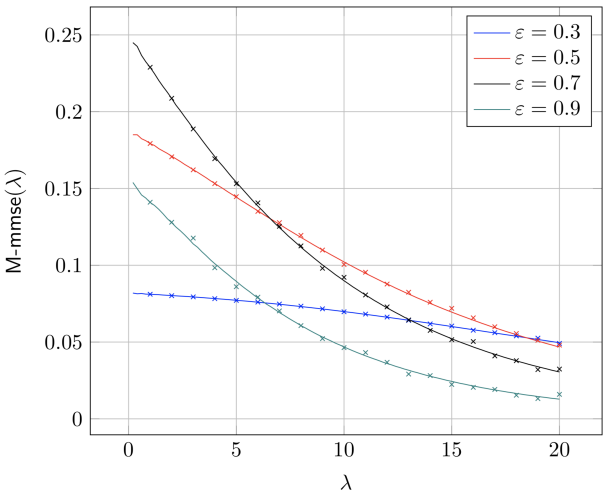


Figure 11: Figure from [18]. The solid curves are the state evolution prediction of the asymptotic spike-MSE of AMP $\lim_{t\to\infty}\lim_{n\to\infty}\mathrm{MSE}^t_{\mathrm{AMP}}$ for different SNR values λ. The crosses mark median MSE incurred by AMP in 100 Monte Carlo runs with $n = 2000$ for the spiked Wigner model with $X_i \sim \mathrm{Ber}(\epsilon)$. In this case there is no hard phase and AMP is always Bayes optimal.

By the independence assumptions and the law of large numbers the "magnetization" $Q(\boldsymbol{x}^*, \boldsymbol{b}^t)$ must concentrate onto its asymptotic expected value q_*^t and the overlap $Q(\boldsymbol{b}^t, \boldsymbol{b}^t)$ onto q^t as $n \to +\infty$ and with high probability. The limit overlap sequences are therefore solutions of the *state evolution recursions*. Let $X^* \sim P_X$, $Z \sim \mathcal{N}(0,1)$.

$$\text{State evolution: let } q_*^0 = q^0 = \epsilon \ll 1. \text{ For } t \in \mathbb{N}_{\geq 0}\text{:}$$
$$\begin{cases} q_*^{t+1} = \mathbb{E}\big[\eta(\lambda q_*^t X^* + \sqrt{\lambda q^t} Z) X^*\big], \\ q^{t+1} = \mathbb{E}\big[\eta(\lambda q_*^t X^* + \sqrt{\lambda\, q^t} Z)^2\big]. \end{cases}$$

The initialization $q_*^0 = q^0 = \epsilon$ is called *non-informative*. It corresponds to no a priori information about the signal. Having $\epsilon > 0$ allows to break a possible trivial fixed-point $q^t = 0$ that may appear due to the $\pm$ global symmetry in the problem (the data $\boldsymbol{y}$ is invariant by a global sign change of $\boldsymbol{x}^*$). This trivial fixed point may prevent state evolution to start. Running AMP on any real problem, this fixed point is always spontaneously broken by random fluctuations[20].

By the law of large numbers one can then track functions of the AMP fields. With high probability,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \phi(b_i^{t+1}, x_i^*) = \mathbb{E}\,\phi\big(\eta(\lambda q_*^t X^* + \sqrt{\lambda q^t} Z), X^*\big).$$

[20] There is now a more principled way to remove the possible trivial fixed point of state evolution, which corresponds to a spectral initialization of the AMP cavity means, see [19].

