Statistical mechanics of inference problems with correlated patterns : an incursion between the replica method and message passing algorithms

Alia Abbara

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à École Normale Supérieure

Statistical mechanics of inference problems with correlated patterns: an incursion between the replica method and message passing algorithms

Soutenue par
Alia Abbara
Le 10 novembre 2020

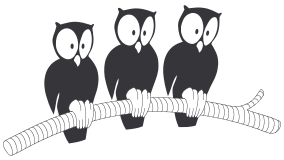
École doctorale n°564
Physique en Ile-de-France

Spécialité
Physique Théorique

Composition du jury :

Marc Mézard ENS	*Président du jury*
Rémi Gribonval ENS de Lyon / INRIA	*Examinateur*
Lenka Zdéborová EPFL	*Examinatrice*
Manfred Opper TU Berlin	*Rapporteur*
Ramji Venkataramanan University of Cambridge	*Rapporteur*
Florent Krzakala EPFL	*Directeur de thèse*

ENS | PSL



Seldon suddenly felt weary. It seemed as though this misinterpretation of his theory was constantly going to occur. Perhaps he should not have presented his paper. (...)

"My mathematical analysis implies that order must underlie everything, however disorderly it may appear to be, but it does not give any hint as to how this underlying order must be found. Consider — Twenty-five million worlds, each with its overall characteristics and culture, each being significantly different from all the rest, each containing a billion or more human beings who each have an individual mind, and all the worlds interacting in innumerable ways and combinations! However theoretically possible a psychohistoric analysis may be, it is not likely that it can be done in any practical sense."

Isaac Asimov, *Prelude to foundation*, 1988.

Main contributions

This thesis covers part of my work as a Ph.D. student between October 2017 and October 2020 at École Normale Supérieure in Paris, under the supervision of Florent Krzakala. Several sections and chapters and adapted from the following works, where one can also find codes for numerical implementation.

The first paper is mainly a statistical mechanics work, as it revisits a usual problem in a more delicate setting, and invokes a mixture of several methods and results from spin glass theory.

1. A. Abbara, Y. Kabashima, T. Obuchi and Y. Xu. Learning performance in inverse Ising problems with sparse teacher couplings. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (7), 073402.

This work is the fruit of a project led by Professor Yoshiyuki Kabashima (Tokyo University) and collaborators. We consider the inverse Ising problem, where one knows samples of spin configurations and tries to infer the couplings. We focus on the teacher-student scenario in the asymptotic limit, which was already analyzed in a previous work by Opper et. al., but now deal with the case of sparse teacher weights. This adds a major difficulty, since student weights are not sparse a priori. We design an ansatz for student weights, and combine the cavity and replica method to compute the error achieved by the pseudo likelihood estimator. We compare this result to simulations on several types of graphs.

The three following works divert from pure statistical mechanics and turn to inference problems, exploiting their links with message passing algorithms to characterize optimal reconstruction performance.

2. A. Abbara, A. Baker, F. Krzakala and L. Zdeborová. On the universality of noiseless linear estimation with respect to the measurement matrix. *Journal of Physics A: Mathematical and Theoretical*, 53 (16), 164001, 2020

This work looks into noiseless sparse linear regression, i.e. noiseless compressed sensing, a very simple setting that can be seen as a building block towards more complicated schemes. The initial motivation was to compare phase transitions in terms of reconstruction error in the teacher-student scenario, for different types of sensing matrices. We study two cases: the teacher-student scenario which exhibits for i.i.d. matrices a transition between a hard phase and an easy one, and the ℓ_1 recovery case with the well-known Donoho–Tanner transition. We use vector approximate message passing (VAMP) as solver, and find that many types of structured matrices (but not all) share the same transitions, thus shedding light on a universal behavior. This universality is easily clarified for rotationally invariant sensing matrices, but there is no clear explanation for other more complicated matrices.

3. C. Gerbelot, A. Abbara, and F. Krzakala. Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices. Volume 125 of *Proceedings of Machine Learning Research*, pages 1682-1713, 2020.

We consider penalized linear regression, where the regularization is a convex function. We provide a theoretical proof to characterize reconstruction performance in the teacher-student scenario, involving convex optimization and random matrix theory tools. The data matrix is taken to be rotationally invariant, which goes beyond the usual Gaussian data setting. We prescribe an oracle version of VAMP to solve the reconstruction problem. We build on its analytic tractability, provided by state evolution equations which specify statistical properties of the algorithm's estimates through its iterations, and show that Oracle VAMP converges if the regularization is strong enough. An analytic continuation argument is used to extend the result to regimes where VAMP is non-convergent. State evolution equations are shown to match, at their fixed point, the replica formula for the corresponding inference problem. Under reasonable assumptions, the proof shows that the heuristic replica prediction is rigorous.

4. C. Gerbelot, A. Abbara, and F. Krzakala. Asymptotic errors for teacher-student convex-generalized linear models (or: How to prove Kabashima's replica formula). *arXiv preprint arXiv:2006.06581*, 2020.

This paper extends the previous proof to generalized linear models with convex penalty, again with rotationally invariant data matrices. It demonstrates a somewhat difficult replica formula first derived by Kabashima in 2008, providing rigorous characterization for reconstruction performance. This time, we lean on an oracle version of multi-layer vector approximate message passing (MLVAMP) and its state evolution. To deal with non-linearity in the convergence proof, we need to refine our analysis. We resort to a dynamical system approach, where a non-linear system is recast into a linear one, and its convergence is determined through a linear matrix inequality.

The last paper in this list builds a bridge between statistical theory of learning and statistical physics, highlighting a connection between two quantities that are often encountered in each field.

5. A. Abbara, B. Aubin, F. Krzakala, and L. Zdeborová. Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning. Volume 107 of *Proceedings of Machine Learning Research*, pages 27–54, 2020.

In this work, we consider binary classification problems, in particular the Rademacher complexity which provides a worst-case scenario bound on the generalization gap, since it does not depend on the rule used to generate labels. We note that it is closely linked to the ground state energy of the corresponding problem and to the Gardner capacity, which are familiar quantities in statistical physics. The Rademacher complexity can thus be computed in some cases thanks to the replica method. We compare it to the best-case scenario, i.e. the teacher-student reconstruction error. We show how such a connection can benefit both communities, and deepen intuition about bounds on generalization performance.

Index of notations and abbreviations

AMP	Approximate message passing
ADMM	Alternating direction method of multipliers
BP	Belief propagation
DCT	Discrete cosine transform
EP	Expectation propagation
ER	Erdős–Rényi
i.i.d.	identically and independently distributed
IS	Interaction screening
LASSO	Least absolute shrinkage and selection operator
LMMSE	Least minimum mean squared error
MAP	Maximum a posteriori
MC	Monte Carlo
MCMC	Monte Carlo Markov chain
ML	Maximum likelihood
MLVAMP	Multi-layer vector approximate message passing
MSE	Mean squared error
PL	Pseudo likelihood
RF	Random features
RR	Regular random
RS	Replica-symmetric
1RSB	1-step replica symmetry breaking
2RSB	2-step replica symmetry breaking
RSS	Residual sum of squares
SE	State evolution
SK	Sherrington Kirkpatrick
SVD	Singular value decomposition
SVM	Support vector machine
TAP	Thouless Anderson Palmer
VAMP	Vector approximate message passing
VC	Vapnik Chervonenkis

$\delta(\cdot)$	Kronecker/Dirac delta
Tr	Trace operator
(i, j)	network edge or neighbor spins
$\sim$	sampled from
$\stackrel{\mathrm{d}}{=}$	equal in distribution
$\approx$	approximately equal
$\cong$	equal up to a constant factor
$\propto$	proportional to
$\mathcal{N}(0, \Delta)$	Gaussian distribution of mean 0 and variance Δ
p / P / $\mathbb{P}$	Probability

Contents

Introduction

A stroll through the statistical physics field

When talking to a non-specialist audience, it seems that statistical physics counts among the fields of physics that are less heard of. And yet, it is an essential tool to understand the behavior of large systems. Most people have in fact brushed upon it through thermodynamics lectures, as thermodynamic laws are a consequence of its fundamental representations. A sketchy way to present its scope of applications would be stating that statistical physics deals with systems that contain a very large number of elements. Say that you want to describe the properties of a gas of particles stored inside a box. Physics gives us an equation of motion for each particle, including their interactions, so we could try and solve all those equations and obtain a complete description of what is going on inside the box. Of course, this would be completely over the top: the number of particles involved is of the order of the number of Avogadro, i.e. 10^{23}, so there is no way that we could technically solve all those equations. In fact, obtaining a simulation of a few seconds is already a very hard task that requires a huge amount of computing power. What do we do then, if a simple box of gas renders us helpless? We need to turn to a statistical description of the particles, to move from a *microscopic* (i.e. at the scale of particle interaction) to a *macroscopic* description (around the size of the box). Besides, we are not actually interested in knowing the movement of every single particle. Instead, we would like to know some characteristics of the system as a whole: for instance, we usually describe a gas by its temperature and pressure, quantities which need to appear through our description of the system. The foundations to statistical mechanics were laid out in the second half of the 19^{th} century by Maxwell, Gibbs [63], and of course Boltzmann. In this introduction, we will give a mere glimpse of the discipline, simply to set the background that will be needed through the chapters.

The key to soothe down the serious headache of dealing with 10^{23} particles is to adopt a probabilistic description. Instead of studying every single element of the system, we will focus on the probability of the system being in a given configuration. A *configuration*, or a *microstate*, is defined by all the dynamic variables needed to specify the state of each particle. The ensemble of all configurations is the *phase space*. We also know the Hamiltonian $\mathcal{H}$ of the system, i.e. the sum of all kinetic and potential energies of the system. Of course, the state of the system depends on time, and many problems in statistical mechanics include a dynamical study. We might be interested in *observables* evaluated on the system, some of them being proportional to the system size i.e. *extensive*, while the other are *intensive*. Since the system has a very large size, it is reasonable in many cases to assume that taking the average with respect to observation time of an observable stabilizes: we say that the system is in its (Gibbs) *equilibrium state*, and most of our work will be framed in this context. Feynman described equilibrium as the state when *all the "fast" things have happened and all the "slow" things not* [51]. A fundamental assumption of statistical physics is that all configurations have the same probability after equilibrium is reached. One attempt to explain this is the *ergodicity hypothesis*: it basically states that each microstate has been visited enough times so as to construct locally a uniform measure in the phase space. Averaging any observable on a sufficiently long time would then be the same as doing an average over the phase space. However, ergodicity is very hard to prove and is a challenging topic in mathematics [175], and even if the ergodicity hypothesis holds, the time

needed to do a sufficient spanning of the phase space is incredibly large... Another way of justifying the uniform distribution over the space phase is that many observables actually have the same value for "most" microstates (except an exponentially smaller number of configurations), therefore many different measures on the phase space would still be an acceptable sampling of the equilibrium measure, and still yield the right result for the observable average. In this sense, the uniform measure is one among many measures that do the job.

We can now define the number of configurations (or the corresponding volume in the phase space) that have a given energy E as $W(E)$. It grows exponentially with the number of particles, hence it is convenient to use its logarithm. This is none other than the *entropy*[1] of the system, whose existence was postulated by Boltzmann [147], that reads $S(E) = k_B \log W(E)$ where k_B is the Boltzmann constant and has the dimension of an energy divided by a temperature ($k_B \simeq 1.38 \times 10^{-23}\mathrm{JK}^{-1}$). Entropy is an extensive function of the energy. It is often described as the disorder of the system, which can be a bit confusing. Entropy depends on the number of configurations that yield the system's energy. During a spontaneous reaction such as an exchange of heat between a warm cup of tea and cold air, due to the very large number of molecules, there are many more microstates where the energy is "spread out" between the drink and air molecules, than those where energy stays stored in the drink. Entropy thus increases, and the tea becomes lukewarm. Without going into detail, we remind the reader of the second law of thermodynamics: if we consider an isolated system, entropy can only increase (or stay constant for a reversible operation). While it is hard to develop an intuition about entropy, it carries deep meaning and shows up in several fields of physics[2].

From there we can also define the *temperature* T as $\frac{\partial S}{\partial E} \equiv \frac{1}{T}$. Note that those statistical quantities actually coincide with the thermodynamic ones and allow to properly recover the meanings we are used to. When two isolated systems are put in contact, they exchange energy, and equilibrium is reached when their temperatures are the same, as in Figure 1. Another usual

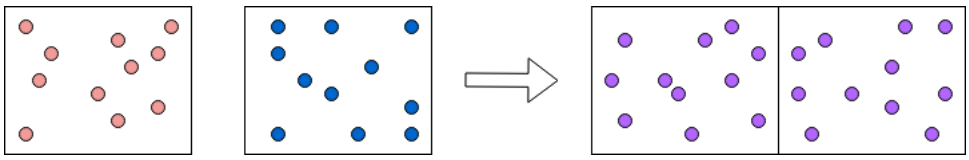


FIGURE 1: Left: two isolated systems with different temperatures, which are then put in contact. Right: after a long enough observation time, the reunion of the two systems reaches thermal equilibrium, as both their temperatures stabilize at the same value.

setting is having a system that can exchange energy with a much larger macroscopic system, often called a thermostat. The reunion of these two systems is isolated, hence has constant energy. The large system is so huge that it has an absolute temperature T, and imposes it to his little brother. The probability p_n of finding the small system in one of its configurations of energy E_n then reads $p_n = \frac{1}{Z}\exp(-\frac{E_n}{k_B T})$, where $Z = \sum_n \exp(-\frac{E_n}{k_B T}) = \sum_E W(E)\exp(-\frac{E}{k_B T})$ is called the *partition function*. We often replace the factor $\frac{1}{k_B T}$ by β. The average value of the small system energy is called *internal energy*. It reads $U = \frac{1}{Z}\sum_n E_n e^{-\beta E_n}$ and also verifies $U = -\frac{\partial \log Z}{\partial \beta}$. An elegant property of this internal energy is that its fluctuation depends on the

[1] Shannon's entropy, which was introduced in 1948, was named this way due to its similarities with Boltzmann's entropy. It is in fact one among many fruitful parallels that can be drawn between physics and information theory.

[2] A mesmerizing question that involves entropy is the so-called *arrow of time* first underlined in astrophysics [47]: why do we observe time flowing in only one direction, even though most classical and quantum laws of physics are time-symmetric? The second principle of thermodynamics does specify a preferred direction of evolution, since entropy can only increase: time can only go forward...

system size, and is of the order of $1/\sqrt{N}$. N being very large, internal energy is known with precision. A follow-up definition is the *free energy* $F \equiv U - TS$, which verifies $F = -\frac{1}{\beta} \log Z$. It corresponds to the energy that is actually available for the system to perform work at constant temperature, i.e. that does not get lost through heat, hence the word "free". At equilibrium, a system will minimize its free energy.

This formalism allows to explain *phase transitions* between different phases of matter, when one external parameter is modified. At the heart of these transitions lies the *competition between energy and entropy*. A famous example is the Ising model, which describes N spins lying on a network's vertices. Each spin can have value ± 1, and neighboring spins interact. The energy of one configuration is defined as $E = -J \sum_{(i,j)} S_i S_j - B \sum_i S_i$. J is a coupling constant that tends to align spins, and B is a magnetic field. Note that replacing constant J by random couplings $\{J_{ij}\}$ gives rise to interesting and complex behaviors, and studying them is a purpose of a whole field called *spin glass* physics. In our simple model, there are 2^N possible configurations, and each one has a probability $\frac{1}{Z} e^{-\beta E}$ with E its energy. At large temperature $\beta \to \infty$, the configurations with smaller energy dominate the partition function, and the system is in its *ground state*. At small temperature $\beta \to 0$, all configurations are equiprobable. The majority of microstates have average magnetization $M = \frac{1}{N} \sum_i S_i$ equal to zero. The system will be in one of these configurations and have the corresponding energy, such that its entropy is maximal. Starting from aligned spins for intermediate values of temperature, flipping some of them will increase energy but also increase entropy: it is not clear which configurations are preferred by the system. It turns out that for a dimension equal to or larger than 2, the Ising model displays a phase transition. Below a critical temperature T_c, magnetization is non-null (it is the *ferromagnetic* phase), but becomes equal to zero for beyond T_c (*paramagnetic* phase). The magnetization is a good tracker of the state of the system, and studying it allows to characterize the phase transition: it is called an *order parameter*[3].

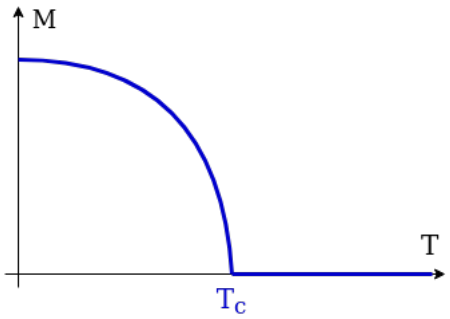


FIGURE 2: Shape of the average magnetization M as function of the temperature for the two-dimensional Ising model. At $T = T_c$, the system displays a phase transition as M becomes equal to zero.

Theoretical understanding of inference problems

In recent years, machine learning techniques that deal with huge datasets have improved in a spectacular way. In particular, *deep learning* [91] relies on artificial neural nets with several layers, and has achieved many successes in computer vision, speech recognition, image processing, recommendation systems, spanning health and industrial applications. Machine learning

[3] An intriguing property of continuous phase transitions is the existence of *critical exponents*, which provide a power law describing the behavior of order parameters at the transition. It seems that those exponents only depend on some features of the physical system, and can be classified in universality classes [90].

advances are at the heart of the so-called "fourth industrial revolution" [142]. State-of-the-art neural nets are made up of numerous layers of artificial neurons, which are simple units that perform a possibly non-linear operation based on the information received from their neighbors. Layers of neurons are connected through synaptic weights, defining a specific architecture. The resulting system is determined through a tremendous amount of parameters, which makes its behavior almost impossible to visualize.

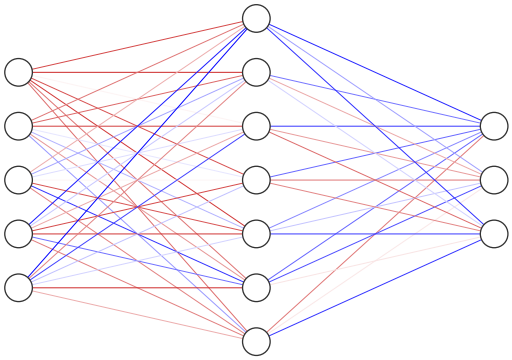


FIGURE 3: Example of fully-connected neural network. The input layer (left) is connected to one hidden layer (middle), itself connect to the output layer (right). The weights colors indicate their various strengths.

Yet, some network structures possess an important *representation power*: they become able to extract significant features from the data. For instance in *supervised learning*, a network is trained on examples of associated input-output, and needs to learn the underlying rule and gain the ability to *generalize* outside the training set. To train the network, we can use the widely-spread method of *stochastic gradient descent*, updating weights to go towards the minimum of a loss function that measures the network's mistakes on the training set. In this example, the goal is to minimize a function (which can be seen as an energy) with a very large number of parameters, as seen in the simple example of Figure 4. The mathematics involved [34] are demanding and far from the intuition that we might extract from low dimension settings. The advent of very complicated phenomena due to large dimension is common when we attempt to analyze neural networks, and is referred to as the *curse of dimensionality*.

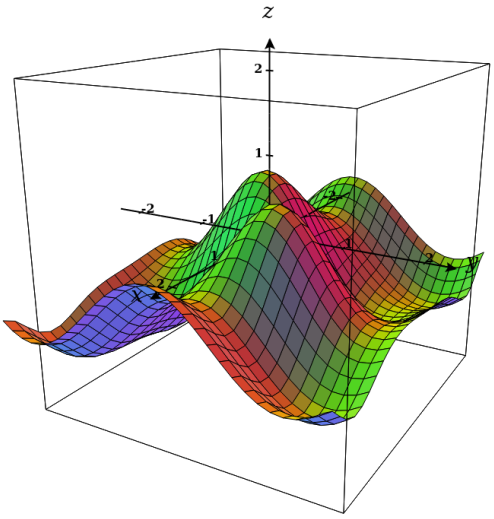


FIGURE 4: Example of energy landscape in 3 dimensions with several local minima.

While the high number of parameters tends to transform neural nets into black boxes, a silver lining is that statistical physics actually thrives in this large-size thermodynamic limit, that we will call the *asymptotic limit*. In fact, statistical physics has a long standing tradition in studying neural nets since the 80s, and more generally *inference* problems where we try to recover an underlying distribution through analysis of data. A seminal work by Hopfield [75] made an explicit connection between the field of spin glass physics and neural nets, and used analytical approach from disordered systems to solve the now-called *Hopfield model* with random input patterns. The *perceptron*, which corresponds to a one-layer neural network that performs binary classification, was studied by Gardner [58] for random input data.

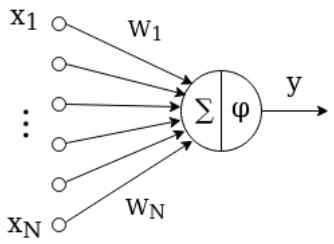


FIGURE 5: Single perceptron with inputs $x_1, ..., x_N$ connected through weights $W_1, ..., W_N$. The output is $y = \varphi(\sum_{i=1}^{N} W_i x_i)$, i.e. the weighted sum of inputs to which we apply a non-linearity φ, often a sigmoid.