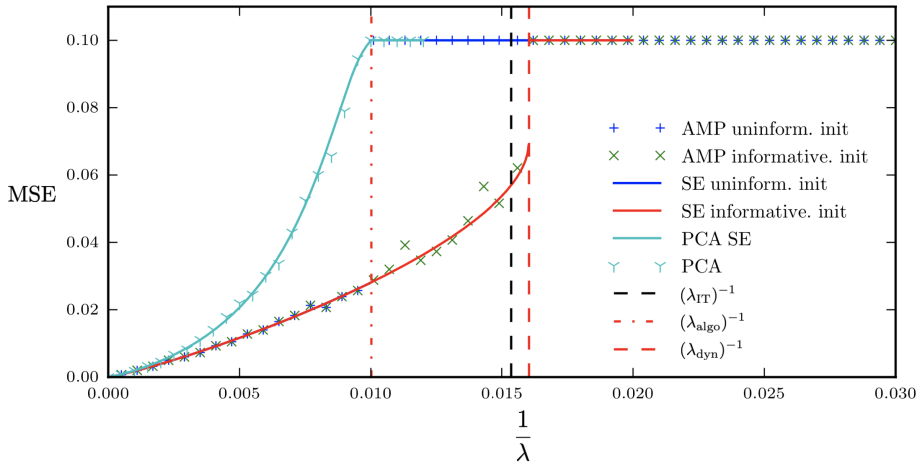


Figure 12: Figure from [8]. Comparison between the state evolution and the fixed point of the Low- RAMP algorithm, for the spiked Gauss–Bernoulli model of sparse PCA with rank one and sparsity $\rho = 0.1$. The phase transitions stemming from state evolution are $1/\lambda_{\text{algo}} = 0.01$, $1/\lambda_{\text{IT}} = 0.0153$, and $1/\lambda_{\text{dyn}} = 0.0161$. The points are the fixed points of the AMP algorithm run on one typical instance of the problem of size $n = 20000$. Blue pluses are the MSE reached from an uninformative initialization $q^0 = \epsilon \ll 1$. Green crosses are the MSE reached from the informative initialization $q^0 = \rho$.

When the denoiser is the MMSE denoiser (i.e. the assumed λ matches the true SNR in the model and the prior is known), i.e., in the planted SK model,

$$\eta(B) = \frac{\sum_{x=\pm 1} x \exp\{xB\}}{\sum_{x=\pm 1} \exp\{xB\}} = \tanh B \equiv \langle X \rangle_B,$$

then $\eta(\lambda x^* q_*^t + z\sqrt{\lambda q^t})$ is the posterior expectation for a scalar Gaussian channel $y = \lambda q_*^t x^* + \sqrt{\lambda q^t} z$ which is, not by coincidence, similar to the denoising model appearing in the replica symmetric potential. Therefore the Nishimori identity implies

$$\mathbb{E}\big[\eta(\lambda q_*^t X^* + \sqrt{\lambda q^t} Z) X^*\big] = \mathbb{E}\big[\eta(\lambda q_*^t X^* + \sqrt{\lambda q^t} Z)^2\big].$$

State evolution in the Bayesian optimal setting therefore simplifies to a recursion over a single parameter:

State evolution (Bayesian optimal setting): let $q^0 = \epsilon \ll 1$. For $t \in \mathbb{N}_{\geq 0}$:

$$q^{t+1} = \mathbb{E}\big[\eta(\lambda q^t X^* + \sqrt{\lambda q^t} Z) X^*\big].$$

This implies that almost surely

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} \phi(b_i^{t+1}, x_i^*) = \mathbb{E}\, \phi\big(\eta(\lambda q^t X^* + \sqrt{\lambda q^t} Z), X^*\big).$$