A method borrowed from statistical physics, called the *replica method*, allowed to determine analytically how many patterns could be stored by N units, i.e. the *capacity* of the network. Note that those initial works describe simple inference problem or take strong assumptions, in particular assuming randomness properties of the data. Therefore, their results might not directly apply to the complicated, very structured real-world data used in cutting edge applications of machine learning. However, this is not an admission of defeat: starting with building blocks is essential to moving on to more complicated schemes. As the interest of statistical physics in machine learning has been revived, physical methods have also evolved, see [56] for a recent review of mean-field inference methods for neural networks. Statistical physics' current topics of research include – among many others – the generalization ability of networks, training algorithms through mean-field methods, clarifying the role of depth in networks by studying signal propagation through layers, or shedding light on dynamics of stochastic gradient descent in simple models.

A striking fact that calls for a statistical physics approach is the presence of phase transitions in some inference problems [88]. These transitions can be purely information-theoretical, or can describe an algorithmic behavior. A common scenario is the following: there are N variables, and we know M observations or samples, from which we want to extract information about the variables. We look at the regime where both M and N are large, in fact we take them both going to infinity, but such that their ratio $\alpha = M/N$ remains finite. A first question is to know the theoretical bounds: in what case do we have enough information to recover the variables? Say that our specific question can be tackled through an analytical approach. In some settings, we observe that the problem undergoes a sudden transition from solvable to unsolvable when varying some parameters, as shown in Figure 6. This can happen when the data become too sparse, or too noisy. The transition happens at a critical value α_{IT} (which might be unknown); which provides an information-theoretic threshold: inference is successful if the ratio of measurements to unknowns is larger than α_{IT}, and fails otherwise. Another question is computational efficiency: how can we design an algorithm that allows to solve the problem in polynomial time? It turns out that a phase transition may also occur in computational behavior, bringing to light another threshold α_C (which also might be unknown), such that $\alpha_C > \alpha_{\mathrm{IT}}$.

When $\alpha > \alpha_C$, the algorithm allows to solve the problem, but it fails to do so below α_C, even though inference is theoretically possible. The region between α_C and α_{IT} is called the *hard phase*: in this area, we would like to find algorithms that push the computational solvable to unsolvable transition more towards α_{IT}, or to find a lower bound for the transition threshold. Statistical physics can be informative for both questions. For some problems, it offers on one hand an analytical approach with proper order parameter, which explains the information-theoretical transition. On the other hand, it can provide insight about computational bounds by describing algorithmic behavior as a dynamic system. This will be illustrated with *message passing algorithms* [45] further down in our chapters. Note that several methods from physics provide heuristic non-rigorous results, that still carry deeper intuition. As in many examples through history, mathematics endeavors to prove these results, and build a solid theory around them; this duality will be illustrated in our manuscript.

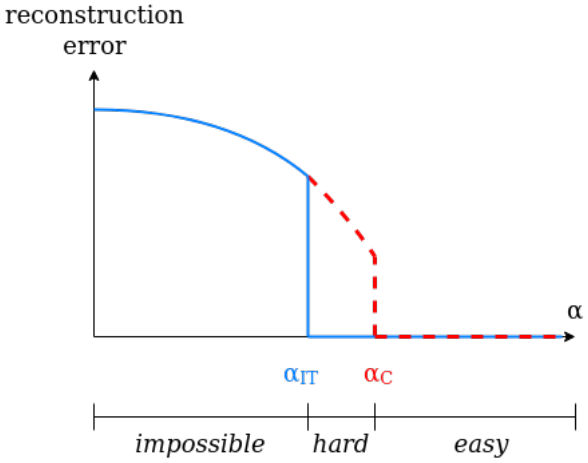


FIGURE 6: A typical phase diagram for an inference problem where the goal is to reconstruct a signal. The reconstruction error is plotted with respect to α, the measurements to signal ratio. The blue line represents optimal theoretical bound: at α_{IT}, the reconstruction error goes to zero, which can be explained as a physical phase transition. The dashed red line represents an algorithmic behavior, where the error goes to zero at α_C. The resulting diagram is divided in three parts: impossible, hard, and easy.

Research in machine learning is very active and channels efforts from many communities. Statistical learning theory provides rigorous bounds on worst-case scenarios, while statistical physics – relying on Bayesian inference – describes *typical* cases, which are the most probable. Optimization (convex, or non-convex) concentrates on extremizing energy or loss functions, and provides a mathematical analysis of corresponding algorithms. Information theory lays out a theoretical framework to measure the amount of information transmitted through neural networks... Keeping track of those plentiful directions of research is certainly hard, but also exciting as frontiers between disciplines become blurry, allowing for exchange of ideas, just like Shannon's entropy was named after Boltzmann's.

In this thesis, we will focus on a small portion of interaction between statistical physics and inference questions, in particular exploiting the physics toolbox developed in the context of spin glass theory, with the *replica* and *cavity* methods. We will build up on the relation between those methods and message passing algorithms, and go beyond the usual framework of Gaussian identically and independently distributed data, which was the first playground of statistical physics results. One concern will be to justify and prove heuristic physics result, exploiting optimization and random theory elements. Our work deals with simple linear networks, but also with more complex generalized linear models, i.e. multi-layer networks.

Organization of the thesis

In Chapter 1, we stay in the physics field and start by introducing spin glass physics, to illustrate the *replica method* through a well-known example, the Sherrington-Kirkpatrick model. We then apply this method on an inverse Ising problem with sparse teacher weights. In Chapter 2, we drift towards signal processing and machine learning, and address the inference problem of linear regression or *compressed sensing*. We introduce *message passing algorithms* and show that noiseless compressed sensing displays universal transitions for a large class of data matrices. Chapter 3 leans towards mathematics and optimization, as it turns to convex penalized linear regression, and provides a mathematical proof of the validity of the associated replica free energy, relying on a convergence proof of *Vector approximate message passing algorithm*. Chapter 4 is a generalization, since it shows the validity of the replica free energy for a generalized linear model with a convex penalty, relying on a dynamical systems inspired analysis of *Multi-layer vector approximate message passing*. Finally, Chapter 5 builds a bridge between statistical learning theory – which provides worst-case bounds on the generalization gap in classification problems – and statistical physics, by highlighting the link between the *Rademacher complexity* and the physical *ground state energy*. Each chapter will be followed by a summary of results and arising questions.

Chapter 1

The replica method and the inverse Ising problem with sparse weights

1.1 The replica method for spin glass models

1.1.1 Introduction to spin glass models

Spin glass models were originally introduced to describe some metal alloys such as copper-magnesium or gold-iron. They differ from standard ferromagnets where magnet spins align; or paramagnet where magnets spins are anti-aligned: in spin glasses, both behaviors compete to produce a non-regular pattern. The magnetic elements are modeled by N variables, that we call spins. Each spin is described by a scalar parameter S_i, and we define $\mathbf{S} = \{S_i\}_{i=1,\dots,N}$. Those spins interact with each others in a given way, and this interaction will influence how they behave together. Spin glasses are in fact very complex to study and show various phenomena of interest, both in-equilibrium and off-equilibrium. A nice and broad introduction to spin glasses in statistical physics can be found in [107].

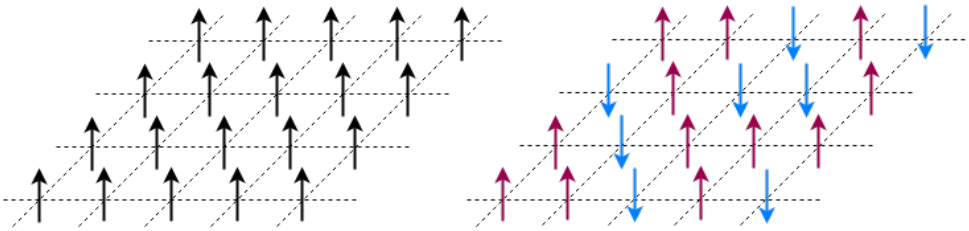


FIGURE 1.1: Schematic spin system where spins lie on a square lattice and are represented by up or down arrows. Left: ferromagnetic state where spins are aligned. Right: random spin glass state.

The interaction between spins are modeled by the coupling variables $\mathbf{J} = \{J_{ij}\}_{i,j=1,..,N}$ which are distributed according to some probability distribution $P(\mathbf{J})$. The spins are subjected to local fields $\mathbf{h} = \{h_i\}_{i=1,\dots,N}$. For a given set $\mathbf{J}$, the system is described by the Ising Hamiltonian, already mentioned in the introduction:

$$\mathcal{H}_{\mathbf{J}}(\{S_i\}) = -\sum_{i,j=1}^{N} J_{ij} S_i S_j - \sum_{i=1}^{N} h_i S_i. \tag{1.1}$$

A typical phenomenon in spin glass systems is *frustration*: depending on the sign of couplings J_{ij}, two spins will tend to be aligned or anti-aligned to minimize the energy of the system. However, a given spin might be under two opposing constraints from two of its neighbours, and it is not clear which configuration is the most energetically favorable, as shown in Fig. 1.2. From

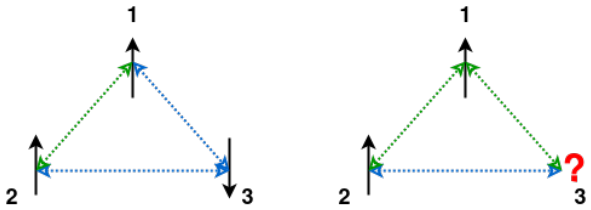


FIGURE 1.2: Representation of a system of 3 spins with possible frustration. The green dashed arrows symbolize positive couplings, which tend to align spins, while blue dashed arrows symbolize negative couplings which tend to anti-align spins. Left: the spin configurations satisfy the constraints imposed by the couplings. Right: Frustrated system. Spins 1 and 2 are aligned, but the positive coupling J_{13} pushes spin 3 to be up-oriented, while negative J_{23} pushes it to be down-oriented.

the Hamiltonian, we can define the partition function

$$\mathcal{Z}_{\mathbf{J}} = \sum_{\mathbf{S}} e^{-\beta \mathcal{H}_{\mathbf{J}}(\mathbf{S})} \tag{1.2}$$

with β an inverse temperature, and the average free energy

$$\Phi_{\mathbf{J}} = -\frac{1}{\beta N} \log \mathcal{Z}_{\mathbf{J}}. \tag{1.3}$$

Note that $\Phi_{\mathbf{J}}$ depends on a precise set of couplings: hence it differs for every spin glass sample! Therefore it is hard to draw general statements from it. Luckily, a standard thermodynamic argument shows that the free energy is self-averaging: it will reach the same value for any set of couplings $\mathbf{J}$ with a non-vanishing probability. Basically, a quantity is self-averaging when it concentrates around its average value and its variance goes to zero. The free energy is of order 1, and taking the average on the couplings distribution yields

$$\overline{\Phi_{\mathbf{J}}^2}^{\mathbf{J}} - (\overline{\Phi_{\mathbf{J}}}^{\mathbf{J}})^2 = \mathcal{O}\left(\frac{1}{N}\right). \tag{1.4}$$

We can thus appropriately define the average value of the free energy density:

$$\Phi = \overline{\Phi_{\mathbf{J}}}^{\mathbf{J}} = -\frac{1}{\beta N} \overline{\log \sum_{\mathbf{S}} e^{-\beta \mathcal{H}_{\mathbf{J}}(\mathbf{S})}}^{\mathbf{J}}. \tag{1.5}$$

At first glance, we see that computing the average free energy is no easy task: you need to first calculate $\mathcal{Z}_{\mathbf{J}}$ for any configuration of the couplings, then take the logarithm of this complicated integral, and finally average over the couplings. This is called the *quenched* average (the first step is done with a frozen choice of couplings $\mathbf{J}$). A simpler computation would be to perform the average over the couplings before taking the logarithm, i.e.

$$-\frac{1}{\beta N} \log \overline{\sum_{\mathbf{S}} e^{-\beta \mathcal{H}_{\mathbf{J}}(\mathbf{S})}}^{\mathbf{J}} \tag{1.6}$$

which is called the *annealed* average. In some fortunate cases, the quenched and annealed averages can actually be the same, but this is not true in general. There is no escaping: we need to find a way to evaluate the very difficult quenched free energy.

While strenuous, this task would be very informative: the free energy carries useful information about the state of the system, and will be written in terms of parameters that hold physical

meaning, and can help understand the configuration of the spins.

1.1.2 Goal of the replica method

The replica method aims at simplyfing the free energy computation, and relies on the following replica trick:

$$\log \mathcal{Z} = \lim_{n\to 0} \frac{\mathcal{Z}^n - 1}{n}. \tag{1.7}$$

The idea is essentially to replace the computation of $\log \mathcal{Z}$ by $\mathcal{Z}^n$. The dubious reader might object that this does not seem much simpler! In fact, in some settings it is indeed feasible to calculate $\mathcal{Z}^n$, as you will see in the following examples. We have not written the average free energy yet, so we add a few elements to reach

$$\Phi = -\frac{1}{\beta N} \lim_{N\to\infty} \lim_{n\to 0} \frac{\overline{\mathcal{Z}^n}^{\mathbf{J}} - 1}{n}. \tag{1.8}$$

To do this calculation properly, we should a priori make sure that the limit $n \to 0$ exists, and we should take it before the limit $N \to \infty$. However, in practice we cannot respect those constraints: we will have to deviate from this agenda to reach an analytic result.

1.1.3 Example of replica-symmetric calculation: the Sherrington-Kirkpatrick model

The Sherrington-Kirkpatrick model [148] is a celebrated spin glass model that displays infinite-range interactions. Its Hamiltonian reads

$$\mathcal{H}_{\mathbf{J}}(\mathbf{S}) = -\sum_{i<j} J_{ij} S_i S_j - \sum_i h S_i \tag{1.9}$$

with h a constant external field, and the couplings satisfy

$$\overline{J_{ij}} = 0 \qquad \overline{J_{ij}^2} = \frac{1}{N} \qquad J_{ij} = J_{ji}. \tag{1.10}$$

We will walk through the main steps of the calculations to provide an overview of the use of the replica method. We take the couplings as Gaussian variables, but note that the result only depends in fact on their first two moments.

Averaging the replicated partition function

We start off by writing $\overline{\mathcal{Z}^n}^{\mathbf{J}}$, that we call the replicated partition function. The idea is to introduce n replicas of the system, that we denote with the subscript a:

$$\overline{\mathcal{Z}^n}^{\mathbf{J}} = \sum_{\mathbf{J}} P(\mathbf{J}) \sum_{\mathbf{S_i^a}} \exp\left\{ \beta \sum_a \sum_{i<j} J_{ij} S_i^a S_j^a + \beta \sum_a \sum_i h S_i^a \right\}. \tag{1.11}$$

where

$$\sum_{\mathbf{S_i^a}} = \sum_{a=1}^{n} \sum_{i=1}^{N} \sum_{S_i^a = \pm 1}.$$

Taking the couplings to be Gaussian, we can compute the average

$$\overline{\mathcal{Z}^n}^{\mathbf{J}} = \int \prod_{i<j} \left(dJ_{ij} \frac{e^{-\frac{J_{ij}^2}{2N}}}{\sqrt{2\pi N}} \right) \sum_{\mathbf{S_i^a}} \exp\left\{ \beta \sum_a \sum_{i<j} J_{ij} S_i^a S_j^a + \beta \sum_a \sum_i h S_i^a \right\} \tag{1.12}$$

$$= \sum_{\mathbf{S_i^a}} \exp\left\{ \frac{\beta^2}{2N} \sum_{i<j} \left(\sum_{a=1}^{n} S_i^a S_j^a \right)^2 + \beta h \sum_i \sum_a S_i^a \right\}. \tag{1.13}$$

Note that

$$\sum_{i<j} \left(\sum_{a=1}^{n} S_i^a S_j^a \right)^2 = \sum_{i<j} \sum_a (S_i^a S_j^a)^2 + 2 \sum_{i<j} \sum_{a<b} S_i^a S_i^b S_j^a S_j^b \tag{1.14}$$

$$= \frac{nN(N-1)}{2} + \sum_{a<b} \left(\sum_i S_i^a S_j^b \right)^2 - \sum_{a<b} \sum_i (S_i^a S_j^b)^2 \tag{1.15}$$

$$= \frac{nN(N-1)}{2} - \frac{n(n-1)N}{2} + N^2 \sum_{a<b} \left(\sum_i \frac{S_i^a S_i^b}{N} \right)^2 \tag{1.16}$$

Keeping only terms of dominant order in N and replacing in the average free energy above, we reach

$$\overline{\mathcal{Z}^n}^{\mathbf{J}} = \sum_{\mathbf{S_i^a}} \exp\left\{ \beta^2 \frac{Nn}{4} + \beta^2 \frac{N}{2} \sum_{a<b} \left(\sum_i \frac{S_i^a S_i^b}{N} \right)^2 + \beta N h \sum_i \sum_a \frac{S_i^a}{N} \right\}. \tag{1.17}$$

This first step is already interesting: averaging out on the couplings yields an integral that depends only on the quantities $\sum_i \frac{S_i^a S_i^b}{N}$ and $\sum_i \frac{S_i^a}{N}$. The first one quantifies the overlap between the spins from replicas a and b, and the second the average magnetization of the spins for replica a: we begin seeing quantities that carry *physical meaning*. We would like to write the free energy as a function of those variables, instead of individual spins. Using the following identity[1]:

$$\exp\left\{ \frac{\beta^2 N}{2} \sum_i \left(\frac{S_i^a S_i^b}{N} \right)^2 \right\} = \int \mathrm{d}Q_{ab} \sqrt{\frac{\beta^2 N}{2\pi}} e^{-\frac{\beta^2 N}{2} Q_{ab}^2 + \beta^2 N \sum_i \frac{S_i^a S_i^b}{N} Q_{ab}}, \tag{1.18}$$

the replicated partition function becomes

$$\begin{aligned}
\overline{\mathcal{Z}^n}^{\mathbf{J}} &= \sum_{\mathbf{S_i^a}} \int \prod_{a<b} \left(\mathrm{d}Q_{ab} \sqrt{\frac{\beta^2 N}{2\pi}} \right) \\
&\quad \times \exp\left\{ -\sum_{a<b} \frac{\beta^2 N}{2} Q_{ab}^2 + \beta^2 N \sum_{a<b} Q_{ab} \sum_i \frac{S_i^a S_i^b}{N} + \beta^2 \frac{Nn}{4} + \beta N h \sum_i \sum_a \frac{S_i^a}{N} \right\} \\
&= \int \prod_{a<b} \left(\mathrm{d}Q_{ab} \sqrt{\frac{\beta^2 N}{2\pi}} \right) \exp\left\{ -N \left[\frac{\beta^2}{N} \sum_{a<b} Q_{ab}^2 - \frac{\beta^2 n}{4} \right] \right\} \left(\sum_{\mathbf{S^a}} \exp\left\{ \beta h \sum_a S^a + \beta^2 \sum_{a<b} Q_{ab} S^a S^b \right\} \right)^N \\
\overline{\mathcal{Z}^n}^{\mathbf{J}} &= \int \prod_{a<b} \left(\mathrm{d}Q_{ab} \sqrt{\frac{\beta^2 N}{2\pi}} \right) e^{-N \Phi_n[\mathbf{Q}]}
\end{aligned} \tag{1.19}$$

[1] (1.18) is simply a Gaussian integral where we have reverted the usual right-hand and left-hand terms, but is often referred to as the Hubbard-Stratonovitch transform (which might unnecessarily impress the reader), as it was first used by Hubbard and Stratonovitch in quantum mechanics. It contains the idea of converting a system of particles with two-body interactions, into a system of independent particles interacting with a field. This trick is widely used in replica calculations since it allows to "decouple" replicas.

where $\mathbf{Q} = \{Q_{ab}\}$ is a $n \times n$ matrix and

$$\Phi_n[\mathbf{Q}] = -\frac{n\beta^2}{4} + \frac{\beta^2}{2}\sum_{a<b} Q_{ab}^2 - \log \sum_{\mathbf{S^a}} \exp\left\{\beta h \sum_a S^a + \beta^2 \sum_{a<b} Q_{ab} S^a S^b\right\}. \tag{1.20}$$

Taking the limit $N \to \infty$ first

At this stage, we are tempted to use a saddle-point method on (1.19). This means that we first take the limit $N \to \infty$ before the limit $n \to 0$: a priori it might be wrong to do this. To progress further, we assume that we are allowed to reverse those limits, and get to

$$f = -\lim_{n\to 0} \frac{1}{\beta n} \min_{\mathbf{Q}} \Phi_n\left[\mathbf{Q}\right]. \tag{1.21}$$

This step is also tricky: we are looking for the $\frac{n(n-1)}{2}$ values inside matrix $\mathbf{Q}^*$ that minimize function Φ_n, but recall that n is bound to go to zero. We would then consider a negative number of parameters: this does not soundly make sense, and something could go wrong when taking $n \to 0$. Still, we feign ignorance and go on looking for $\mathbf{Q}^*$ satisfying

$$\forall a < b, \quad \left.\frac{\partial \Phi_n}{\partial Q_{ab}}\right|_{\mathbf{Q}^*} = 0 \tag{1.22}$$

$$\forall a < b, \qquad Q_{ab}^* = \frac{S^a S^b \exp\left\{\beta h \sum_a S^a + \beta^2 \sum_{a<b} Q_{ab}^* S^a S^b\right\}}{\sum_{\mathbf{S^a}} \exp\left\{\beta h \sum_a S^a + \beta^2 \sum_{a<b} Q_{ab}^* S^a S^b\right\}}. \tag{1.23}$$

The right-hand term can be seen as an average $\langle\cdot\rangle_{\mathbf{Q}}$ with respect to the partition function

$$\mathcal{Z}(\mathbf{Q}, \mathbf{S^a}) = \sum_{\mathbf{S^a}} e^{-\beta \mathcal{H}[\mathbf{Q}, \mathbf{S^a}]} \tag{1.24}$$

where the Boltzmann measure is on Hamiltonian

$$\mathcal{H}[\mathbf{Q}, \mathbf{S^a}] = -h \sum_a S^a - \beta \sum_{a<b} Q_{ab} S^a S^b. \tag{1.25}$$

Saddle-point equations thus read

$$\forall a < b, \quad Q_{ab}^* = \langle S^a S^b \rangle_{\mathbf{Q}^*}. \tag{1.26}$$

We are now focusing on the set of parameters $\mathbf{Q}^*$, which hold the meaning of average overlaps between two replicas. In particular S_{aa}^*, for $a = 1, ..., n$ is called the self-overlap of replica a. For our spin system, the self-overlap is always equal to one.