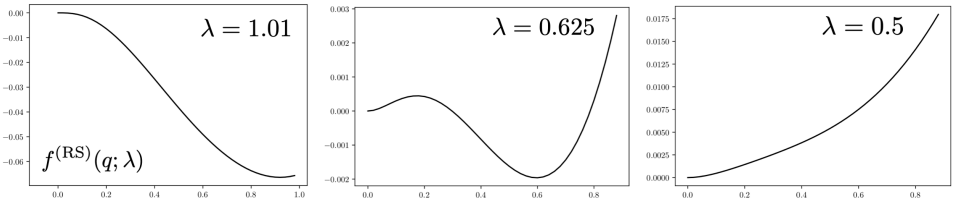


Figure 13: Figure from [7]. Plot of the free-energy single-letter potential $f^{(\mathrm{RS})}(q;\lambda)$ as a function of q for different SNR values λ. The porential $f^{(\mathrm{RS})}(q;\lambda)$ is the function minimized in (36). **Left (easy phase):** The SNR is above the algorithmic threshold $\lambda_{\mathrm{algo}} = 1$, so AMP yields allows optimal estimation with a MMSE equal to $\rho^2 - q_0^2$, where q_0 is the global minimizer of the potential. Optimal estimation being possible at low computational cost this regime is called *easy*. **Middle (hard phase):** Below the algorithmic threshold state evolution initialized from $q^{t=0} = 0$ will remain trapped in the local minimizer of the potential. AMP (as any known efficient algorithm) is therefore sub-optimal. This defines the *hard phase*. **Right (impossible phase):** Finally for very low SNR even the MMSE is high (q_0 is low), so inference is impossible for all algorithms (efficient or not).

Note that the fixed point equation associated with state evolution corresponds to the stationary condition (30) of the replica symmetric potential. This emphasizes a deep link between the potential function $i^{(\mathrm{RS})}$, related to the *static/thermodynamic* properties of the model, and the state evolution related to the *dynamic* properties of an algorithm.

All this analysis can be turned onto a theorem thanks to the conditionning technique developed by Bolthausen for the SK model, and then adapted by Bayati and Montanari for high dimensional regression with random features, followed by Rangan et al. for spiked matrix models.

MSE optimality of AMP. The spike-MSE of AMP is

$$\mathrm{MSE}^t_{\mathrm{AMP}} \equiv \frac{1}{n^2}\|\boldsymbol{x}^* \otimes \boldsymbol{x}^* - \boldsymbol{b}^t \otimes \boldsymbol{b}^t\|_{\mathrm{F}}^2 = \frac{1}{n^2}\big(\|\boldsymbol{x}^*\|_2^4 + \|\boldsymbol{b}^t\|_2^4 - 2(\boldsymbol{x}^* \cdot \boldsymbol{b}^t)^2\big).$$

Using state evolution and the independence of the parameters x_i^* we have almost surely (recall $\rho \equiv \mathbb{E}_{P_X}[(X^*)^2]$)

$$\lim_{n\to+\infty} \mathrm{MSE}^t_{\mathrm{AMP}} = \rho^2 + (q^t)^2 - 2(q^t_*)^2$$

which becomes, in the Bayesian optimal setting,

$$\lim_{n\to+\infty} \mathrm{MSE}^t_{\mathrm{AMP}} = \rho^2 - (q^t)^2.$$

Comparing this formula with the MMSE formula of Corollary 1, we can assert: *whenever the fixed point q^∞ of state evolution reached from the non-informative*

initialization matches the global minimizer $q_0(\lambda, \rho)$ of the replica symmetric potential, then AMP is optimal in the sense that its estimate $\boldsymbol{b}^\infty \otimes \boldsymbol{b}^\infty$ of the spike almost surely leads to the spike-MMSE in the limit $n \to +\infty$:

$$\text{If } q^0 = \epsilon \text{ and } q^\infty = q_0(\lambda, \rho)$$
$$\text{then } \lim_{t\to+\infty} \lim_{n\to\infty} \text{MSE}^t_{\text{AMP}} = \lim_{n\to\infty} \text{MMSE}(\boldsymbol{X}^* \otimes \boldsymbol{X}^* | \boldsymbol{Y}).$$

As a consequence we can simply predict the asymptotic performance of AMP by plotting the replica symmetric potential, see Fig 13. The high SNR region where the potential has a single minimum at high q corresponds to the *easy phase*: there exists a polynomial complexity algorithm, AMP, that matches the optimal performance. When a second local minimum exists at low q state evolution will find the corresponding fixed point and remain stuck there. AMP is therefore sub-optimal as any known algorithm. This defines the *hard phase*, associated with a computational-to-statistical gap. The entrance in the hard phase happens at the so-called *algorithmic (or dynamical) threshold*. It is a fundamental open question in average case complexity to know wether this phase is fundamentally hard for all sub-exponential complexity algorithms or not, and if AMP can be overcomed in this regime. Finally at SNR values lower than the *information theoretic threshold* even the MMSE is high, so that there exist no algorithm (efficient or not) able to perform good inference: there is simply not enough information in the data to extract the signal. This easy-hard-impossible phase diagram is very generic in high-dimensional inference problems.