Choosing an ansatz

We now reach a key point of the calculation. Solving the saddle-point equations over the whole space of possible values of $\mathbf{Q}$ is too hard: we need to assume a reasonable parametrization of the matrix $\mathbf{Q}$. Remember that we introduced the replicas of the system as a formal trick to write the replicated partition function: they are all equivalent. In particular, the function $\Phi_n[\mathbf{Q}]$ should be invariant if we exchange lines or columns of the matrix. The simplest and natural ansatz is thus the replica-symmetric (RS) ansatz: $Q_{ab} = Q_{cd}$ for any $a \neq b$, and $c \neq d$. This parametrization states that

$$Q_{ab} = q_0 + (1 - q_0)\delta_{ab} \tag{1.27}$$

such that

$$Q_{aa} = 1 \text{ for } a = 1, ..., n \tag{1.28}$$
$$Q_{ab} = q_0 \text{ for } a \neq b \tag{1.29}$$

where q_0 is called the *order parameter*. Let us plug this into:

$$-\frac{1}{\beta n}\Phi(q) = -\frac{1}{\beta n}\left\{-\frac{n\beta^2}{4} + \frac{\beta^2}{4}\sum_{a<b} q_0^2 - \log\left[\sum_{\mathbf{S^a}} e^{\beta^2 q_0 \sum_{a<b} S^a S^b + \beta h \sum_a S^a}\right]\right\}. \tag{1.30}$$

Using

$$e^{\beta^2 q_0 \sum_{a<b} S^a S^b} = e^{\beta^2 q (\sum_a S^a)^2 - \beta^2 q_0 \sum_a (S^a)^2} \tag{1.31}$$
$$= e^{-n\beta^2 q_0} \int \mathrm{D}z e^{\beta\sqrt{q_0}\sum_a S^a} \tag{1.32}$$

with $\mathrm{D}z = \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$, we can decouple the sum on the replicas, reaching

$$-\frac{1}{\beta n}\Phi(q) = -\frac{1}{\beta n}\left\{-\frac{n\beta^2}{4} + \frac{n(n-1)\beta^2}{4} q_0^2 - \log\left[\int \mathrm{D}z (2\cosh(\beta\sqrt{q_0}z + \beta h))^n\right]\right\}. \tag{1.33}$$

Taking the limit $n \to 0$

Up till now, we were considering an integer number of replicas. To take the limit $n \to 0$, we assume that there is an analytic continuation of the value of $-\frac{1}{\beta n}\Phi(q)$ for real values of n, which yields

$$\Phi = -\frac{\beta}{4}(1 - q_0)^2 - \int \mathrm{D}z \log[2\cosh(\beta\sqrt{q_0}z + \beta h)], \tag{1.34}$$

and the corresponding saddle-point equation is

$$q_0 = \int Dz \tanh^2(\beta\sqrt{q_0}z + \beta h). \tag{1.35}$$

Let us take a step back and check where we are: we have computed the average free energy, as a function of one parameter q_0. The state of our system is described by the minimum of this free energy: we only need to minimize it with respect to q_0. To reach this stage, we have done several moves that could possibly lead to a mistake: we have reverted limits, chosen a particular ansatz, and assumed an analytic continuation of free energy for $n \to 0$.

Solving (1.35) for $h = 0$ shows that $q_0 = 0$ is the only solution for $\beta < 1$, but another non-trivial solution exists for $\beta > 1$. Since β plays the role of an inverse temperature, it means that the system undergoes a phase transition between a regime of high-temperature, where it has a paramagnetic phase, and a regime of low temperature, the spin glass phase. At $h \neq 0$, there is no phase transition.

We could stop here and be happy with our result, but since we have done several fishy steps, we need to be extra careful and double-check it. In fact, something does go wrong with our calculation, when we take the limit $n \to 0$. We already see that taking $n \to 0$ makes Φ_n a function of a negative number of parameters, which does not mean much, but we need some amount of faith – and some amount of craziness – to go through replica calculations. To evaluate the correctness of our solution we can probe the stability of the fixed point, by computing the

eigenvalues of the Hessian matrix

$$\left(\frac{\partial^2 \Phi_n[\mathbf{Q}]}{\partial Q_{cd} Q_{ef}}\right)_{\{c,d\},\{e,f\}} \in \mathbb{R}^{\frac{n(n-1)}{2} \times \frac{n(n-1)}{2}}$$

which is well defined for integer values of n, and then taking $n \to 0$ to see how it affects them. If matrix $\mathbf{Q}$ is a minimum of Φ_n, the eigenvalues of the Hessian matrix must all be strictly positive at this point. However, a detailed analysis in the replica-symmetric ansatz [39] provides all eigenvalues of the Hessian matrix, and shows that for any value of h, one of them becomes negative as $n \to 0$ if the temperature is sufficiently low. Therefore our solution is not valid anymore as the fixed-point becomes unstable and is therefore not a global minimum, and we need to pick a different ansatz for low temperature regime. Another way of seeing that something went wrong is to compute the entropy of the system. The entropy is positive by definition, being the logarithm of the number of configurations, however taking the RS ansatz we find a negative entropy at zero temperature. To understand how we can choose an appropriate ansatz for the matrix $\mathbf{Q}$ and what it would mean, we will have to dig deeper into the physical meaning of overlap parameters.

1.1.4 An overview of replica symmetry breaking

Pure states and ergodicity breaking

Until now, we simply went with the flow of the calculation, but let us see what statistical physics tell us about the behavior of this disordered spin system. By minimizing the free energy, we are looking for the Gibbs equilibrium state reached by the system. In fact, this equilibrium state can be seen as a mixture of several *pure states*, which themselves cannot be split: they form a basis of all states. In a pure state, correlations between spins need to vanish as the distance between them goes to infinity, which is a priori not the case for the equilibrum state. At low temperature, the system undergoes a breaking of ergodicity: only a subpart of the space of configurations is explored. In that case, several pure states coexist, and we can rewrite the average $\langle\cdot\rangle$ on the Hamiltonian as:

$$\langle\cdot\rangle = \sum_\gamma w_\gamma \langle\cdot\rangle_\gamma \tag{1.36}$$

where γ denotes pure states, and the average $\langle\cdot\rangle_\gamma$ is an average over the Boltzmann measure for all configurations belonging to the pure state γ:

$$\langle\cdot\rangle_\gamma = \frac{1}{\mathcal{Z}_\gamma} \sum_{\mathbf{S}\in\gamma} \cdot \; e^{-\beta \mathcal{H}_\mathbf{J}(\mathbf{S})} \tag{1.37}$$

$$\mathcal{Z}_\gamma = \sum_{\mathbf{S}\in\gamma} e^{-\beta \mathcal{H}_\mathbf{J}(\mathbf{S})}. \tag{1.38}$$

Note that those averages depend on the disordered couplings $\mathbf{J}$.

Physical parameters

- Edwards-Anderson parameter

We would like to define a nice parameter to describe a given configuration of the spins. For magnets, the magnetization $\frac{1}{N}\sum_{i=1}^N \langle S_i\rangle$ comes in handy, but since the couplings $\mathbf{J}$ are unbiased, the spins end up frozen in all directions, and averaging the magnetization on the disorder would

simply give 0. Therefore, we prefer using the Edwards-Anderson parameter:

$$q_{\mathrm{EA}} = \frac{1}{N}\sum_{i=1}^{N} \overline{\langle S_i\rangle^2}^{\mathbf{J}} \tag{1.39}$$

which is non-zero if the magnetizations are locally non-zero.

- Overlap between two spin configurations

Another parameter we are interested in the overlap, which allows to compare two spin configurations, and which naturally shows in the replica calculation. For two spin configurations $\mathbf{S}, \mathbf{S}'$, the overlap reads

$$q_{SS'} = \frac{1}{N}\sum_{i=1}^{N} S_i S_i'. \tag{1.40}$$

- Overlap between two states of the system

We can also define an overlap between states γ and η, which themselves contain several configurations, as

$$q_{\gamma\eta} = \frac{1}{N}\sum_{i=1}^{N} \langle S_i\rangle_\gamma \langle S_i\rangle_\eta \tag{1.41}$$

which unfolds as

$$q_{\gamma\eta} = \sum_{\mathbf{S}\in\gamma}\sum_{\mathbf{S}'\in\eta} \frac{1}{\mathcal{Z}_\gamma \mathcal{Z}_\eta} e^{-\beta\mathcal{H}_{\mathbf{J}}(\mathbf{S})} e^{-\beta\mathcal{H}_{\mathbf{J}}(\mathbf{S}')} q_{SS'}. \tag{1.42}$$

Hence the overlap between two states of the system is the same as taking overlaps between all configurations between those two states, and summing on their statistical weight. For a given system, we would like to know the overlaps between pure states: to do so we define their probability distribution $P_{\mathbf{J}}(q)$. Interestingly, we can rewrite q_{EA} as a sum on pure states γ, η:

$$q_{\mathrm{EA}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\gamma,\eta} \overline{w_\gamma w_\nu \langle S_i\rangle_\gamma \langle S_i\rangle_\eta}^{\mathbf{J}} = \sum_{\gamma,\eta} w_\gamma w_\eta q_{\eta\gamma} = \int \mathrm{d}q\; q\; \overline{P_{\mathbf{J}}(q)}^{\mathbf{J}}. \tag{1.43}$$

Physical meaning of the overlap matrix Q

Besides, note that q_{EA} can be written in the same fashion as a replica calculation, taking two replicas c and d of the system, and $\mathbf{S^c}$, $\mathbf{S^d}$ both being the set of all configurations:

$$q_{\mathrm{EA}} = \overline{\frac{1}{\mathcal{Z}_{\mathbf{J}}^2}\sum_{\mathbf{S^c}}\sum_{\mathbf{S^d}} e^{-\beta\mathcal{H}_{\mathbf{J}}(\mathbf{S^c})} e^{-\beta\mathcal{H}_{\mathbf{J}}(\{S^d\})} q_{S^cS^d}}^{\mathbf{J}} \tag{1.44}$$

$$= \lim_{n\to 0}\frac{1}{N}\sum_{a=1}^{n}\sum_{\mathbf{S^a}} \overline{e^{-\beta\sum_{a=1}^{n}\mathcal{H}_{\mathbf{J}}(\mathbf{S^a})}\sum_{i=1}^{N} S_i^c S_i^d}^{\mathbf{J}} \tag{1.45}$$

$$= \lim_{n\to 0} \overline{\langle S^c S^d\rangle}^{\mathbf{J}}. \tag{1.46}$$

In the replica-symmetric ansatz, this is $\overline{\langle S^c S^d\rangle}^{\mathbf{J}} = Q_{cd}^{(\mathrm{RS})}$, the value of Q_{cd} at the saddle-point solving (1.35), in the replica-symmetric case. Note that (c, d) could be any pair of different indices. Hence this relation makes sense in the RS ansatz, but becomes problematic in a non-symmetric ansatz where the overlaps could have different values. In fact, in that case we need to broaden our calculation to include more than one saddle-point. Indeed, if the matrix $\mathbf{Q}$ is not

invariant under permutations of the replicas, then any permutation provides a different saddle-point with the same free energy. We would hence need to take into account all those equivalent saddle-points, and sum on all of them. Hence the general case reads:

$$q_{\mathrm{EA}} = \lim_{n\to 0} \frac{2}{n(n-1)} \sum_{a<b} Q_{ab}. \tag{1.47}$$

Matching this with equation (1.41), we get

$$P(q) \equiv \overline{P_{\mathbf{J}}(q)}^{\mathbf{J}} = \lim_{n\to 0} \frac{2}{n(n-1)} \sum_{a<b} \delta(q - Q_{ab}). \tag{1.48}$$

Finally, we see that the fraction of elements equal to q in the matrix $\mathbf{Q}$ gives the average probability that two pure states of the system have overlap q. Hence, the parameters have a clear physical meaning, and correspond to possible overlaps among pure states. In the RS ansatz, we only have a single possible overlap $q_0 = Q_{ab}$ for all $a \neq b$, which means that only one pure state exists, and the overlap q_0 is the self-overlap of this single pure state. At low temperature, as ergodicity breaking occurs, several pure states appear and hence give rise to several overlap values. Bearing this in mind, we can try to formulate another ansatz for the matrix $\mathbf{Q}$.

1-step replica symmetry breaking

Now that we see the link between the overlap matrix $\mathbf{Q}$ and overlaps between pure states, we can turn to a more sophisticated ansatz. The puzzle is still very hard to solve: we have no idea how many pure states there are, what their self-overlap is, how many configurations they include... Once again, we will try to design a somewhat naive ansatz and see how well it performs (notably by comparing it to simulations), hoping that it will be well enough to encapsule the physical reality. The simplest – although not that simple – ansatz is the first step of replica symmetry breaking (1RSB), which was introduced by Parisi [123, 124]. Say that among the n replicas, there exists n/m groups of m replicas each. Of course, n needs to be a multiple of m for this to make sense. Let us further assume that the configurations within one state all have the same overlap q_1, and that configurations within two different states all have the same overlap q_0. Basically, the phase space is divided into clusters, each cluster being a state containing a number of configurations, and all clusters having the same internal overlap and mutual overlap, as shown in Fig. 1.3. The replicas reproduce this structure. For instance, if we take $n = 9$ and $m = 3$, the 1RSB matrix $\mathbf{Q}$ reads:

$$\mathbf{Q} = \begin{pmatrix} 1 & q_1 & q_1 & & & & & & \\ q_1 & 1 & q_1 & & \cdots & & & q_0 & \\ q_1 & q_1 & 1 & & & & & & \\ & & & 1 & q_1 & q_1 & & & \\ & \cdots & & q_1 & 1 & q_1 & & \cdots & \\ & & & q_1 & q_1 & 1 & & & \\ & & & & & & 1 & q_1 & q_1 \\ & q_0 & & & \cdots & & q_1 & 1 & q_1 \\ & & & & & & q_1 & q_1 & 1 \end{pmatrix}. \tag{1.49}$$

Very well, but we still need to figure out what will happen when we take the mysterious $n \to 0$ limit, since $m < n$. According to (1.48), we have

$$P(q) = \frac{n-m}{n-1}\delta(q - q_0) + \frac{m-1}{n-1}\delta(q - q_1) \tag{1.50}$$

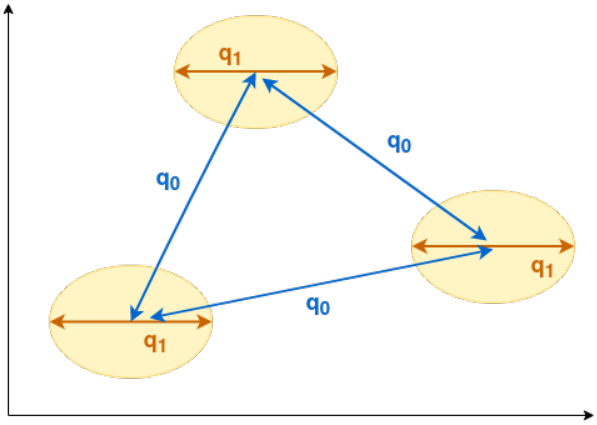


FIGURE 1.3: Schematic representation of the phase space for 1-step replica symmetry breaking. The axes are some 2D projection of the configuration phase space. The yellow blobs represent configuration clusters. The self overlap q_1 is the same within each cluster, while q_0 is the mutual overlap between any pair of clusters.

and taking the limit $n \to 0$, we get to

$$P(q) = m\delta(q - q_0) + (1 - m)\delta(q - q_1). \tag{1.51}$$

Since we are striving to keep some meaning into the quantities we are manipulating, we would like to keep $P(q)$ a probability, hence smaller than 1, which implies that $0 \leq m \leq 1$. Besides, we would like $0 \leq q_0 \leq q_1$ since two configurations belonging to the same state should be "closer". Our recipe is now complete: we need to plug this ansatz inside the replica calculation, to write the free energy as a function of q_0, q_1, m; and to extremize it to reach a set of saddle-point equations. There are still some difficulties along the way, but we will not detail them here since we were mainly interested in explaining the idea between replica calculations. However, it is worth noting that the result from the 1RSB ansatz is very close to numerical experiments, and much more satisfactory than the RS wrong result in the low temperature spin glass phase.

Beyond 1RSB

While the 1RSB ansatz performs well for the SK model, it is clear that is still a simplified way to model the phase space. We could keep repeating the same plan by splitting clusters into sub-clusters, and adding a new overlap q_2: this is two-step replica symmetry breaking. in practice, the computation keeps getting harder, and it is not necessarily useful to increase complexity since the 2RSB scheme does not always show a significant difference from the 1RSB result.

Although somewhat hard to use at first try, the replica method is a very powerful tool to reach a heuristic solution for many problems. While it clearly lacks mathematical rigor, a huge effort was made in order to prove replica results in several settings, and we will come back to this later in this thesis. The take-away message is that the replica method manages to capture some very deep and fundamental properties of disordered systems, and to incorporate them accurately in its structure: it is no wonder that it has been used so consistently in the past 40 years, and is still a stepping stone to formulate many questions, without completely unveiling its secrets. Let us stray away from the replica method for some time: it will show up again in a few paragraphs as a way of tackling some inference problems.

1.2 An introduction on statistical inference

In *A study in scarlet*, by Sir Arthur Conan Doyle, Sherlock Holmes makes his first appearance as Dr Watson reads one of his articles. The latter states that "From a drop of water, a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard one of the other. (...) By a man's finger nails, by his coat-sleeve, by his boot, by his trouser knees, by the callosities of his forefinger and thumb, by his expression, by his shirt cuffs – by each of these things a man's calling is plainly revealed. That all united should fail to enlighten the competent enquirer in any case is almost inconceivable." Watson does not buy it, and indignantly comments "What ineffable twaddle!" We may want to agree: Holmes' deductions often seem phenomenal and somewhat far-fetched. Holmes then argues that this is no empty gibberish, since detectives often come to him, "lay all the evidence before [him],and [he is] generally able, by the help of [his] knowledge of the history of crime, to set them straight." The reason why Holmes' findings might not always convincing is the little amount of information that he has. The more data he has on a person, the more believable his deductions become. Nevertheless, he points out that he is able to draw conclusions thanks to his "special knowledge". What makes Holmes such a gifted investigator is his shrewd use of inference.

Statistical inference describes the process of deducing information about a distribution of underlying data, from the knowledge of observations. Let us see how to formulate this in mathematical terms, taking up the approach of Bayesian inference. We consider a set of variables $\mathbf{x} = \{x_i\}_{i=1,\dots,N}$ and measurements $\mathbf{y} = \{y_\mu\}_{\mu=1,\dots,M}$ that contain some information on the variables.

- We assume that the data is distributed according to $P(\mathbf{x})$, the *prior distribution*, that we may only know partially. Note that the data does not effectively have to be a random variable, we merely use the probability notation to describe a belief about the values of $\mathbf{x}$.
- The way in which the observations are generated or derive from the data is $P(\mathbf{y}|\mathbf{x})$, the *likelihood distribution*.
- The conditional probability of having data $\mathbf{x}$ given observations $\mathbf{y}$ is $P(\mathbf{x}|\mathbf{y})$, the *posterior distribution*.

They are linked through the well-known Bayes formula

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}. \tag{1.52}$$

For Holmes and Watson, the variables could be information about the suspect's life and whereabouts, while the observations are all the clues that are so brilliantly interpreted by Sherlock. The "special knowledge" he mentioned lies in distributions $P(\mathbf{x})$ and $P(\mathbf{y}|\mathbf{x})$, and his deductions are a proxy for $P(\mathbf{x}|\mathbf{y})$. Apart from profiling a culprit, inference problems show up in many fields: machine learning, social science, information theory, biology, signal processing... They can be very complex and challenging, fuel a large amount of research, and applications are everywhere. Statistical inference is particularly exciting in our modern era of big data: for many problems we can obtain a huge amount of information, thousands, millions of measurements. How accurate would Sherlock Holmes become if he also had access to the social media accounts, the GPS localization, and the message history of the suspect! Even Watson would not do too bad... We will hence focus on the large-dimensional thermodynamic limit: we will consider $M, N \to \infty$ but will often keep their ratio $\alpha \equiv M/N$ of order 1.

Teacher-student scenario

Within the Bayesian inference set-up, some problems can be written in the teacher-student scenario.

i) We assume that a teacher generates the variables $\mathbf{x_0}$ from a probability distribution $P_T(\mathbf{x_0})$, then the measurements are obtained through a likelihood probability $P_T(\mathbf{y}|\mathbf{x_0})$. The original variables $\mathbf{x_0}$ are called the *ground truth*.

ii) A student then starts from the observations $\mathbf{y}$, and has some knowledge through distributions $P_S(\mathbf{x})$ and $P_S(\mathbf{y}|\mathbf{x})$, and would like to recover $\mathbf{x_0}$.