When there is a hard phase/a computational-to-statistical gap, the phase transitions separating the different regimes are of the first-order type with discontinuous MSE's as in Fig. 8. When there is no hard phase AMP is always Bayes optimal. Its performances matches the MMSE which continuously degrades as the SNR gets smaller, see Fig 11.

A Proof of inequality (42)

Let us drop the indices in the bracket $\langle - \rangle_\epsilon$. We start by proving

$$-2\,\mathbb{E}\big\langle Q(\mathcal{L} - \mathbb{E}\langle\mathcal{L}\rangle)\big\rangle = \mathbb{E}\big\langle (Q - \mathbb{E}\langle Q\rangle)^2\big\rangle + \mathbb{E}\big\langle (Q - \langle Q\rangle)^2\big\rangle\,. \tag{71}$$

Using the definitions $Q \equiv \frac{1}{n}\boldsymbol{x}\cdot\boldsymbol{x}^*$ and (41) gives

$$\begin{aligned} 2\,\mathbb{E}\big\langle Q(\mathcal{L} - \mathbb{E}\langle\mathcal{L}\rangle)\big\rangle = \frac{1}{n^2}\Big\{ &\mathbb{E}\Big[X_i^*\langle X_i X_j^2\rangle - 2X_i^* X_j^* \langle X_i X_j\rangle - \frac{\tilde{Z}_j}{\sqrt{\epsilon}} X_i^* \langle X_i X_j\rangle\Big] \\ &- \mathbb{E}[X_i^*\langle X_i\rangle]\,\mathbb{E}\Big[\langle X_j^2\rangle - 2X_j^*\langle X_j\rangle - \frac{\tilde{Z}_j}{\sqrt{\epsilon}}\langle X_j\rangle\Big]\Big\}\,. \end{aligned} \tag{72}$$

The Gaussian integration by part formula $\mathbb{E}[\tilde{Z}_j g(\tilde{Z}_j)] = \mathbb{E}\, g'(\tilde{Z}_j)$ yields

$$\mathbb{E}\Big[\frac{\tilde{Z}_j}{\sqrt{\epsilon}} X_i^* \langle X_i X_j\rangle\Big] = \mathbb{E}[X_i^*\langle X_i X_j^2\rangle - X_i^*\langle X_i X_j\rangle\langle X_j\rangle]\,,$$

as well as

$$\mathbb{E}\Big[\frac{\tilde{Z}_j}{\sqrt{\epsilon}}\langle x_j\rangle\Big] = \mathbb{E}[\langle x_j^2\rangle - \langle x_j\rangle^2]$$

These simplify (72) to

$$\begin{aligned}2\,\mathbb{E}\big\langle Q(\mathcal{L}-\mathbb{E}\langle\mathcal{L}\rangle)\big\rangle = \frac{1}{n^2}\sum_{i,j=1}^{n}\big\{&\mathbb{E}[X_i\langle x_j\rangle\langle x_i x_j\rangle - 2X_iX_j\langle x_i x_j\rangle] \\ &- \mathbb{E}[X_i\langle x_i\rangle]\,\mathbb{E}[\langle x_j\rangle^2 - 2X_j\langle x_j\rangle]\big\}\,. \end{aligned} \tag{73}$$

The Nishimori identity implies $\mathbb{E}[\langle x_j\rangle^2] = \mathbb{E}[X_j\langle x_j\rangle]$ and

$$\mathbb{E}[X_i\langle x_j\rangle\langle x_i x_j\rangle] = \mathbb{E}[\langle x_i\rangle\langle x_j\rangle\langle x_i x_j\rangle] = \mathbb{E}[\langle x_i\rangle\langle x_j\rangle X_iX_j]\,.$$