Estimators

From here comes an important question: how do we measure the performance of an estimator $\hat{\mathbf{x}}$ that aims at recovering $\mathbf{x_0}$? There are several options, but we will often use the following quantities:

- The *maximum a posteriori (MAP)* estimator verifies

$$\hat{\mathbf{x}}^{\mathrm{MAP}} = \arg\max_{\mathbf{x}} P_S(\mathbf{x}|\mathbf{y}) \tag{1.53}$$

- We could also try to minimize the *mean squared error (MSE)* between the signal and the ground truth defined as

$$\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x_0}) = \frac{1}{N}\mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x_0}\|_2^2] \tag{1.54}$$

  where the average is on the prior distribution $P_T(\mathbf{x_0})$. When the ground truth is not known, we need to assume that the signal $\mathbf{x}$ is sampled from the distribution $P_S(\mathbf{x}|\mathbf{y})$, on which we average to get $\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x})$.

Bayes optimal versus mismatched case

The distributions used by the student may or may not coincide with those of the teacher. Clearly, it would be ideal if the student knows the ground truth prior distribution and likelihood since it would give him more accurate information about the data, but it is often not the case. Thus we distinguish the two settings:

- If $P_S(\mathbf{x}) = P_T(\mathbf{x})$ and $P_S(\mathbf{y}|\mathbf{x}) = P_T(\mathbf{y}|\mathbf{x})$, we are in the *Bayes optimal* case.
- If $P_S(\mathbf{x}) \neq P_T(\mathbf{x})$ and/or $P_S(\mathbf{y}|\mathbf{x}) \neq P_T(\mathbf{y}|\mathbf{x})$, we are in the *mismatched* case.

Nishimori identity Say that we are considering an inference problem in a Bayes optimal teacher-student setting. $\mathbf{x_0}$ is again the ground truth, $\mathbf{y}$ the measurement given by the teacher, and let us take $\mathbf{x^a}, \mathbf{x^b}, \mathbf{x^c}$ three configurations sampled independently from the posterior distribution $P(\mathbf{x}|\mathbf{y})$. Take f any function of two configurations. We can then write

$$\mathbb{E}[f(\mathbf{x^a}, \mathbf{x^b})] = \int \mathrm{d}\mathbf{x^a}\,\mathrm{d}\mathbf{x^b}\,\mathrm{d}\mathbf{y} P(\mathbf{x^a}|\mathbf{y})P(\mathbf{x^b}|\mathbf{y})P(\mathbf{y})f(\mathbf{x^a}, \mathbf{x^b}) \tag{1.55}$$

$$\mathbb{E}[f(\mathbf{x}^*, \mathbf{x^c})] = \int \mathrm{d}\mathbf{x_0}\,\mathrm{d}\mathbf{x^c}\,\mathrm{d}\mathbf{y} P(\mathbf{x_0}, \mathbf{x^c}, \mathbf{y})f(\mathbf{x_0}, \mathbf{x^c}). \tag{1.56}$$

Note that

$$P(\mathbf{x_0}, \mathbf{x^c}, \mathbf{y}) = P(\mathbf{x^c}|\mathbf{x_0}, \mathbf{y})P(\mathbf{x_0}, \mathbf{y}) \tag{1.57}$$
$$= P(\mathbf{x^c}|\mathbf{y})P(\mathbf{y}|\mathbf{x_0})P(\mathbf{x_0}) \tag{1.58}$$
$$= P(\mathbf{x^c}|\mathbf{y})P(\mathbf{x_0}|\mathbf{y})P(\mathbf{y}) \tag{1.59}$$

where we used the independence of $\mathbf{x^c}$ and $\mathbf{x_0}$, and Bayes formula. Thus we reach

$$\mathbb{E}[f(\mathbf{x_0}, \mathbf{x^c})] = \int \mathrm{d}\mathbf{x_0}\, \mathrm{d}\mathbf{x^c} P(\mathbf{x^c}|\mathbf{y})P(\mathbf{x_0}|\mathbf{y})P(\mathbf{y})f(\mathbf{x_0}, \mathbf{x^c}) \tag{1.60}$$

and finally the *Nishimori identity*

$$\mathbb{E}[f(\mathbf{x}^*, \mathbf{x^c})] = \mathbb{E}[f(\mathbf{x^a}, \mathbf{x^b})]. \tag{1.61}$$

This identity implies

$$(\mathbf{x}^*, \mathbf{x^c}) \overset{\mathrm{d}}{=} (\mathbf{x^a}, \mathbf{x^b}) \tag{1.62}$$

where $\overset{\mathrm{d}}{=}$ is the equality in probability law. In particular, the overlap between the ground truth and any replica of the system, and the overlap between two distinct replicas, are the same. Thus there is only one possible value of overlap between any two different replicas of the system, which means that a Bayes optimal teacher student setting is always replica-symmetric.

1.3 Inverse Ising problem with sparse teacher couplings

This section is adapted from [3].

1.3.1 Introduction of the inverse Ising problem

We come back to an Ising model consisting of N spin variables $\mathbf{S} = \{S_i\}_{i=1,\dots,N}$ with symmetric couplings $\mathbf{J} = \{J_{ij}\}_{i,j=1,\dots,N}$, subjected to local fields $\mathbf{H} = \{H_i\}_{i=1,\dots,N}$ with Hamiltonian

$$\mathcal{H}(\mathbf{S}|\mathbf{J}, \mathbf{H}) = -\sum_{i<j}^{N} J_{ij}S_iS_j - \sum_{i=1}^{N} H_iS_i \tag{1.63}$$

and probability distribution

$$P_{\text{Ising}}(\mathbf{S}|\mathbf{J}, \mathbf{H}) = \frac{1}{\mathcal{Z}_{\text{Ising}}} e^{\sum_{i<j}^{N} J_{ij}S_iS_j + \sum_{i=1}^{N} H_iS_i}, \tag{1.64}$$

where $\mathbf{J} \in \mathbb{R}^{N\times N}$ and $\mathbf{H} \in \mathbb{R}^N$ are the couplings and external fields. The inverse Ising problem consists in inferring the couplings and external fields from a given dataset of spin samples $D^M \equiv \{\mathbf{S}^{(\mu)}\}_{\mu=1}^{M}$, where $\mathbf{S}^{(\mu)} = \{S_i^{(\mu)}\}_{i=1,\dots,N}$ is a sample of spin configuration, and M denotes the dataset size.

The inverse Ising problem is attracting more and more attention with the increasing interest in machine learning technologies. One recent application spurring this trend is for retinal neurons [141, 150], and subsequent applications to a wide range of systems have been conducted [159, 70, 176, 177, 166, 160, 161, 162], showing the potential usefulness of the inverse Ising framework.

The maximum likelihood estimator

A typical estimator for couplings and local fields is the *maximum likelihood* (ML) estimator defined by

$$\left\{\hat{\mathbf{J}}^{\mathrm{ML}}(D^M), \hat{\mathbf{H}}^{\mathrm{ML}}(D^M)\right\} = \underset{\mathbf{J},\mathbf{H}}{\arg\min} \left\{ -\sum_{\mu=1}^{M} \log P_{\mathrm{Ising}}\left(\mathbf{S}^{(\mu)}|\mathbf{J},\mathbf{H}\right)\right\}. \tag{1.65}$$

This canonical estimator enjoys some nice properties in the asymptotic limit. First, it is *consistent*: it converges in probability to the true parameter. Besides, it is *unbiased*: its average does converge to the true parameter. The ML estimator is thus a good one, but is not always appropriate for the inverse Ising framework as it has an exponentially large computational cost. Certain approximations and/or algorithms must be tailored to ease this difficulty and meet the demands of advanced applications, which has been attempted in previous studies [24, 81, 156, 74, 26, 144, 136, 108, 35, 9, 172, 94, 173].

The pseudo likelihood and local estimators

One of the most effective examples is the *pseudo likelihood* (PL) method [24, 9]. The idea is to replace the distribution P_{Ising} by another one that approximates it, and makes the estimator easier to compute. To do this, we define $\mathbf{S}_{\setminus \mathbf{i}} \equiv \{S_j\}_{j\neq i}$ the set of $N-1$ spins which does not include spin i; $\mathbf{H_i} = \{H_j\}_{j\neq i}$ and $\mathbf{J_i} = \{J_{ij}\}_{j=1,..,N}$. For each spin S_i, we introduce the conditional distribution

$$P\left(S_i|\mathbf{S}_{\setminus i}, \mathbf{J}_i, H_i\right) = \frac{1}{Z_i} e^{S_i\left(\sum_{j(\neq i)} J_{ij}S_j + H_i\right)} \tag{1.66}$$

where the normalization reads

$$Z_i = 2\cosh\left(\sum_{j(\neq i)} J_{ij}S_j + H_i\right). \tag{1.67}$$

The PL estimator is obtained separately for each spin S_i by

$$\left\{\hat{\mathbf{J}}_i^{\mathrm{PL}}(D^M), \hat{H}_i^{\mathrm{PL}}(D^M)\right\} = \underset{\mathbf{J}_i,H_i}{\arg\min} \left\{ -\sum_{\mu=1}^{M} \log P\left(S_i^{(\mu)}|\mathbf{S}_{\setminus i}^{(\mu)}, \mathbf{J}_i, H_i\right)\right\} \tag{1.68}$$

$$= \underset{\mathbf{J}_i,H_i}{\arg\min} \left\{ \sum_{\mu=1}^{M} \ell^{\mathrm{PL}}\left(S_i^{(\mu)} h_i(\mathbf{S}_{\setminus i}^{(\mu)}, \mathbf{J}_i, H_i)\right)\right\}, \tag{1.69}$$

where we have used functions

$$h_i(\mathbf{S}_{\setminus i}, \mathbf{J}_i, H_i) = \sum_{j(\neq i)} J_{ij}S_j + H_i, \tag{1.70}$$

$$\ell^{\mathrm{PL}}(x) = -x + \log 2\cosh x. \tag{1.71}$$

The PL estimator is consistent in the asymptotic limit, and it is *local*: it enables us to treat large systems because local couplings directly connected to a given spin are isolated from the other couplings, and thus can be estimated independently. Each coupling vector $\mathbf{J}_i$ can be assessed independently from the others with low (polynomial) computational cost. However, it remains a simplification of the complete setting, and loses the coupling symmetry $J_{ij} = J_{ji}$: in general the estimated couplings $\hat{J}_{ij}^{\mathrm{PL}}$ and $\hat{J}_{ji}^{\mathrm{PL}}$ will not be equal.

Note that we could do a broader analysis on any local cost function ℓ replacing ℓ^{PL} in (1.69). A remarkable advantage of local learning is its theoretical tractability in high-dimensional settings; indeed recent theoretical analyses based on the replica method revealed the tight limit of inference accuracy in the asymptotic limit [12, 11, 22].

Teacher-student scenario

We focus on the inverse Ising problem in the teacher-student scenario. The dataset D^M of M spin samples is assumed to be composed of independently identically distributed (i.i.d.) samples from a teacher Ising model with couplings $\mathbf{J}^*$ and external fields $\mathbf{H}^*$. A student Ising model attempts to infer the teacher couplings and fields from the dataset. For each site i, we measure the inference accuracy by the residual sum of squares (RSS) between the teacher couplings $\mathbf{J}_i^*$ and student's estimator $\hat{\mathbf{J}}_i$:

$$\mathcal{E} = ||\mathbf{J}_i^* - \hat{\mathbf{J}}_i||_2^2 = \sum_{j(\neq i)} \left(J_{ij}^* - \hat{J}_{ij}\right)^2 = R^* - 2\rho + R, \tag{1.72}$$

where we defined the following three macroscopic parameters:

$$R^* = \sum_{j(\neq i)} \left(J_{ij}^*\right)^2, \tag{1.73a}$$

$$R = \sum_{j(\neq i)} \left(\hat{J}_{ij}\right)^2, \tag{1.73b}$$

$$\rho = \sum_{j(\neq i)} J_{ij}^* \hat{J}_{ij}. \tag{1.73c}$$

Those parameters will naturally appear in the statistical physics analysis. We would like to evaluate the performance of the pseudo likelihood. We can compute it numerically, and evaluate the residual sum of squares, but we are also interested in finding a way to assess it analytically, which might also be helpful to compute the error of different local estimators and compare them. Previous studies of [12, 11, 22] focused on fully-connected Ising models. In high-dimensional settings, however, sparsely-connected models are more interesting because the inference accuracy is expected to be much better than the dense case. Our goal is to handle the sparsely-connected case. To do this, we investigate the teacher-student scenario using the replica method by drawing on previous studies [12, 11, 22], but refine the theoretical treatment in [12] to deal with the teacher with sparse connections.

Statistical mechanics analysis: general framework

We will go through the statistical mechanics formulation developed in [12] to analyze the theoretical performance of local learning models. For simplicity of theoretical treatment, we assume the absence of external fields both in the teacher and student models. The analysis deals with any local cost function ℓ that depends on the variable $S_i^{(\mu)} h_i(\mathbf{S}_{\setminus i}^{(\mu)}, \mathbf{J}_i, H_i)$, for instance ℓ^{PL} for the pseudo likelihood. All spins can be treated equivalently, so we pick any spin and name it spin 0, and reorder the rest as spins 1 to $N-1$. To lighten notations, we rename $\mathbf{J}$ the coupling vector $\mathbf{J_0} = \{J_{0j}\}_{j=1,\dots,N}$. The starting idea is to introduce the Hamiltonian and Boltzmann distribution induced by ℓ, instead of the natural Ising distributions which relate to

the maximum-likelihood estimator. Focusing on the zeroth spin, we thus consider

$$\mathcal{H}(\mathbf{J}|\mathbf{D}^M) = \sum_{\mu=1}^{M} \ell\left(S_0^{(\mu)} h(\mathbf{S}_{\setminus 0}^{(\mu)}, \mathbf{J})\right) \tag{1.74}$$

$$P(\mathbf{J}|\mathbf{D}^M) = \frac{1}{\mathcal{Z}} e^{-\beta \mathcal{H}(\mathbf{J}|\mathbf{D}^M)}, \tag{1.75}$$

where

$$h(\mathbf{S}_{\setminus 0}, \mathbf{J}) = \sum_{j=1}^{N-1} J_j S_j \tag{1.76}$$

$$\mathcal{Z} = \operatorname*{Tr}_{\mathbf{J}} e^{-\beta \mathcal{H}(\mathbf{J}|\mathbf{D}^M)}. \tag{1.77}$$

β is an inverse temperature, and $\mathrm{Tr}_{\mathbf{J}}$ denotes the integration with respect to $\mathbf{J}$ with an appropriate measure. For instance in the Sherrington-Kirkpatrick model, the couplings were taken to be Gaussian variables. Depending on our knowledge of the model, we should adapt the prior distribution on the couplings. Since we have no particular information here, the integration is done on the uniform measure: we simply need to rescale the couplings (this will become clear through the computation), so $\mathrm{Tr}_{\mathbf{J}} = \int \sqrt{N}\, \mathrm{d}\mathbf{J}$.

In the limit $\beta \to \infty$, the Boltzmann distribution converges to a pointwise measure on the estimator $\hat{\mathbf{J}} = \arg\min_{\mathbf{J}} \left\{ \sum_{\mu=1}^{M} \ell\left(S_0^{(\mu)} h(\mathbf{S}^{(\mu)}, \mathbf{J})\right)\right\}$. This is good news for us: we are precisely interested in characterizing this estimator. We focus on the asymptotic limit $M, N \to \infty$ while keeping $\alpha = M/N = \mathcal{O}(1)$. The average (on spin samples) free energy is self-averaging and reads:

$$\Phi = -\lim_{N\to\infty} \frac{1}{\beta N} \left[\log \mathcal{Z}\right]_{D^M} \tag{1.78}$$

where the square brackets $[\cdot]_{D^M}$ denote the average over the dataset, i.e. over the teacher Ising model:

$$[\cdot]_{D^M} = \sum_{\mathbf{S}^{(1)},\cdots,\mathbf{S}^{(M)}} (\cdot) \prod_{\mu=1}^{M} P_{\mathrm{Ising}}(\mathbf{S}^{(\mu)}|\mathbf{J}^*). \tag{1.79}$$

We will resort to the replica method to perform the difficult average over D^M, through the following replica trick:

$$f = -\lim_{N\to\infty} \frac{1}{\beta N} \left[\log Z\right]_{D^M} = -\lim_{N\to\infty} \frac{1}{\beta N} \lim_{n\to 0} \frac{1}{n} \log \left[Z^n\right]_{D^M}. \tag{1.80}$$

Assuming that n is a positive integer, we can rewrite $\left[\mathcal{Z}^n\right]_{D^M}$ as

$$\left[\mathcal{Z}^n\right]_{D^M} = \operatorname*{Tr}_{\{\mathbf{J}^a\}_{a=1}^n} \left\{ \sum_{\mathbf{S}^{(\mu)}} P_{\mathrm{Ising}}(\mathbf{S}^{(\mu)}|\mathbf{J}^*) e^{-\beta \sum_{a=1}^n \ell\left(S_0^{(\mu)} h(\mathbf{S}_{\setminus 0}^{(\mu)}, \mathbf{J}^a)\right)} \right\} \tag{1.81}$$

$$= \operatorname*{Tr}_{\{\mathbf{J}^a\}_{a=1}^n} \left\{ \sum_{\mathbf{S}} P_{\mathrm{Ising}}(\mathbf{S}|\mathbf{J}^*) e^{-\beta \sum_{a=1}^n \ell\left(S_0 h(\mathbf{S}_{\setminus 0}, \mathbf{J}^a)\right)} \right\}^M. \tag{1.82}$$

We now introduce variables $\{h^a = \sum_{j=1}^{N-1} J_j^a S_j\}_{a=1,\ldots,n}$ and $h^* = \sum_{j=1}^{N-1} J_j^* S_j$, which are called *cavity fields* since they are obtained after removing a spin from the sum, hence introducing a

"cavity". Using those variables, we get

$$[\mathcal{Z}^n]_{D^M} = \underset{\{\mathbf{J}^a\}_{a=1}^n}{\mathrm{Tr}} \left\{ \sum_{\mathbf{S}} \int \mathrm{d}h^* \prod_{a=1}^{n} \mathrm{d}h^a \right. \\ \left. \times \delta\left(h^* - \sum_{j=1}^{N-1} J_j^* S_j\right) \prod_{a=1}^{n} \delta\left(h^a - \sum_{j=1}^{N-1} J_j^a S_j\right) P_{\mathrm{Ising}}(\mathbf{S}|\mathbf{J}^*) e^{-\beta \sum_{a=1}^n \ell(S_0 h^a)} \right\}^M . \tag{1.83}$$

We perform the sum over all spins $S_1, ..., S_{N-1}$ except S_0, yielding the joint distribution $P_{\mathrm{cav}}(h^*, \{h^a\}_{a=1}^n | \mathbf{J}^*, \{\mathbf{J}^a\}_{a=1}^n)$ of the cavity fields:

$$[\mathcal{Z}^n]_{D^M} = \underset{\{\mathbf{J}^a\}_{a=1}^n}{\mathrm{Tr}} \left\{ \sum_{S_0} \int \mathrm{d}h^* \prod_{a=1}^{n} \mathrm{d}h^a P_{\mathrm{cav}}(h^*, \{h^a\}_{a=1}^n | \mathbf{J}^*, \{\mathbf{J}^a\}_{a=1}^n) \frac{1}{Z_0} e^{S_0 h^*} e^{-\beta \sum_{a=1}^n \ell(S_0 h^a)} \right\}^M \tag{1.84}$$

with the normalization constant Z_0 is defined as

$$Z_0 = \int \mathrm{d}h^* P_{\mathrm{cav}}(h^*|\mathbf{J}^*) 2\cosh h^* . \tag{1.85}$$

Our integrating variables have now become the cavity fields. To proceed further with the computation, we need to specify the functional form of the cavity field distribution, and perform the average over it. When the teacher is a fully-connected model, this analysis was done in [12]. We will review this result, and then build upon it to tackle the sparsely-connected case.