These further simplify (73) to

$$\begin{aligned}&2\,\mathbb{E}\big\langle Q(\mathcal{L}-\mathbb{E}\langle\mathcal{L}\rangle)\big\rangle \\ &\quad= \frac{1}{n^2}\sum_{i,j=1}^{n}\big\{\mathbb{E}[\langle x_i\rangle\langle x_j\rangle X_iX_j - 2X_iX_j\langle x_i x_j\rangle] + \mathbb{E}[X_i\langle x_i\rangle]\,\mathbb{E}[X_j\langle x_j\rangle]\big\} \\ &\quad= \mathbb{E}[\langle Q\rangle^2] - 2\,\mathbb{E}\langle Q^2\rangle + \mathbb{E}[\langle Q\rangle]^2 \\ &\quad= -\big(\mathbb{E}\langle Q^2\rangle - \mathbb{E}[\langle Q\rangle]^2\big) - \big(\mathbb{E}\langle Q^2\rangle - \mathbb{E}[\langle Q\rangle^2]\big)\end{aligned}$$

which is (71).

The identity (71) just proven then implies

$$\begin{aligned}2\big|\mathbb{E}\big\langle Q(\mathcal{L}-\mathbb{E}\langle\mathcal{L}\rangle)\big\rangle\big| &= 2\big|\mathbb{E}\big\langle (Q-\mathbb{E}\langle Q\rangle)(\mathcal{L}-\mathbb{E}\langle\mathcal{L}\rangle)\big\rangle\big| \\ &\geq \mathbb{E}\big\langle (Q-\mathbb{E}\langle Q\rangle)^2\big\rangle\,.\end{aligned}$$

An application of the Cauchy-Schwarz inequality then gives

$$2\big\{\mathbb{E}\big\langle (Q-\mathbb{E}\langle Q\rangle)^2\big\rangle\,\mathbb{E}\big\langle(\mathcal{L}-\mathbb{E}\langle\mathcal{L}\rangle)^2\big\rangle\big\}^{1/2} \geq \mathbb{E}\big\langle (Q-\mathbb{E}\langle Q\rangle)^2\big\rangle\,.$$

This ends the proof of (42).

B AMP for the spiked Wigner model with generic prior

References

[1] M. Mézard and A. Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

[2] L. Zdeborová and F. Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[3] N. Macris and R. Urbanke. Statistical physics for communication and computer science. https://documents.epfl.ch/groups/i/ip/ipg/www/2014-2015/Statistical_Physics_for_Communications_and_Computer_Science/statphys-24-08-2015.pdf.

[4] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[5] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, page arXiv:1908.05355, Aug 2019.

[6] C. Shannon. A mathematical theory of communication. http://www.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

[7] L. Miolane. Phase transitions in spiked matrix estimation: information-theoretic analysis. *arXiv preprint arXiv:1806.04343*, 2018.

[8] T. Lesieur, F. Krzakala, and L. Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

[9] F. Guerra and F. Toninelli. The infinite volume limit in generalised mean field disordered models. *Markov Proc. Rel. Fields*, 9(2):195–2017, 2003.

[10] F. Guerra and F. Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230(1):71–79, 2002.

[11] F. Guerra. An introduction to mean field spin glass theory: methods and results. *Mathematical Statistical Physics*, 2005.

[12] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin-Glass Theory and Beyond*, volume 9 of *Lecture Notes in Physics*. World Scientific, Singapore, 1987.

[13] J. Barbier, C. L. Chan, and N. Macris. Concentration of multi-overlaps for random ferromagnetic spin models. *arXiv preprint arXiv:1901.06521*, 2019.

[14] M. Talagrand. *Mean Field Models for Spin Glasses. Volume I: Basic Examples*. Springer Verlag, 2011.

[15] M. Talagrand. *Mean Field Models for Spin Glasses. Volume II: Advanced Replica-Symmetry and Low Temperature*. Springer Verlag, 2011.

[16] J. Barbier and N. Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability Theory and Related Fields*, 174(3-4):1133–1185, 2019.

[17] T. Richardson and R. Urbanke. *Modern coding theory*. Cambridge university press, 2008.

[18] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse pca. In *IEEE Int. Symp. on Inf. Theory*, pages 2197–2201, 2014.

[19] A. Montanari and R. Venkataramanan. Estimation of low-rank matrices via approximate message passing. *arXiv preprint arXiv:1711.01682*, 2017.