Revisiting the fully-connected case

When the teacher is a fully-connected model, we assume that dependencies between spins are so weak that the central limit theorem applies to $\{h^a = \sum_{j=1}^{N-1} J_j^a S_j\}_a$ and $h^* = \sum_{j=1}^{N-1} J_j^* S_j$, which can be considered as multivariate Gaussian variables. In [12], the authors assumed that data are sampled from the paramagnetic phase of a teacher network, and that replica symmetry (RS) holds in both the student and teacher systems, which is true for convex cost functions, e.g. in the case of pseudo likelihood. Under these assumptions, the following four order parameters are sufficient to describe the average free energy:

$$Q^* \equiv \sum_{i,j} C_{ij}^{\backslash 0} J_i^* J_j^* \tag{1.86a}$$

$$Q \equiv \sum_{i,j} C_{ij}^{\backslash 0} J_i^a J_j^a \tag{1.86b}$$

$$q \equiv \sum_{i,j} C_{ij}^{\backslash 0} J_i^a J_j^b, \ (a \neq b) \tag{1.86c}$$

$$m \equiv \sum_{i,j} C_{ij}^{\backslash 0} J_i^* J_j^a, \tag{1.86d}$$

with $\mathbf{C}^{\backslash \mathbf{0}}$ the correlation matrix between the spins:

$$C_{ij}^{\backslash 0} = \langle S_i S_j \rangle^{\backslash 0} - \langle S_i \rangle^{\backslash 0} \langle S_j \rangle^{\backslash 0} = \langle S_i S_j \rangle^{\backslash 0} , \tag{1.87}$$

where $\langle \cdots \rangle^{\backslash 0}$ denotes the average over the teacher Ising model without the zeroth spin; the last equality is due to the paramagnetic assumption. From these parameters, we can write the

covariances of the cavity fields (1.86) as

$$\left\langle (h^*)^2 \right\rangle^{\backslash 0} = Q^* \tag{1.88}$$

$$\left\langle h^* h^a \right\rangle^{\backslash 0} = m \tag{1.89}$$

$$\left\langle h^a h^b \right\rangle^{\backslash 0} = Q\delta_{ab} + (1-\delta_{ab})q. \tag{1.90}$$

We can rewrite $[Z^n]_{D^M}$ as

$$[Z^n]_{D^M} = \int \mathrm{d}Q\, \mathrm{d}q\, \mathrm{d}m\; e^{N\mathcal{S}\left(\mathbf{C}^{\backslash \mathbf{0}}, \mathbf{J}^*, Q, q, m\right) + M \log L(Q^*, Q, q, m)}, \tag{1.91}$$

where

$$\begin{aligned} e^{N\mathcal{S}\left(\mathbf{C}^{\backslash \mathbf{0}}, \mathbf{J}^*, Q, q, m\right)} \equiv & \operatorname*{Tr}_{\{\mathbf{J}^a\}_{a=1}^n} \prod_{a=1}^n \left\{ \delta\left(Q - \sum_{i,j} C_{ij}^{\backslash 0} J_i^a J_j^a\right) \delta\left(m - \sum_{i,j} C_{ij}^{\backslash 0} J_i^* J_j^a\right) \right\} \\ & \times \prod_{a<b} \delta\left(q - \sum_{i,j} C_{ij}^{\backslash 0} J_i^a J_j^b\right) \end{aligned} \tag{1.92}$$

$$L(Q^*, Q, q, m) \equiv \sum_{s_0} \int \mathrm{d}h^* \prod_{a=1}^n \mathrm{d}h^a P_{\mathrm{cav}}(h^*, \{h^a\}_{a=1}^n | Q^*, Q, q, m) \frac{1}{Z_0} e^{S_0 h^*} e^{-\beta \sum_{a=1}^n \ell(S_0 h^a)}. \tag{1.93}$$

Deferring the detailed computations to appendix A.1, we immediately have the result in the limit $n \to 0$:

$$\lim_{n\to 0} \frac{1}{n} \mathcal{S}\left(\mathbf{C}^{\backslash \mathbf{0}}, \mathbf{J}^*, Q, q, m\right) = \frac{1}{2}\left\{ \frac{Q - m^2/Q^*}{Q-q} + \log 2\pi + \log(Q-q) - \frac{1}{N} \operatorname{Tr} \log \mathbf{C}^{\backslash \mathbf{0}} \right\}, \tag{1.94}$$

$$\lim_{n\to 0} \frac{1}{n} \log L\left(Q^*, Q, q, m\right) = \int \mathrm{D}z\; e^{\sqrt{\frac{m^2}{q}} z - \frac{1}{2}\frac{m^2}{q}} \log \int \mathrm{D}v\; e^{-\beta \ell\left(\sqrt{Q-q} v + \sqrt{q} z\right)}. \tag{1.95}$$

Further, we take the limit $\beta \to \infty$, which requires the scaling of $\chi \equiv \beta(Q-q)$ to remain of order 1. After straightforward calculations, we get

$$\begin{aligned} \Phi\left(\beta \to \infty\right) &= \lim_{\beta\to\infty} -\frac{1}{\beta}\left(\lim_{n\to 0} \frac{1}{n} \mathcal{S}\left(\mathbf{C}^{\backslash \mathbf{0}}, \mathbf{J}^*, Q, q, m\right) + \alpha \lim_{n\to 0} \frac{1}{n} \log L\left(Q^*, Q, q, m\right) \right) \\ &= -\operatorname*{Extr}_{Q,\chi,m} \left\{ \frac{1}{2} \frac{Q - m^2/Q^*}{\chi} + \alpha \int \mathrm{D}z \max_y \left(-\frac{\left(y - \sqrt{Q} z - m\right)^2}{2\chi} - \ell(y) \right) \right\}. \end{aligned} \tag{1.96}$$

The extremization condition yields the following equations of state:

$$0 = \frac{1}{\chi} - \frac{\alpha}{\sqrt{Q}} \int \mathrm{D}z\; z \left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}}, \tag{1.97a}$$

$$0 = -\frac{m}{Q^* \chi} - \alpha \int \mathrm{D}z \left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}}, \tag{1.97b}$$

$$0 = -\frac{1}{\chi^2}\left(Q - \frac{m^2}{Q^*}\right) + \alpha \int \mathrm{D}z \left(\left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}} \right)^2, \tag{1.97c}$$

where

$$\hat{y}(z, Q, \chi, m) = \arg\max_{y} \left(-\frac{(y - \sqrt{Q}z - m)^2}{2\chi} - \ell(y) \right). \tag{1.98}$$

Solving (1.97), we obtain the saddle-point value of Q, χ, m. A standard technique detailed in appendix A.2 gives the following relation on the macroscopic parameters (1.73):

$$R = \left(Q - \frac{m^2}{Q^*} \right) \frac{1}{N} \mathrm{Tr} \left(\mathbf{C}^{\backslash \mathbf{0}} \right)^{-1} + R^* \left(\frac{m}{Q^*} \right)^2 \tag{1.99a}$$

$$\rho = R^* \frac{m}{Q^*}. \tag{1.99b}$$

We still need to specify the values of Q^*, R^*, and the inverse correlation function $\left(\mathbf{C}^{\backslash \mathbf{0}} \right)^{-1}$. To obtain these quantities, we have to separately solve the direct problem. Once we compute (1.99), we can directly derive the residual sum of squares $\mathcal{E}$ through (1.72) and evaluate the performance of the estimator.

1.3.2 Details of the sparsely-connected case

This section provides the extension of the above result to the sparsely-connected case, which is our main contribution in [3]. To this end, we introduce an ansatz for the estimator's behavior as well as the functional form of the cavity field distribution. Under the ansatz, the cavity field is decomposed into a signal and a noise, and it is shown that the noise part obeys essentially the same equations of state as the fully-connected case. To complete the computation under the ansatz, we will need the tree-like structure of the coupling network of the teacher model.

Difficulty of the sparse case and oracle estimator

In the fully-connected case, the cavity fields could be seen as Gaussian thanks to the central limit theorem. However, this result does not hold in the sparse case, where the sums inside the cavity fields contain a finite number of non-zero terms. The distribution of h^* actually becomes the sum of a few pointwise measures, which is far from Gaussian. Hence, we need a new ansatz to handle the cavity field distribution in the sparse case.

To find an idea of how to overcome this, let us consider an ideal situation where we know which couplings are non-zero. We assume that the zeroth spin is connected to $c = \mathcal{O}(1)$ neighboring spins, and introduce two sets of indices $\Omega = \{i | J_i^* \neq 0,\ i \in \{1, \cdots, N-1\}\}$ and $\bar{\Omega} = \{i | J_i^* = 0,\ i \in \{1, \cdots, N-1\}\}$, where Ω ($\bar{\Omega}$) is called the active (inactive) set; $|\Omega| = c$ and $|\bar{\Omega}| = N - 1 - c$. If we know Ω and $\bar{\Omega}$ in advance, then the inference should be conducted only on $\{J_i | i \in \Omega\}$. Accordingly, the number of variables to be inferred is just $c = \mathcal{O}(1)$; hence, the dataset size $M = \mathcal{O}(N)$ is sufficiently large. In this ideal case, an estimator behaves as

$$\hat{J}_i^{\mathrm{oracle}} = \begin{cases} J_i^* + \Delta_i & (i \in \Omega) \\ 0 & (i \in \bar{\Omega}) \end{cases}, \tag{1.100}$$

and we call this an *oracle* estimator. Δ_i is the "error" from the true solution, taken as a random variable. In the local learning class with appropriate cost functions such as pseudo likelihood [77], Δ_i is considered to have zero mean and a variance that decreases at the rate of $\mathcal{O}(\frac{1}{M}) = \mathcal{O}(\frac{1}{N})$. The RSS then reads

$$\mathcal{E} = \sum_{i \in \Omega} \Delta_i^2 = \mathcal{O}(\frac{1}{N}) \tag{1.101}$$

and vanishes in the asymptotic limit.

Ansatz on the estimator

Based on these observations about the oracle estimator, we assume that the (non-oracle) estimator obtained from consistent cost functions obeys the following form:

$$\hat{J}_i \doteq \begin{cases} \bar{J}_i + \Delta_i & (i \in \Omega) \\ \Delta_i & (i \in \bar{\Omega}) \end{cases}, \tag{1.102}$$

where we again assume that Δ_i is a random variable which is asymptotically zero mean with variance scaled as $\mathcal{O}(N^{-1})$; the correlations among $\{\Delta_i\}_i$ are also assumed to be sufficiently weak. The quantity $\bar{J}_i$ is interpreted as the mean value of the estimator and will deviate from the true value J_i^* owing to the extensive number of noise terms $\{\Delta_i\}_i$. The values of $\{\bar{J}_i\}_{i\in\Omega}$ are

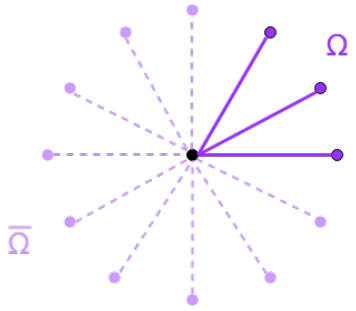


FIGURE 1.4: Ansatz on the estimator in the sparsely-connected case. The center blackdot is S_0, the three dark dots represent spins $\{S_i, i \in \Omega\}$ and the light dots are spins $\{S_i, i \in \bar{\Omega}\}$. The dark lines are the couplings $\hat{J}_i = \bar{J}_i + \Delta - i$ for $i \in \Omega$, while the light dashed lines symbolize the couplings that reduce to noise.

later computed by taking the minimization condition of the free energy as the order parameters. The applicable range of this ansatz is discussed in part 1.3.4. With this ansatz, the RSS can be written as

$$\mathcal{E} \approx \sum_{i\in\Omega} \left(J_i^* - \bar{J}_i\right)^2 + \sum_{i\in\bar{\Omega}} \Delta_i^2. \tag{1.103}$$

There are two non-negligible contributions to the RSS coming from the bias in Ω, and the noise in $\bar{\Omega}$. The RSS remains finite even in the limit $N \to \infty$, while it vanished in the ideal oracle case. The cavity field can also be decomposed as

$$h^a = h_\Omega + h_\Delta^a \tag{1.104}$$

$$h_\Omega \equiv \sum_{j\in\Omega} \bar{J}_j S_j \tag{1.105}$$

$$h_\Delta^a \equiv \sum_j \Delta_j^a S_j \approx \sum_{j\in\bar{\Omega}} \Delta_j^a S_j, \tag{1.106}$$

we call h_Δ^a is termed as the "noise" part.

Properties of h_Ω and h_Δ^a

We will assume that h_Ω and h_Δ^a are asymptotically independent in the limit $N \to \infty$. Let us explain why this assumption is reasonable. Our network has a tree-like structure, we can

thus define the generation g of a spin S from Ω as the shortest path length between S and any spin in Ω along the network. As g grows, the correlation with $\{S_i | i \in \Omega\}$ decays exponentially fast, while the number of spins belonging to generation g exponentially increases, as shown in Fig. 1.5. If the correlation decay is sufficiently faster than the increase of the spins, then the majority of spins in the network can be regarded as uncorrelated with Ω. Some terms in h_Δ^a are certainly correlated with h^*, but their contribution would then be of order $\mathcal{O}(1/\sqrt{N})$ because $\Delta_i = \mathcal{O}(1/\sqrt{N})$, and the number of correlating terms is $\mathcal{O}(1)$ thanks to the fast decay of correlations. Hence, the contribution of correlating terms vanishes and the uncorrelated majority with Ω completely dominates h_Δ^a in the thermodynamic limit. These observations indicate that

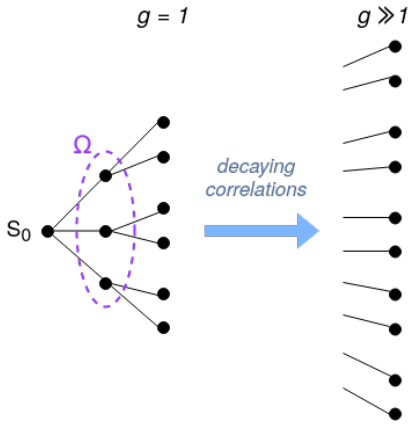


FIGURE 1.5: Representation of a tree-like network. The spin generation g (expressing its distance from S_0) grows from left to right. As g increases, the number of spins becomes exponentially large, and their correlation with S_0 decays exponentially too. The independance of h_Ω and h_Δ^a holds if the decrease in correlation is sufficiently faster.

(1.82) can be now decomposed as follows:

$$
\begin{aligned}
[\mathcal{Z}^n]_{D^M} &= \operatorname*{Tr}_{\{\mathbf{J}^a\}_{a=1}^n} \left\{ \sum_{\mathbf{S}} \int \prod_{a=1}^n \mathrm{d}h^a P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*) e^{-\beta \sum_{a=1}^n \ell(S_0 h^a)} \right\}^M \\
&\approx \operatorname*{Tr}_{\{\mathbf{\Delta}^a\}_{a=1}^n} \left\{ \sum_{\mathbf{S}} \int \prod_{a=1}^n \mathrm{d}h_\Delta^a P_{\text{Ising}}(\mathbf{s}|\mathbf{J}^*)\, \delta \left(h_\Delta^a - \sum_{j \in \bar{\Omega}} \Delta_j^a S_j \right) e^{-\beta \sum_{a=1}^n \ell\left(S_0 \left(\sum_{j\in\Omega} \bar{J}_j S_j + h_\Delta^a\right)\right)} \right\}^M
\end{aligned}
$$

where we performed the variable transformation $\mathbf{\Delta}^a = \mathbf{J}^a - \bar{\mathbf{J}}$ and neglected the contribution $\sum_{j\in\Omega} \Delta_j^a S_j$ in h_Δ^a following (1.106). We then denote $\mathbf{S}_\Omega = \{S_i | i \in \Omega\}$ and $\mathbf{S}_{\bar{\Omega}} = \{S_i | i \in \bar{\Omega}\}$, and perform the sum over $\mathbf{S}_{\bar{\Omega}}$, yielding the joint distribution $P(S_0, \mathbf{S}_\Omega, \{h_\Delta^a\}_a | \mathbf{J}^*, \{\mathbf{\Delta}^a\}_a)$:

$$
[\mathcal{Z}^n]_{D^M} = \operatorname*{Tr}_{\{\mathbf{\Delta}^a\}_{a=1}^n} \left\{ \sum_{S_0, \mathbf{S}_\Omega} \int \prod_{a=1}^n \mathrm{d}h_\Delta^a P(S_0, \mathbf{S}_\Omega, \{h_\Delta^a\}_a | \mathbf{J}^*, \{\mathbf{\Delta}^a\}_a) e^{-\beta \sum_a \ell\left(S_0 \left(\sum_{j\in\Omega} \bar{J}_j S_j + h_\Delta^a\right)\right)} \right\}^M \tag{1.107}
$$

We now invoke the asymptotic absence of correlation between h_Δ^a and $(S_0, \mathbf{S}_\Omega)$ to reach

$$
\begin{aligned}
[\mathcal{Z}^n]_{D^M} \approx \operatorname*{Tr}_{\{\mathbf{\Delta}^a\}_{a=1}^n} \Bigg\{ \sum_{S_0, \mathbf{S}_\Omega} \int \prod_{a=1}^n \mathrm{d}h_\Delta^a P_{\text{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) P_{\text{cav}}(\{h_\Delta^a\}_a | \{\mathbf{\Delta}^a\}_a) \\
\times e^{-\beta \sum_{a=1}^n \ell\left(S_0 \left(\sum_{j\in\Omega} \bar{J}_j S_j + h_\Delta^a\right)\right)} \Bigg\}^M .
\end{aligned} \tag{1.108}
$$

We can apply the central limit theorem to noise parts $\{h_\Delta^a\}_a$, that we can consider as Gaussian variables. Like in the fully-connected case, two order parameters describing their covariances are introduced:

$$Q \equiv \sum_{i,j} C_{ij}^{\backslash 0} \Delta_i^a \Delta_j^a, \tag{1.109}$$

$$q \equiv \sum_{i,j} C_{ij}^{\backslash 0} \Delta_i^a \Delta_j^b, \ (a \neq b). \tag{1.110}$$

Average free energy in the sparsely-connected case

We do not need counterparts of parameters m and Q^* from the densely-connected case here, because the dependence on $(S_0, \mathbf{S}_\Omega)$ is separately and explicitly treated in the present formulation. Then,

$$[\mathcal{Z}^n]_{D^M} \approx \int \mathrm{d}Q\, \mathrm{d}q\ e^{N\mathcal{S}(\mathbf{C}^{\backslash \mathbf{0}}, Q, q) + M \log L(\mathbf{J}^*, \bar{\mathbf{J}}, Q, q)}, \tag{1.111}$$

where

$$e^{N\mathcal{S}(\mathbf{C}^{\backslash \mathbf{0}}, Q, q)} \equiv \operatorname*{Tr}_{\{\mathbf{\Delta}^a\}_{a=1}^n} \prod_{a=1}^n \delta\left(Q - \sum_{i,j} C_{ij}^{\backslash 0} \Delta_i^a \Delta_j^a\right) \prod_{a<b} \delta\left(q - \sum_{i,j} C_{ij}^{\backslash 0} \Delta_i^a \Delta_j^b\right) \tag{1.112}$$

$$\begin{aligned} L(\mathbf{J}^*, \bar{\mathbf{J}}, Q, q) = \sum_{S_0, \mathbf{S}_\Omega} \int \prod_{a=1}^n \mathrm{d}h_\Delta^a P_{\mathrm{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) P_{\mathrm{cav}}(\{h_\Delta^a\}_{a=1}^n | Q, q) \\ \times e^{-\beta \sum_{a=1}^n \ell\left(S_0\left(\sum_{j\in\Omega} \bar{J}_j S_j + h_\Delta^a\right)\right)}. \end{aligned} \tag{1.113}$$

Again, using the techniques in appendix A.1 we get

$$\lim_{n\to 0} \frac{1}{n} \mathcal{S}\left(\mathbf{C}^{\backslash \mathbf{0}}, Q, q\right) = \frac{1}{2}\left\{\frac{Q}{Q-q} + \log 2\pi + \log(Q-q) - \frac{1}{N} \operatorname{Tr} \log C^{\backslash 0}\right\}, \tag{1.114}$$

$$\begin{aligned} \lim_{n\to 0} \frac{1}{n} \log L\left(\mathbf{J}^*, \bar{\mathbf{J}}, Q, q\right) = \sum_{S_0, \mathbf{S}_\Omega} P_{\mathrm{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) \\ \times \int \mathrm{D}z \log \int \mathrm{D}v\ e^{-\beta \ell\left(S_0\left(\sum_{j\in\Omega} \bar{J}_j S_j + \sqrt{Q-q} v + \sqrt{q} z\right)\right)}. \end{aligned} \tag{1.115}$$

Recalling the finite scaling of $\chi = \beta(Q-q)$ and taking the $\beta \to \infty$ limit, we get

$$\begin{aligned} \Phi = -\operatorname*{Extr}_{Q,\chi} \Bigg\{ \frac{1}{2}\frac{Q}{\chi} \\ + \alpha \sum_{S_0, \mathbf{S}_\Omega} P_{\mathrm{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) \int \mathrm{D}z \max_y \left[-\frac{\left(y - S_0(\sqrt{Q} z + \sum_{j\in\Omega} \bar{J}_j S_j)\right)^2}{2\chi} - \ell(y) \right] \Bigg\}. \end{aligned} \tag{1.116}$$

The extremization condition with respect to Q and χ gives

$$0 = \frac{1}{\chi} - \frac{\alpha}{\sqrt{Q}} \sum_{S_0, \mathbf{S}_\Omega} P_{\mathrm{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) S_0 \int \mathrm{D}z\, z \left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}}, \tag{1.117a}$$

$$0 = -\frac{Q}{\chi^2} + \alpha \sum_{S_0, \mathbf{S}_\Omega} P_{\mathrm{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) \int \mathrm{D}z \left(\left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}}\right)^2, \tag{1.117b}$$

where

$$\hat{y}(z, S_0, \mathbf{S}_\Omega | Q, \chi, \{\bar{J}_j\}_{j\in\Omega}) = \arg\max_y \left\{ -\frac{\left(y - S_0\left(\sqrt{Q}z + \sum_{j\in\Omega}\bar{J}_j S_j\right)\right)^2}{2\chi} - \ell(y) \right\}. \quad (1.118)$$

The mean estimates $\{\bar{J}_j\}_{j\in\Omega}$ are also evaluated by the extremization condition. The result for $\bar{J}_j$ is given by

$$0 = \sum_{S_0,\mathbf{S}_\Omega} P_{\text{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) \int \mathrm{D}z \frac{y^* - S_0(\sqrt{Q}z + \sum_{j\in\Omega}\bar{J}_j S_j)}{2\chi} S_0 S_j \quad (1.119)$$

$$= \sum_{S_0,\mathbf{S}_\Omega} P_{\text{Ising}}(S_0, \mathbf{S}_\Omega | \mathbf{J}^*) \int \mathrm{D}z \left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}} S_0 S_j. \quad (1.120)$$

If we solve (1.117), (1.120), we obtain parameters $Q, \chi, \{\bar{J}_i\}_{i\in\Omega}$, and we can evaluate the residual sum of squares which is expressed in the present setting as

$$\mathcal{E} \approx \sum_{i\in\Omega}(J_i^* - \bar{J}_i)^2 + \sum_{i\in\bar{\Omega}}\Delta_i^2 = \sum_{i\in\Omega}(J_i^* - \bar{J}_i)^2 + \frac{Q}{N}\,\mathrm{Tr}\left(\mathbf{C}^{\backslash \mathbf{0}}\right)^{-1}. \quad (1.121)$$

We still need to know $\left(\mathbf{C}^{\backslash \mathbf{0}}\right)^{-1}$, and $P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*)$ also needs to be assessed in the sparsely-connected case. We turn to the direct problem to evaluate these quantities.

1.3.3 Properties of the direct problem

The inverse problem essentially requires certain information from its direct problem counterpart. In the fully-connected case, two-body quantities such as $\left(\mathbf{C}^{\backslash \mathbf{0}}\right)^{-1}$ and $\sum_{i,j} C_{ij}^{\backslash 0} J_i^* J_j^*$ are sufficient. However, in the sparse case, higher-order information is needed because the central limit theorem does not fully describe the dominant terms in the system. Hence, the functional form of $P_{\text{Ising}}(S_0, \mathbf{S}_\Omega)$ becomes necessary, as seen in (1.116). Techniques for computing such quantities in the sparse case largely advanced in the '90–'00s. We will use a portion of the results to compute the necessary quantities, and more detailed techniques can be found in [120, 105]. For our work, we rely on the assumptions that the teacher model is in the paramagnetic phase and the external fields are absent.

Marginal distribution of the teacher model

The marginal distribution $P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*)$ is computed by marginalizing the whole distribution $P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*)$ with respect to $\mathbf{S}_{\bar{\Omega}}$. In general, this operation requires nontrivial computations and the resultant distribution becomes dependent on parameters among the marginalized spins. However, under the present assumptions, such dependencies do not exist and the expression becomes rather simple:

$$P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*) = \frac{1}{Z_{\Omega_c}} e^{S_0 \sum_{j\in\Omega} J_j^* S_j} \quad (1.122)$$

$$Z_{\Omega_c} = \sum_{S_0,\mathbf{S}_\Omega} e^{S_0 \sum_{j\in\Omega} J_j^* S_j} \quad (1.123)$$

where Ω_c denotes the union index set of 0 and Ω. This form is applied to (1.117), (1.120) to obtain the order parameters.

Inverse correlation function

We will now compute the inverse correlation function $(\mathbf{C}^{\backslash \mathbf{0}})^{-1}$; the superscript $^{\backslash 0}$ is not essential so we discard it and treat the whole system. The so-called *Gibbs free energy* G will serve our purpose. It is similar to the free energy we have defined previously, but adds a term that depends on its argument $\mathbf{m}$:

$$G(\mathbf{m}) \equiv \max_{\theta} \left\{ \theta^\top \mathbf{m} - \log Z(\theta) \right\}, \tag{1.124}$$

where $Z(\theta) = \sum_{\mathbf{S}} e^{\sum_{i<j}^{N} J_{ij} S_i S_j + \sum_{i=1}^{N} H_i S_i + \sum_i \theta_i S_i}$. A precious identity is the equality between the Hessian of G, and the inverse correlation function:

$$\left(C^{-1}\right)_{ij} = \frac{\partial^2 G}{\partial m_i \partial m_j} \tag{1.125}$$

We may thus focus on computing G.

For sparsely-connected graphs in which loops can be neglected, the Gibbs free energy corresponds to the so-called *Bethe free energy*, which consists of two contributions corresponding to factor and variable nodes, in the asymptotic limit. More insight on the Bethe free energy will be given in 2.1.3. In our case, G is known to have the following form:

$$G_{\text{Bethe}}(\mathbf{m}) = \sum_{e \in E} \operatorname*{Tr}_{\mathbf{S}_e} b_e(\mathbf{S}_e) \log \left(\frac{b_e(\mathbf{S}_e)}{e^{J_e \prod_{i \in e} S_i}} \right) + \sum_{i=1}^{N} (c_i - 1) \mathcal{S}_m(m_i) \tag{1.126}$$

where c_i denotes the connectivity or the number of edges connecting to node i, and $\mathcal{S}_m(m)$ is the entropy conditioned by the magnetization m:

$$\mathcal{S}_m(m) = -\frac{1-m}{2} \log \frac{1-m}{2} - \frac{1+m}{2} \log \frac{1+m}{2}, \tag{1.127}$$

and e and E denote an edge and the set of edges, respectively. With a slight abuse of notation, e also represents the index set of variable nodes connected to the edge e, allowing us to use expressions like $J_e \prod_{i \in e} S_i$ and $\mathbf{S}_e$ denoting the spins connected to the edge e. The factor b_e represents the marginal distribution of $\mathbf{S}_e$ which can be parametrized as

$$b_e(\mathbf{S}_e) \propto e^{J_e \prod_{i \in e} S_i} \prod_{i \in e} \frac{e^{S_i h_{i \to e}}}{2 \cosh h_{i \to e}} \tag{1.128}$$

where $h_{i \to e}$ is an auxiliary external field (also usually called cavity field) necessary to match the average $\langle S_i \rangle$ with the given value m_i. Specifying the node indices in the edge e as $e = (i, j)$, we can explicitly write equations to be satisfied as

$$\tanh^{-1}(m_i) = h_{i \to e} + \tanh^{-1}\left(\tanh(J_e) \tanh(h_{j \to e})\right) \tag{1.129a}$$

$$\tanh^{-1}(m_j) = h_{j \to e} + \tanh^{-1}\left(\tanh(J_e) \tanh(h_{i \to e})\right). \tag{1.129b}$$

In general, $\{h_{i \to e}\}_{i,e}$ can have a complicated dependence relation. As a result, the computation of the Hessian of G becomes difficult, although numerically doable, and we cannot have a compact analytic form of the inverse correlation function. Fortunately, under the paramagnet and no external field assumptions, we can assume that h and m are small, and linearize (1.129) with

respect to them, yielding

$$h_{i\to e}=\frac{m_i-\tanh(J_e)m_j}{1-\tanh^2(J_e)}\qquad h_{j\to e}=\frac{m_j-\tanh(J_e)m_i}{1-\tanh^2(J_e)}. \tag{1.130}$$

Inserting this into (1.126) and expanding it with respect to m up to the second order, we get

$$G_{\text{Bethe}}(m)\approx\sum_{e\in E}\left\{\sum_{i\in e}\frac{m_i^2}{2(1-\tanh^2(J_e))}-\frac{\tanh(J_e)}{1-\tanh^2(J_e)}\prod_{i\in e}m_i\right\}-\sum_i(c_i-1)\frac{1}{2}m_i^2+\text{cst}. \tag{1.131}$$

From here, we compute the Hessian and the expression of the inverse correlation function:

$$\left(C^{-1}\right)_{ij}=\left(\sum_{k\in\partial i}\frac{1}{1-\tanh^2(J_{ik})}-(c_i-1)\right)\delta_{ij}-\frac{\tanh(J_{ij})}{1-\tanh^2(J_{ij})}(1-\delta_{ij}), \tag{1.132}$$

where ∂i denotes the index set of nodes connected to i. This expression can be applied even if there is no edge for (i,j) (namely $J_{ij}=0$).

1.3.4 Applicable range of the ansatz

We now discuss the applicable range of the ansatz (1.102). This ansatz is a strong statement, since it allows us to set aside any possible biases of the estimator components outside the active set Ω. When is it valid and how does it relate to the tree-like network structure?

Zero gradient conditions for Ω

To answer these questions, we rethink (1.120) which stated for every $j\in\Omega$

$$0=\sum_{S_0,\mathbf{S}_\Omega}P_{\text{Ising}}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*)\int\mathrm{D}z\left.\frac{\partial\ell}{\partial y}\right|_{y=\hat{y}}S_0S_j. \tag{1.133}$$

An important observation is that this equation is merely the zero-gradient condition of ℓ with respect to J_j ($j\in\Omega$), averaged over the datasets. Let us justify this equality. Denoting the empirical average on the dataset D^M by $\overline{\cdots}^{D^M}$, and using the statistical mechanical analysis explained so far, we can start by directly writing the zero-gradient condition with respect to J_j for $j\in\Omega$ as

$$0=\overline{\left.\frac{\partial\ell(S_0h(\mathbf{S}_{\setminus 0},\mathbf{J}))}{\partial J_j}\right|_{\mathbf{J}=\hat{\mathbf{J}}(D^M)}}^{D^M}=\overline{\frac{\partial\ell(y)}{\partial y}S_0S_j}^{D^M} \tag{1.134}$$

where we name $y=S_0h(\mathbf{S}_{\setminus 0},\hat{\mathbf{J}})=S_0(\sum_i\hat{J}_iS_i)$. We would like this equation to yield the same result as (1.120). We replace the estimator $\hat{\mathbf{J}}$ with the average over (1.75) (we will then take $\beta\to\infty$ to recover the estimator properties), take the average over the dataset, and introduce n replicas. These manipulations take us to

$$\begin{aligned}0=\lim_{\beta\to\infty}\lim_{n\to 0}\operatorname*{Tr}_{\{\mathbf{J}^a\}_{a=1}^n}&\left(\sum_{\mathbf{S}}P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*)e^{-\beta\sum_{a=1}^n\ell(y^a)}\right)^{M-1}\\&\times\left(\sum_{\mathbf{S}}P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*)e^{-\beta\sum_{a=1}^n\ell(y^a)}\frac{\partial\ell(y^1)}{\partial y^1}S_0S_j\right).\end{aligned} \tag{1.135}$$

where $y^a \equiv S_0 \left(\sum_i J_i^a S_j\right)$. The ansatz on the estimator (1.102) and replica symmetry used in 1.3.2, say in short that

$$y^a \overset{\text{ansatz}}{=} S_0 \left\{ \sum_{i\in\Omega} \bar{J}_i S_i + h_\Delta^a \right\} \overset{(\text{RS})}{=} S_0 \left\{ \sum_{i\in\Omega} \bar{J}_i S_i + \sqrt{Q-q} v^a + \sqrt{q} z \right\}. \tag{1.136}$$

where $v^a, z \sim \mathcal{N}(0,1)$. Applying this form and following the same line of computations as in 1.3.2, we get

$$\begin{aligned}
&\left(\sum_{\mathbf{S}} P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*) e^{-\beta \sum_{a=1}^n \ell(y^a)} \frac{\partial \ell(y^1)}{\partial y^1} S_0 S_j \right) \\
&\overset{N\to\infty}{\longrightarrow} \sum_{S_0,\mathbf{S}_\Omega} P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*) \int \mathrm{D}z \left(\int \mathrm{D}v e^{-\beta \ell(y(z,v))} \right)^n \frac{\int \mathrm{D}v^1 e^{-\beta \ell(y(z,v^1))} \frac{\partial \ell(y^1)}{\partial y^1} S_0 S_j}{\int \mathrm{D}v^1 e^{-\beta \ell(y(z,v^1))}} \\
&\overset{n\to 0}{\longrightarrow} \sum_{S_0,\mathbf{S}_\Omega} P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*) \int \mathrm{D}z \frac{\int \mathrm{D}v e^{-\beta \ell(y(z,v))} \frac{\partial \ell(y)}{\partial y} S_0 S_j}{\int \mathrm{D}v e^{-\beta \ell(y(z,v))}} \\
&\overset{\beta\to\infty}{\longrightarrow} \sum_{S_0,\mathbf{S}_\Omega} P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*) \int \mathrm{D}z \left. \frac{\partial \ell(y)}{\partial y} \right|_{y=\hat{y}} S_0 S_j,
\end{aligned} \tag{1.137}$$

and the factor $\left(\sum_{\mathbf{S}} P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*) e^{-\beta \sum_{a=1}^n \ell(y^a)}\right)^{M-1}$ in (1.135) becomes unity when taking the $n \to 0$ limit. Therefore, we correctly recover the same equation as (1.120).

Validity of the ansatz on tree-like graphs

This computation naturally leads to the following question: Should we compute all the zero-gradient conditions not only for Ω but also for $\bar{\Omega}$? This point is important: if the answer is affirmative, then the ansatz (1.102) is insufficient as it only defines $\bar{J}_j$ for the active set $j \in \Omega$. To be consistent, the answer should be a priori considered to be yes in general; hence, we need to take into account the zero-gradient conditions for $k \in \bar{\Omega}$. This implies that the ansatz (1.102) should be modified, and we need to introduce mean estimates $\bar{J}_k$ for $k \in \bar{\Omega}$ in general situations.

Fortunately, if the network is tree-like, we can show that all the zero-gradient conditions are automatically satisfied once those for all $j \in \Omega$ are met. To show this, we recover the external field $\mathbf{H}^*$ for technical reasons. When the external field exists, the student model should also have an external field variable, and hence the replica result is slightly modified. That modification is accomplished by replacing $\sum_{j\in\Omega} \bar{J}_j S_j$ with $\sum_{j\in\Omega} \bar{J}_j S_j + \bar{H}_0$ in (1.116), (1.118) and (1.136). Here, $\bar{H}_0$ denotes the mean estimate of the external field variable acting on the focused spin S_0 of the student model, and is determined by the extremization condition of the free energy, yielding

$$0 = \sum_{S_0,\mathbf{S}_\Omega} P_{\text{Ising}}(S_0, \mathbf{S}_\Omega|\mathbf{J}^*, \mathbf{H}^*) \int \mathrm{D}z \left. \frac{\partial \ell}{\partial y} \right|_{y=\hat{y}} S_0. \tag{1.138}$$

Under this setup, we would like to show consistency of (1.102) on tree-like networks. The first step is to write down the zero-gradient condition for $k \in \bar{\Omega}$. The result of applying the averages and replica method is simply the replacement of S_j with S_k in (1.135). Recalling that

$$P_{\text{Ising}}(\mathbf{S}|\mathbf{J}^*, \mathbf{H}^*) = \frac{1}{\mathcal{Z}_{\text{Ising}}} e^{\sum_{i<j}^N J_{ij}^* S_i S_j + \sum_{i=1}^N H_i^* S_i}, \tag{1.139}$$

a simple computation of its derivative with respect to H_k^* gives the following relation:

$$\sum_{\mathbf{S}} P_{\rm Ising}(\mathbf{S}|\mathbf{J}^*,\mathbf{H}^*)e^{-\beta\sum_{a=1}^n \ell(y^a)}\frac{\partial \ell(y^1)}{\partial y^1}S_0S_k = \frac{\partial}{\partial H_k^*}\left(\sum_{\mathbf{S}} P_{\rm Ising}(\mathbf{S}|\mathbf{J}^*,\mathbf{H}^*)e^{-\beta\sum_{a=1}^n \ell(y^a)}\frac{\partial \ell(y^1)}{\partial y^1}S_0\right)$$
$$+\left(\sum_{\mathbf{S}} P_{\rm Ising}(\mathbf{S}|\mathbf{J}^*,\mathbf{H}^*)e^{-\beta\sum_{a=1}^n \ell(y^a)}\frac{\partial \ell(y^1)}{\partial y^1}S_0\right)\langle S_k\rangle \tag{1.140}$$

where $\langle\cdots\rangle$ denotes the average over $P_{\rm Ising}(\mathbf{S}|\mathbf{J}^*,\mathbf{H}^*)$. Let us look at the last term in the limits $\beta\to\infty, n\to 0, N\to\infty$. The coefficient before $\langle S_k\rangle$ converges to the right-hand side of (1.138), giving zero. Meanwhile, in the same limit, the first term can be transformed as

$$\frac{\partial}{\partial H_k^*}\left(\sum_{\mathbf{S}} P_{\rm Ising}(\mathbf{S}|\mathbf{J}^*,\mathbf{H}^*)e^{-\beta\sum_{a=1}^n \ell(y^a)}\frac{\partial \ell(y^1)}{\partial y^1}S_0\right)$$
$$\to \frac{\partial}{\partial H_k^*}\left(\sum_{S_0,\mathbf{S}_\Omega} P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*,\mathbf{H}^*)\int \mathrm{D}z\left.\frac{\partial \ell(y)}{\partial y}\right|_{y=\hat{y}} S_0\right), \tag{1.141}$$

and the dependence on H_k^* appears only in the marginal distribution $P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*,\mathbf{H}^*)$. On tree-like networks, the marginal distribution necessarily takes the following form:

$$P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*,\mathbf{H}^*) = \frac{1}{Z}e^{S_0(\sum_{i\in\Omega}J_i^*S_i+H_0^*)+\sum_{i\in\Omega}h_i^{\backslash 0}S_i} \tag{1.142}$$

where $h_i^{\backslash 0}$ is the effective field obtained by marginalizing the descendant spins of i, and is usually termed as cavity field. An important point of (1.142) is the absence of higher-order interactions among active set spins because of the tree-like structure. Hence, the dependence on H_k^* appears only through the effective fields $h_i^{\backslash 0}$. Furthermore, owing to the tree-like structure, only one of the effective fields is dependent on H_k^*. Specifying the corresponding index as $j(\in\Omega)$, we get

$$\frac{\partial}{\partial H_k^*}\left(\sum_{S_0,\mathbf{S}_\Omega} P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*,\mathbf{H}^*)\int \mathrm{D}z\left.\frac{\partial \ell(y)}{\partial y}\right|_{y=\hat{y}} S_0\right)$$
$$= \frac{\partial h_j^{\backslash 0}}{\partial H_k^*}\frac{\partial}{\partial h_j^{\backslash 0}}\left(\sum_{S_0,\mathbf{S}_\Omega} P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*,\mathbf{H}^*)\int \mathrm{D}z\left.\frac{\partial \ell(y)}{\partial y}\right|_{y=\hat{y}} S_0\right)$$
$$= \frac{\partial h_j^{\backslash 0}}{\partial H_k^*}\left\{\sum_{S_0,\mathbf{S}_\Omega} P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*)\int \mathrm{D}z\left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}} S_0S_j \right. \tag{1.143}$$
$$\left. - \sum_{S_0,\mathbf{S}_\Omega} P_{\rm Ising}(S_0,\mathbf{S}_\Omega|\mathbf{J}^*,\mathbf{H}^*)\int \mathrm{D}z\left.\frac{\partial \ell}{\partial y}\right|_{y=\hat{y}} S_0\langle S_j\rangle\right\} \tag{1.144}$$
$$= 0.$$

In lines (1.143) and (1.144), we recognize the right-hand side terms of (1.120) and (1.138), which are both equal to zero. Hence, the zero-gradient conditions on the inactive set $\bar{\Omega}$ are satisfied once those of the active set Ω hold, and the ansatz (1.102) holds on tree-like networks.

This proof also provides a perspective for loopy graphs. If loops exist, then higher-order interactions emerge in $P_{\rm Ising}(S_0,\mathbf{S}_\Omega)$; they generally depend on H_k^* in a complex manner and yield some additional terms as a result of differentiation. In such situations, additional mean estimates $\bar{J}_k$ for $k\in\bar{\Omega}$ will be necessary to satisfy the corresponding zero-gradient conditions;

however, treating all variables in $\bar{\Omega}$ is clearly infeasible. Tailoring good approximations in such cases may be interesting in future work, although we show in 1.3.5 an example in which our present theoretical treatment becomes a good approximation even for loopy graphs.

1.3.5 Numerical experiments

In this part, we conduct numerical experiments to check the accuracy of the theoretical computations. The actual behavior of the order parameters and related quantities depends on the details of the coupling ensembles. We treat the regular random (RR) graph and Erdős–Rényi (ER) graph as representative examples of sparse tree-like graphs. The RR graph is characterized by one connectivity parameter c, while the ER graph is characterized by the connection probability p. To keep the generated graph sparse enough in the ER case, we assume the probability is scaled as $p = d/N$, with d the mean degree. Furthermore, we also assume that the couplings of the teacher model have the same probability of taking both signs and the strength is constant: $|J_i^*| = K > 0$. The coupling strength K is assumed to be small enough to satisfy the paramagnet assumption of the teacher model. In particular, for the RR graph, the paramagnetic condition is

$$(c-1)\tanh^2 K < 1, \tag{1.145}$$

while that of the ER one is

$$d\tanh^2 K < 1. \tag{1.146}$$

The derivation of these bounds can be found in [120, 105]. The cost function is fixed to the pseudo likelihood one in the following, as the simplest and commonly used case. The result for the RR graph case is shown below in 1.3.5, and that for the ER graph is in 1.3.5. For comparison, some numerical results on the square lattice are shown in 1.3.5, focusing on the approximation nature of the present theoretical results. Furthermore, as another common cost function, the so-called *interaction screening* (IS) method [172] belonging to the local learning class is examined and compared with pseudo likelihood.

Thanks to the uniformity of the coupling strength, the strength of mean estimates $\{\bar{J}_i\}_{i\in\Omega}$ can also be set to a uniform value $|\bar{J}_i| = \bar{K} = \hat{b}K$, where the bias factor

$$\hat{b} \equiv \bar{K}/K \tag{1.147}$$

is introduced. For the same reason, the marginal distribution can be simplified by again introducing the cavity field $h^* = \sum_{j\in\Omega} J_j^* S_j$ as

$$\sum_{S_0,\mathbf{S}_\Omega} P(S_0,\mathbf{S}_\Omega|\mathbf{J}^*)(\cdots) = \sum_{S_0}\int \mathrm{d}h^* P_{\mathrm{cav}}(h^*|\mathbf{J}^*)\frac{e^{S_0h^*}}{Z_0}(\cdots) \tag{1.148}$$

where $Z_0 = \int \mathrm{d}h^* 2P_{\mathrm{cav}}(h^*|\mathbf{J}^*)\cosh h^*$. If the considered spin's connectivity is c, then the cavity field distribution becomes

$$P_{\mathrm{cav}}(h^*|\mathbf{J}^*) = P_{\mathrm{cav}}(h^*|K,c) \equiv \frac{1}{2^c}\sum_{k=0}^{c}\binom{c}{k}\delta\left(h^* - K(c-2k)\right). \tag{1.149}$$

Applying the reduction (1.148) in (1.117), (1.120) after replacing $\sum_{j\in\bar{\Omega}}\bar{J}_j s_j$ by $\hat{b}h^*$ in (1.118) reduces the computation of mean estimates to that of the bias factor $\hat{b}$. The theoretically evaluated $\hat{b}$ was compared with that obtained by numerical experiments to check the validity of our

theoretical treatment.

The numerical computation of the order parameter Q will be conducted below, but it has some delicate points. In our actual computations, the following procedure was adopted: From the generated teacher model we first compute the inverse correlation function $\left(\mathbf{C}^{\backslash \mathbf{0}}\right)^{-1}$ by the cavity formula (1.132) and numerically invert it to obtain $\mathbf{C}^{\backslash \mathbf{0}}$. Then we introduce $\{\hat{\Delta}_i = \hat{J}_i - \bar{J}_i\}_{i=1}^{N-1}$ from the learning result $\hat{\mathbf{J}}$, and the mean estimate $\bar{\mathbf{J}}$ which is obtained as $\bar{\mathbf{J}} = \hat{b}\mathbf{J}^*$ where the theoretically evaluated value of $\hat{b}$ is inserted. Finally, we get a numerical value of Q through the relation $Q = \sum_{i,j} C_{ij}^{\backslash 0} \hat{\Delta}_i \hat{\Delta}_j$. Although it is also possible to evaluate $\mathbf{C}^{\backslash \mathbf{0}}$ by the Monte Carlo (MC) sampling instead of using formula (1.132), this method is better for controlling fluctuations and reducing computational cost.

The outline of our numerical experiment is as follows. We first generate teacher model, next perform MC sampling, and finally choose a center spin and conduct learning by numerically minimizing the PL cost function (1.68). The obtained estimator is used to compute relevant quantities such as the residual sum of squares (RSS). As a result, there are some distinctive sources of fluctuation in the estimate, but we do not discriminate them below. The error bar is accordingly defined by the standard error coming from those fluctuations. The number of datasets used to compute the error bar is hereafter denoted as N_{set}. More details of the numerical experiment are summarized in A.3.

RR graph case

In the case of the RR graph with connectivity c, using (1.132) the trace of the inverse correlation function becomes

$$\frac{1}{N} \operatorname{Tr} \mathbf{C}^{-1} = \frac{c}{1 - \tanh^2 K} - c + 1. \tag{1.150}$$

Substituting this in conjunction with the parameters obtained by (1.117), (1.120) into (1.121), we obtain the RSS. Below, we compare these theoretical values with the numerically evaluated ones.

We start by comparing the theoretical and numerical values of $\mathcal{E}$, Q, and $\hat{b}$. In Fig. 1.10, these quantities are plotted against α for $K = 0.2$ and 0.4 at $N = 200$ and $c = 3$. In all the plots,

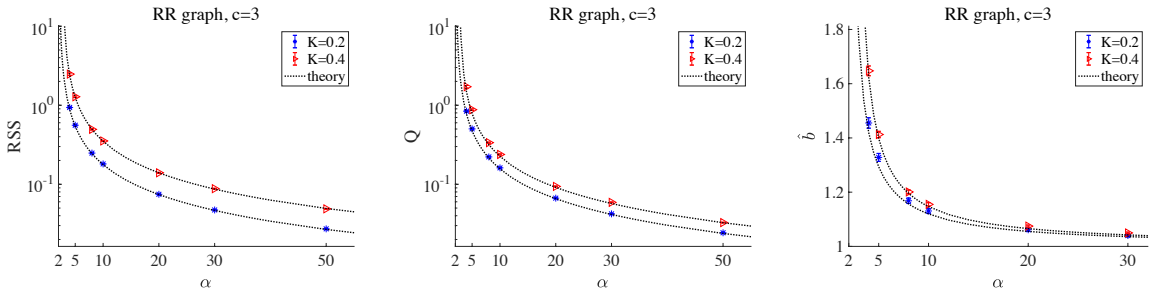


FIGURE 1.6: Plots of $\mathcal{E}$ (left), Q (middle), and $\hat{b}$ (right) against α for $K = 0.2$ and 0.4 at $(N, c) = (200, 3)$. Dotted lines and color markers are the theoretical and numerical values, respectively. The agreement between them is fairly good. The left and middle panels are plotted in the double log scale because $\mathcal{E}$ and Q drastically diverge in the limit $\alpha \to 2$. The error bars obtained from $N_{\text{set}} = 100$ datasets are shown, although they tend to be comparable with the size of markers.

the agreement between the theoretical (dotted lines) and numerical (color markers) results is

fairly good, supporting the validity of our analytical treatment.

Next, we consider the distributions of the estimators in Fig. 1.7, which were normalized as probability distribution functions. The left panel is the distribution of the estimators on the active set Ω. We can observe that two peaks are located around the theoretical prediction $\pm\hat{b}K$. In the middle panel, the estimator distribution on the inactive set $\bar{\Omega}$ is shown, yielding a Gaussian-like distribution with zero mean. Similar behavior is observed for the noise part on the active set, $\{\hat{\Delta}_i = \hat{J}_i - \bar{J}_i\}_{i\in\Omega}$, the distribution of which is given in the right panel. Here, the mean estimates $\{\bar{J}_i\}_{i\in\Omega}$ are computed by multiplying the theoretically evaluated bias $\hat{b}$ by the true coupling $\{J_i^*\}_{i\in\Omega}$. These observations are again consistent with our theoretical analysis.

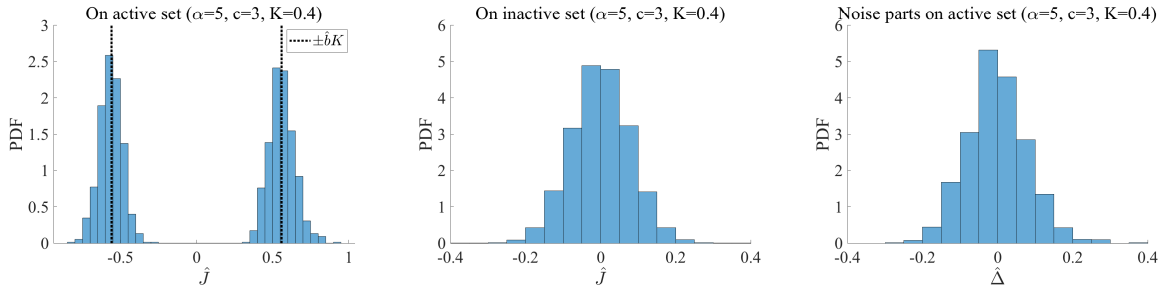


FIGURE 1.7: Distribution of the estimators $\hat{\mathbf{J}}$ on the active and inactive sets are given in the left and middle panels, respectively. The right panel is the distribution of the noise part on the active set, $\{\hat{\Delta}_i = \hat{J}_i - \bar{J}_i\}_{i\in\Omega}$. The system parameters are $(N, K, \alpha, c) = (200, 0.4, 5, 3)$. The middle and right panels imply that the noise parts obey the zero-mean Gaussian distribution and have no discriminative difference between the active and inactive sets. Here, the histograms are generated from $N_{\rm set} = 500$ datasets; from each dataset, the number of obtained estimators is $c = 3$ for Ω while that for $\bar{\Omega}$ is $N - c - 1 = 196$.

Thirdly, we check the finite size effect. In Fig. 1.8, the RSS and rescaled variance (multiplied by N) of the noise parts $\hat{\mathbf{\Delta}} = \hat{\mathbf{J}} - \bar{\mathbf{J}}$ are plotted with respect to system size N, in the upper and lower panels respectively. Although the finite size effect behaves in different ways depending on the parameters and quantities, we can see that the numerical results (markers) fairly match the theoretical values (black dotted lines) as the system size is large. Here, the rescaled variance corresponds to the quantity $Q\,\mathrm{Tr}\left(\mathbf{C}^{\backslash\mathbf{0}}\right)^{-1}/N$ in our theoretical computation, which is consistent with (1.121). These results again confirm the validity of our computations.

Finally, we hint to some noteworthy remarks. The results shown in Figs. 1.7 and 1.8 imply the possibility of an efficient method of debiasing. The bias factor $\hat{b}$ can be computed from our analytical result, and hence we can debias our estimator in an efficient manner. The residual after debiasing $\hat{\mathbf{\Delta}}$ is considered to obey a Gaussian distribution, as shown in Fig. 1.7, and is supported by our analytical computations in A.1. Thus, we can efficiently compute the P-value according to the standard hypothesis testing method, enabling us to judge the relevance of the estimated couplings. Moreover, in the thermodynamic limit $N \to \infty$, we can show that the perfect reconstruction of the teacher's network is possible for any $\alpha > 2$. To do so, we need to evaluate the probability of getting false positives in the estimator. To control false positives, we introduce a constant threshold value $K_{\rm th}(> 0)$, and consider estimated couplings with absolute values less than $K_{\rm th}$ as negligible and set to zero; we independently repeat this procedure for all $i = 1, ..., N$. Let us evaluate the probability of successfully screening out false positives using this method. The observations so far imply, on the inactive set $\bar{\Omega}$, that the estimator behaves

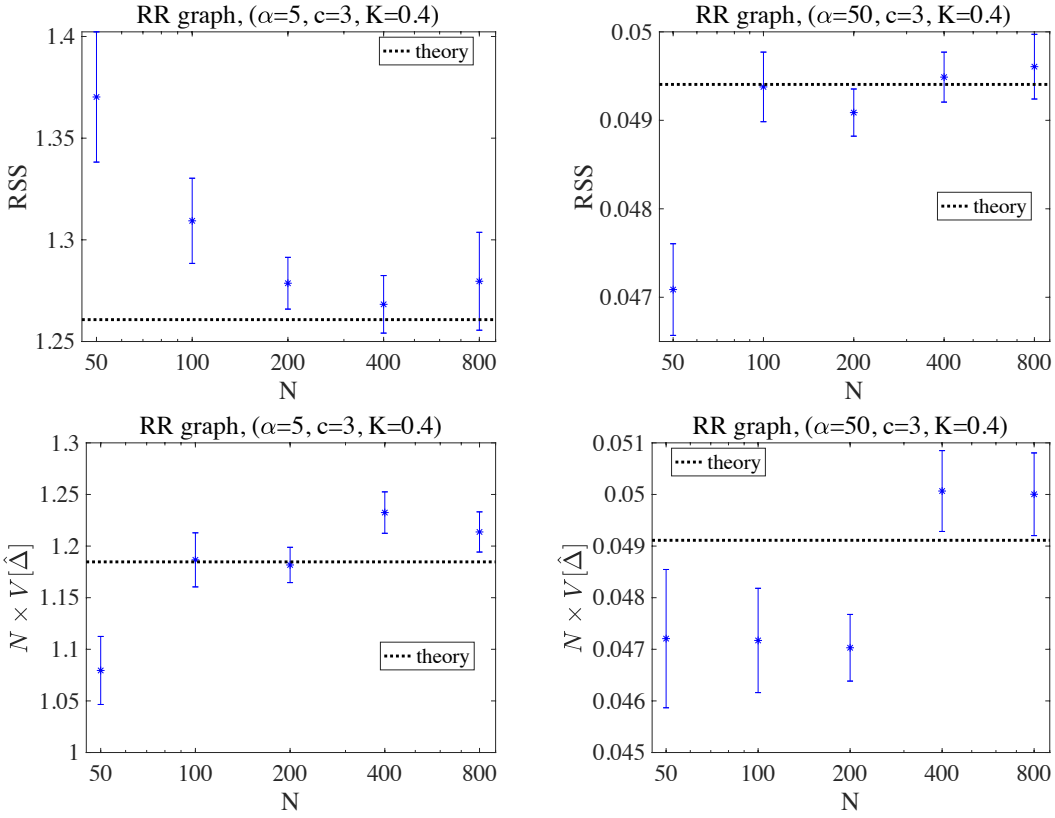


FIGURE 1.8: Top: Plot of $\mathcal{E}$ against the system size N for $\alpha = 5$ (left) and 50 (right) at $(K, c) = (0.4, 3)$. The black dotted lines denote the theoretical result and the markers are the numerical ones. The numerical results tend to converge with the theoretical results as the system size grows, although the finite size effect seems to be different between the left and right panels. The error bar is obtained from $N_{\rm set} = 500$ datasets for $N = 50$–200, $N_{\rm set} = 400$ for $N = 400$, and $N_{\rm set} = 50$ for $N = 800$. Bottom: The rescaled variance (multiplied by N) of the noise part $\hat{\mathbf{\Delta}} = \hat{\mathbf{J}} - \bar{\mathbf{J}}$ is plotted against the system size N. Parameters are the same as those of their counterparts in the upper panels. Although in this closeup scale there is a small gap between the numerical and theoretical results within the one standard error, this gap can be eliminated by taking a larger number of samples. Here, the error bar was obtained using the bootstrap method by considering each realization and component of $\hat{\mathbf{\Delta}}$ as i.i.d..

as

$$\hat{J}_i \sim \mathcal{N}\left(0, \frac{\sigma_i^2}{N}\right), \ (\forall i \in \bar{\Omega}), \tag{1.151}$$

where σ_i^2 is the rescaled variance of the estimate, and verifies $(1/N)\sum_{i\in\bar{\Omega}} \sigma_i^2 \approx Q \,\mathrm{Tr}\left(\mathbf{C}^{\backslash \mathbf{0}}\right)^{-1}/N$. Hence, the probability of successfully screening out these estimators on $\bar{\Omega}$ is

$$\prod_{i\in\bar{\Omega}} \mathrm{Prob}\left(|\hat{J}_i| < K_{\rm th}\right) = \prod_{i\in\bar{\Omega}} \left(1 - 2\int_{\sqrt{\frac{N}{\sigma_i^2}}K_{\rm th}}^{\infty} \mathrm{d}z \, \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}\right) \approx \prod_{i\in\bar{\Omega}} \left(1 - \frac{2}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}\frac{N}{\sigma_i^2}K_{\rm th}^2}}{\sqrt{\frac{N}{\sigma_i^2}}K_{\rm th}}\right) \xrightarrow[N\to\infty]{} 1.$$

The second approximate equality comes from the asymptotic formula of the integral, which holds as $N \to \infty$. The last limiting holds as long as σ_i is bounded from above, because the

exponential factor $\exp(-\frac{1}{2}NK_{\rm th}^2/\sigma_i^2)$ decays fast enough compared with the number of products $|\bar{\Omega}| = N - c - 1$. Hence, we can completely suppress the false positives in the limit $N \to \infty$. Meanwhile, we also desire to accurately reproduce the presence of couplings on Ω. This can be done by tuning the threshold value $K_{\rm th}$ to a smaller value than the true coupling strength K (the mean estimates $\bar{\mathbf{J}}$ are larger than K in the absolute value). In practical situations, we do not know the true coupling strength K in advance, and thus it is nontrivial to correctly tune $K_{\rm th}$. In such cases, it may be better to pick $K_{\rm th}$ by monitoring the distribution of estimators as seen in Fig. 1.7, and to find a value that effectively separates the modes of distribution.

ER graph case

For the ER graph with connection probability $p = d/N$, the evaluation of the order parameters and related quantities is slightly more complex than the RR case because of the distributed nature of the connectivity. In the thermodynamic limit, the distribution of connectivity c in the ER graph obeys the Poisson distribution:

$$P_{\rm po}(c|d) = e^{-d}\frac{d^c}{c!}. \tag{1.152}$$

The trace of the inverse correlation function fortunately becomes simple in the limit:

$$\frac{1}{N}\operatorname{Tr}\mathbf{C}^{-1} \xrightarrow[N\to\infty]{} \sum_{c=0}^{\infty}\left(\frac{c}{1-\tanh^2 K} - c + 1\right)P_{\rm po}(c|d) = \left(\frac{d}{1-\tanh^2 K} - d + 1\right). \tag{1.153}$$

When focusing on spin i with connectivity c_i in the ER graph, its associated order parameters are computed by (1.117) with $P(h^*|c_i)$ defined in (1.149), and the RSS is given by

$$\mathcal{E}_i(c_i) = (1-\hat{b}(c_i))^2K^2 + Q(c_i)\left(\frac{d}{1-\tanh^2 K} - d + 1\right). \tag{1.154}$$

This explicit dependence of the order parameter on c_i is the complex point of the ER case. The mean RSS for the whole network then reads

$$\mathcal{E}_{\rm mean} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}_i \xrightarrow{N\to\infty} \sum_{c=0}^{\infty}\left\{(1-\hat{b}(c))^2K^2 + \left(\frac{d}{1-\tanh^2 K} - d + 1\right)Q(c)\right\}P_{\rm po}(c|d). \tag{1.155}$$

As an interesting departure from the RR case, we here examine the connectivity dependence of our quantities of interest. The plots of $\mathcal{E}(c), Q(c)$, and $\hat{b}(c)$ at $(N, \alpha, d, K) = (400, 10, 4, 0.4)$ are given in Fig. 1.9. In this experiment, we generated ten different ER networks, performed two independent MC samplings, and conducted learning for all $i = 1, \cdots, N$. The error bars were placed using the obtained datasets, and $N_{\rm set}$ varied depending on the connectivity c. The agreement between the theoretical and numerical results is fairly good. Although a slight deviation at large c in $\mathcal{E}(c)$ and $Q(c)$ was observed, it is presumably attributed to the finite size effect, which increased at large c as system size became insufficient to generate nodes with large c. We have tried to control this deviation but found it is difficult to conduct experiments of sufficiently large systems in reasonable time: The generation probability of node with, say, $c = 13$ can be estimated as $P_{\rm po}(13|4) \approx 2 \times 10^{-4}$, and hence for stably generating networks with such large degree nodes we need at least $N \approx 5000$, which is too much in our experiment.

We also computed the mean RSS (1.155) for the whole network. The theoretical value is $\mathcal{E}_{\rm mean} = 0.4780$, while the present experimental value is $\mathcal{E}_{\rm mean} = 0.4907 \pm 0.0041$. The slight difference between these is again attributed to the finite size effect. Here, the theoretical value was obtained by taking the sum of (1.155) up to $c = 20$; the effect of this truncation was found to be small.

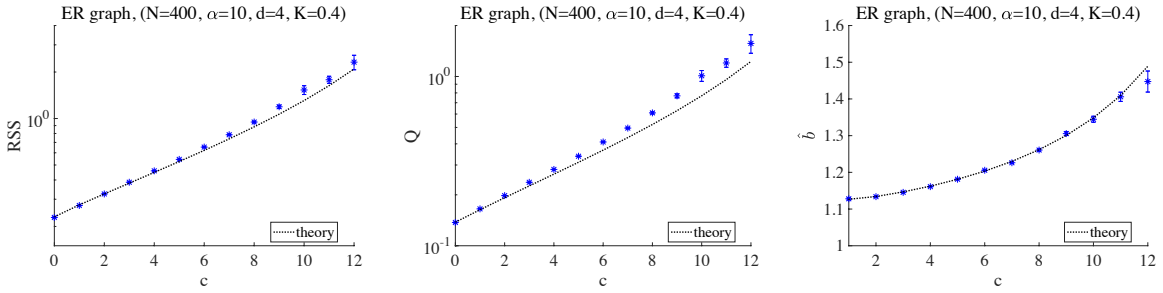


FIGURE 1.9: Plots of $\mathcal{E}$ (left), Q (middle), and $\hat{b}$ (right) against c at $(N, \alpha, d) = (400, 10, 4)$. Black dotted lines and markers are the theoretical and numerical values, respectively; the different colors correspond to different K. The agreement between them is fairly good.

Square lattice case for comparison

We have seen in 1.3.4 that our ansatz holds for tree-like graphs. Usually, carrying out a cavity method in direct problems is known to yield good approximations even for loopy graphs, as long as correlations among spins are weak; it is sometimes referred to as Bethe approximation and will be mentioned in 2.1.3. Here, we compare our theoretical result for $c = 4$ with the simulation result on the square lattice (hence a non tree-like network) with periodic boundary condition. To avoid possible complexity due to frustration, the present teacher couplings were assumed to be all positive and constant, $J_i^* = K > 0, (i \in \Omega)$.

In Fig. 1.10, we plotted $\mathcal{E}$ and $\hat{b}$ against α for $K = 0.2$ on the square lattice of size 20×20, in comparison with our theoretical result (dotted line) computed with the assumption of the tree-like network structure. The agreement between the theoretical and numerical results is very good, which suggests that our theoretical result can be a good approximation even for loopy graphs.

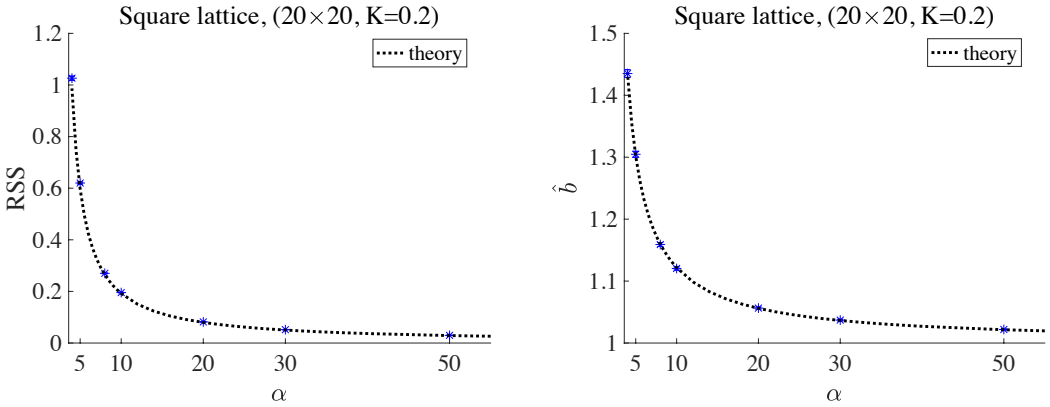


FIGURE 1.10: Plots of $\mathcal{E}$ (left) and $\hat{b}$ (right) against α for $K = 0.2$ on the square lattice of size 20×20. For comparison, the theoretical results derived by assuming the tree-like structure of the coupling network are plotted as the dotted lines. The agreement between the markers (numerical results) and lines is fairly good. The error bars obtained from $N_{\text{set}} = 400$ datasets are shown.

Another interesting phenomenon for loopy graphs is the possible presence of bias in the estimated couplings for spins in $\bar{\Omega}$, as discussed in 1.3.4. To examine this, we show in the upper panels of Fig. 1.11 the distributions of the coupling estimates corresponding to the next nearest neighbors (NNN) from the center spin S_0 in the teacher model for the square lattice (left) and for the RR graph with $c = 4$ (right). To make a fair comparison, the present teacher couplings for the RR graph case are all positive and constant as the square lattice case. These two distributions are very similar, implying that the bias in coupling estimates for remote spins is, even if it exists in loopy graphs, very weak for the present situation. For further quantitative information, the means of those distributions were plotted against the system size in the lower panels. Again, we observed no clear deviation from zero and no significant difference between the two cases of the square lattice and RR graph. These suggest the practicality of the theoretical results for wider situations than tree-like networks.

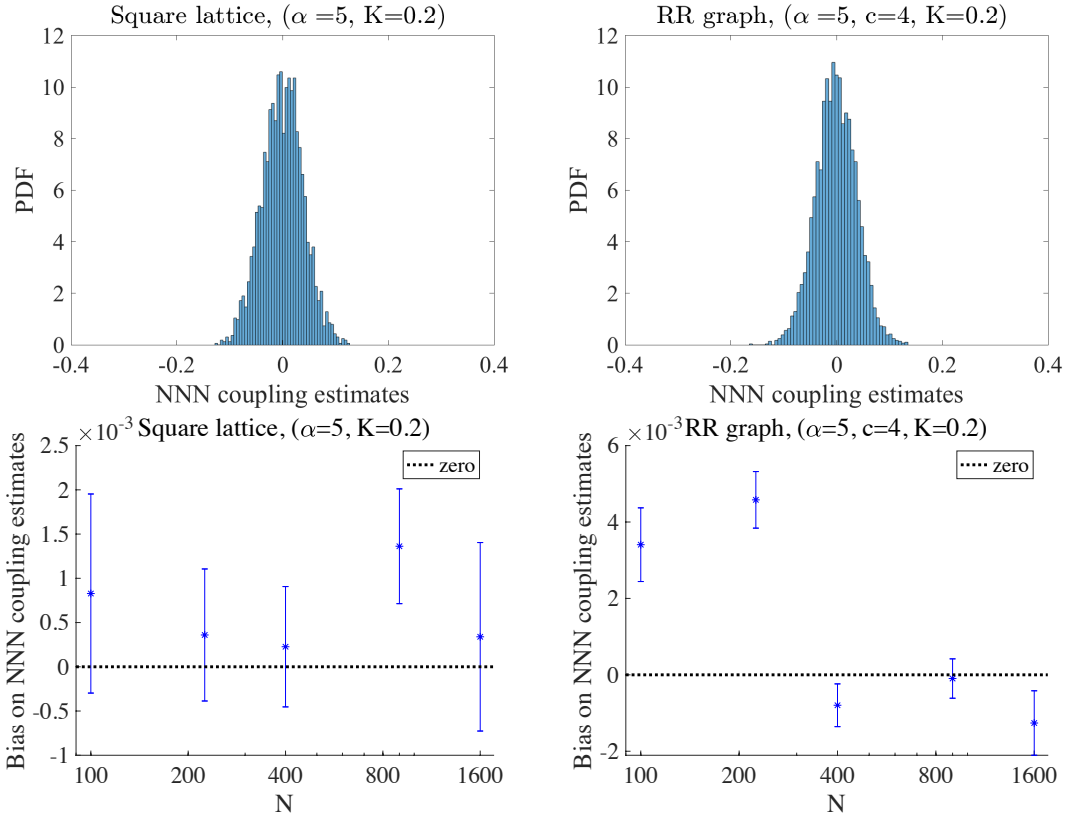


FIGURE 1.11: Top: Distributions of the NNN estimators for the 20×20 square lattice (left) and for the RR graph with $(N, c) = (400, 4)$. In both cases, other parameters are set to be $(K, \alpha) = (0.2, 5)$ and $N_{\rm set} = 400$. No clear positive/negative tendency is observed in both cases. Bottom: Plots of the mean of the NNN estimate distribution against the system size for the square lattice (left) and RR graph (right). Other parameters are similar to those of the corresponding upper panels. The means are quite small, and no clear deviation from zero is observed. The dataset sizes are $N_{\rm set} = 600, 600, 400, 200, 40$ for $N = 100, 225, 400, 900, 1600$, respectively. The error bars are obtained using the bootstrap method.

Comparison with interaction screening

We now examine the interaction screening (IS) cost function [172, 94], which is another common method for the inverse Ising problem, and would like to compare it with the PL method. The

IS cost function is given by

$$\ell^{\mathrm{IS}}(x) = e^{-x}. \tag{1.156}$$

In Fig. 1.12, we plot $\mathcal{E}$, Q, and $\hat{b}$ against α for the RR graph at $(c, K) = (3, 0.4)$, with two theoretical curves of IS (dashed line) and PL (dotted line). Numerical results are also shown to validate the theoretical result of the IS case. The important observation is that the IS result

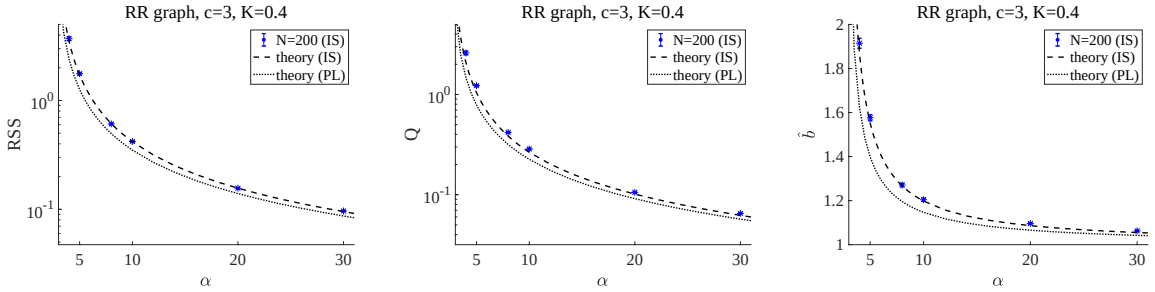


FIGURE 1.12: Comparison of the PL and IS results. Plots of $\mathcal{E}$ (left), Q (middle), and $\hat{b}$ (right) against α at $(c, K) = (3, 0.4)$ for the RR graph are given. The theoretical curves for PL and IS are described by dotted and dashed lines, respectively. The color markers are the numerical results for the IS case at $N = 200$, whose error bars are obtained from $N_{\mathrm{set}} = 100$ datasets. The consistency of the numerical and theoretical results is fairly good. The IS result provides larger values in all these three quantities than those of PL, implying that IS gives larger error, variance, and bias.

consistently gives larger values of $\mathcal{E}$, Q, $\hat{b}$ than those of PL. This implies that the IS method gives larger error, variance, and bias than PL. Although it is known that IS requires a near-optimal number of samples with respect to the dataset size M [172, 94, 139] for perfect reconstruction of the coupling network structure, it does not necessarily mean it holds better accuracy in terms of variance/bias.

1.3.6 Discussion and limits

The crucial assumptions of our treatment are the asymptotic behavior of the estimator (1.102) and the paramagnet assumption of the teacher model, leading to the decoupled distributions of the cavity fields. The former assumption implies that the teacher's couplings can be reconstructed by the student almost perfectly, as discussed at the end of 1.3.5. However, it requires the smallness of coupling strength, implying that strongly-correlated datasets cannot be treated by the proposed theory.

As for perfect reconstruction of the sparse network in the inverse Ising framework, earlier studies reported similar results in empirical and theoretical ways [40, 139, 174, 133]. In particular, a series of analyses by Wainwright and the collaborators [139, 174, 133] derived the necessary and sufficient conditions for the perfect reconstruction in the asymptotic limit, clarifying that the necessary size of the dataset scales as $M = \mathcal{O}(\log N)$ when the maximum degree of the network is bounded from above. Compared to this scaling, our result on the scaling $M = \mathcal{O}(N)$ is rather conservative. Our formulation, however, has some nontrivial advantages by deriving more detailed information about the system. For example, it can deal with the ER graph, whose maximum degree is not bounded, and for which the proofs established in [139, 174, 133] are not applicable. By directly assessing the estimator's fluctuation, our result also clarifies that hypothesis testing can actually achieve perfect reconstruction, which provides another efficient way of

reconstruction than the ℓ_1 regularization used in earlier studies. The explicit computation of the bias on the estimator is also another perk of our approach. In this way, the present formulation can provide finer information.

For handling real-world datasets, finite magnetizations as well as possible loop structures in the network should be taken into account. For such realistic situations, the computation of $(1/N)\,\mathrm{Tr}\,\mathbf{C}^{-1}$ and $P_{\mathrm{cav}}(h^*|\mathbf{J}^*)$ will be more complicated. To evaluate those quantities, advanced techniques such as Bethe approximation [135, 116], high-temperature expansion, and MC samplings will be useful. The ansatz (1.102) should also be modified for the case of loopy graphs, as discussed in 1.3.4. The presented result can be still practical as an approximation for treating such situations, as demonstrated in 1.3.5.

A clear drawback of the estimator treated in this paper is that it is not informative in the region $\alpha \leq 2$, as indicated by the divergent RSS in the limit $\alpha \to 2^+$ shown in 1.3.5 and [12]. To overcome this, the use of regularizations might be promising. The ℓ_1 regularization will be particularly useful to control false positives in the estimated couplings. It is also possible to employ hypothesis testing in conjunction with ℓ_2 regularization [182].

Another interesting extension of the present analysis might be the model-mismatched cases where the student model cannot be equal to the teacher one. Even in such cases, some limited information in the teacher, such as the coupling network structure, might be recovered in some conditions [162]. Pursuing this possibility could provide a better justification for applications of the inverse Ising framework to the analysis of real-world datasets.

Summary of Chapter 1 In this first chapter, we briefly described the Sherrington-Kirkpatrick spin glass model and went over the replica calculation providing a glimpse of its key steps. In particular, we have seen that it relies on several non-rigorous steps and it implies picking an ansatz. In some cases, the replica-symmetric ansatz is correct, but we might have to introduce replica symmetry to describe models with ergodicity breaking. After introducing the framework of Bayesian statistical inference, we proposed a theory to evaluate the reconstruction performance in inverse Ising problems with sparse couplings. To do this, we turned to the pseudo likelihood estimator. A large part of the theory relies on the statistical mechanical formulation in [12], but we refined the theoretical treatment in the cavity method to handle the teacher model with sparse couplings. The resulting expression requires a full functional form of the cavity field distribution, which is far from Gaussian but was obtained by appropriate consideration of the direct problem counterpart. Our theoretical result shows good agreement with numerical experiments conducted on the RR and ER graphs. This agreement holds even for the case of the square lattice, suggesting the practicality of the present result as an approximation for loopy graphs.

Chapter 2

Universality of phase transitions in noiseless linear regression

2.1 Belief propagation

2.1.1 Factor graphs

We consider a graphical model with N variables $\mathbf{x} = (x_1, ..., x_N)$, taking values on a finite set, called the alphabet $\mathcal{X}$. We would like to compute interesting quantities on this model, for instance joint distributions, or conditional distributions of some of its variables. To do so, it is useful to first compute marginal distributions of this model. Naively, it would require summing over all possible configurations, i.e. a number of operations of order $|\mathcal{X}|^N$: the computational cost is exponential and will be unpractical when N becomes large. Belief propagation will help us bypass this issue, through a set of equations which allows us to obtain marginals with a cost of order N. To begin, let us introduce the general framework: we consider a finite bipartite graph with two type of vertices.

- The variable set V includes *variable nodes* $i \in V$, which are symbolized by round vertices.
- The function set F includes *function nodes* $a \in F$, which are symbolized by square vertices.

∂i denotes the function nodes which share an edge with i, while ∂a is the set of variable nodes connected with a. E is the set of edges of the graph. For each function node, we have a function $\psi_a : \mathcal{X}^{|\partial a|} \to \mathbb{R}^+$. The ensemble $\Psi = (V, F, E, \{\psi_a\}_{a \in F})$ is a *factor graph*. On such a graph, we define a probability measure

$$P_\Psi(\mathbf{x}) \equiv \frac{1}{\mathcal{Z}_\Psi} \prod_{a \in F} \psi_a(\mathbf{x}_{\partial a}) \tag{2.1}$$

where $\mathbf{x}_{\partial a}$ is the restriction of variable vector $\mathbf{x}$ to the set ∂a. The normalizing function reads

$$\mathcal{Z}_\Psi = \sum_{\mathbf{x} \in \mathcal{X}^{|V|}} \prod_{a \in F} \psi_a(\mathbf{x}_{\partial a}). \tag{2.2}$$

The one-dimensional marginalized distributions we are interested in read for $i \in V$

$$C_\Psi^{[i]}(x) = \sum_{\mathbf{x}, x_i = x} \prod_{a \in F} \psi_a(\mathbf{x}_{\partial a}). \tag{2.3}$$

2.1.2 An easy example: the Ising chain

To gently dive into the techniques from belief propagation, let us focus for a moment on the one-dimensional ferromagnetic Ising chain, with a constant external field H. The variable are $\mathbf{S} = (S_1, ...S_N)$ which take values in $\mathcal{X} = \{-1, +1\}$. The Boltzmann distribution reads

$$P_{\mathrm{IC}} = \frac{1}{\mathcal{Z}_{\mathrm{IC}}} e^{-\beta \mathcal{H}_{\mathrm{IC}}(\mathbf{S})} \tag{2.4}$$

with β the inverse temperature, and

$$\mathcal{H}_{\mathrm{IC}}(\mathbf{S}) = -\sum_{i=1}^{N-1} S_i S_{i+1} - H \sum_{i=1}^{N} S_i. \tag{2.5}$$

We are interested in the marginal distribution $P_j(S_j)$. It is made up by several contributions, coming from each node that is connected to the variable node S_j on the factor graph: one contribution comes from variable node S_{j-1}, one from variable node S_{j+1}, and one from the function node which models the external field H applied to S_j. We define the following quantities:

$$\hat{\nu}_{\to j}(S_j) = \frac{1}{Z_{\to j}} \sum_{S_1,\dots,S_{j-1}} \exp\left\{ \beta \sum_{i=1}^{j-1} S_i S_{i+1} + \beta H \sum_{i=1}^{j-1} S_i \right\} \tag{2.6a}$$

$$\hat{\nu}_{j\leftarrow}(S_j) = \frac{1}{Z_{j\leftarrow}} \sum_{S_{j+1},\dots,S_N} \exp\left\{ \beta \sum_{i=j}^{N} S_i S_{i+1} + \beta H \sum_{i=j+1}^{N} S_i \right\}. \tag{2.6b}$$

$\hat{\nu}_{\to j}$ and $\hat{\nu}_{j\leftarrow}$ are probability distribution and are thus normalized by $Z_{\to j}$ and $Z_{j\leftarrow}$. From there, we can write the marginal distribution $P_j(S_j)$, as

$$P_j(S_j) \cong \hat{\nu}_{\to j}(S_j)\, e^{\beta H S_j}\, \hat{\nu}_{j\leftarrow}(S_j) \tag{2.7}$$

where $\cong$ denotes equality up to a normalization factor. $\hat{\nu}_{\to j}$ and $\hat{\nu}_{j\leftarrow}$ are called *messages*, they correspond to the marginal distribution of S_j in a model where we have removed all nodes on the right of S_j and all nodes on the left, respectively. It is straightforward to see that messages relate to each other through

$$\hat{\nu}_{\to i+1}(S_{i+1}) \cong \sum_{S_i} \hat{\nu}_{\to i}(S_i) e^{\beta S_i S_{i+1} + \beta H S_i} \tag{2.8a}$$

$$\hat{\nu}_{i-1\leftarrow}(S_{i-1}) \cong \sum_{S_i} \hat{\nu}_{i\leftarrow}(S_i) e^{\beta S_{i-1} S_i + \beta H S_i}. \tag{2.8b}$$

Note that $\hat{\nu}_{\to 1}(\pm 1) = \hat{\nu}_{N\leftarrow}(\pm 1) = \frac{1}{2}$. Combining this with (2.8), we can compute all messages iteratively, with a number of iterations of order N, then obtain marginal distributions through (2.7).

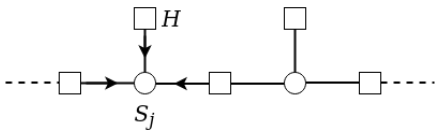


FIGURE 2.1: Factor graph of the one-dimensional Ising chain. The circles are variable nodes i.e. spins, the squares are function nodes. On top, they represent the external field's action. The spin S_j is represented with the three messages it receives: $e^{\beta H S_j}$, $\hat{\nu}_{\to j}$ and $\hat{\nu}_{j\leftarrow}$.

2.1.3 BP on tree-like networks

BP equations

In a similar fashion, we will define messages along the edges of the factor graph, focusing on those which are *tree-like*. For each edge (i, a) between a variable node and a factor node, there are two messages $\nu_{i\to a}(x_i)$ and $\hat{\nu}_{a\to i}(x_i)$, which are both probability distributions. More precisely,

$\hat{\nu}_{i\to a}(x_i)$ coincides with the single variable x_i marginal $C^{[i]}_{\Psi_{i\to a}}$ in the modified graphical model $\Psi_{i\to a}$ which does not include factor a. $\nu_{i\to a}(x_i)$ is the distribution of x_i in the graphical model $\Psi_{a\to i}$ where all factors in ∂i except a have been discarded. Belief propagation (BP) is an update rule of these messages which reads, up to normalization factors:

$$\nu^{(t+1)}_{i\to a}(x_i) \cong \prod_{b\in\partial i\backslash a} \hat{\nu}^{(t)}_{b\to i}(x_i) \tag{2.9a}$$

$$\hat{\nu}^{(t)}_{a\to i}(x_i) \cong \sum_{\mathbf{x}_{\partial a\backslash i}} \psi_a(\mathbf{x}_{\partial a}) \prod_{j\in\partial a\backslash i} \nu^{(t)}_{j\to a}(x_j). \tag{2.9b}$$

Note that if $\partial i\backslash a$ is empty, then $\nu_{i\to a}(\pm 1) = \frac{1}{2}$; and if $\partial a\backslash i$ is empty, $\hat{\nu}_{a\to i} = \psi_a$. We thus obtain one equation for each oriented edge of a graph, i.e. a total of $2|E|$ equations on as many messages. At each time, we can also evaluate the estimate $\nu_i(x_i)$ of the marginal distribution $P_i(x_i)$ of variable x_i by summing on all incoming messages:

$$\nu^{(t)}_i(x_i) \cong \prod_{a\in\partial i} \hat{\nu}^{(t-1)}_{a\to i}(x_i). \tag{2.10}$$

A great fact is the following rigorous result: if D_Ψ is the diameter of Ψ, i.e. the maximal distance between two variable nodes on the factor graph, then BP equations converge for $t > D_\Psi$. We usually say that **BP is exact on trees**. Adding a star subscript to designate the fixed point messages of BP equations, for $t > D_\Psi$ and for any variable node i:

$$\nu^{(t)}_{i\to a} = \nu^{*}_{i\to a} \tag{2.11}$$

$$\hat{\nu}^{(t)}_{a\to i} = \hat{\nu}^{*}_{a\to i} \tag{2.12}$$

$$\nu^{(t)}_i(x_i) = P_i(x_i). \tag{2.13}$$

The first appearance of these equations was in the context of coding theory, but they were also found in statistical physics (in relation to the Bethe approximation), and in the artificial intelligence community. An overview of BP equations in the light of statistical physics can be found in [105], while [183] provides an educational approach.

Simplification for pairwise models

In the following parts, we will often be interested in pairwise models i.e. graphical models where all factor nodes have degree 2. A pairwise model can be represented by a graph $G = (V, E)$ over variable nodes. The edge (i, j) joins variables i and j which are the arguments of a function ψ_{ij}, and function nodes correspond to edges. The corresponding probability distribution is

$$P_{\text{pair}}(\mathbf{x}) = \frac{1}{\mathcal{Z}_{\text{pair}}} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j). \tag{2.14}$$

Besides, the BP equations reduce to only messages $\nu_{i\to(i,j)}$, that we rename $\nu_{i\to j}$ and become

$$\nu^{(t+1)}_{i\to j} \cong \prod_{k\in\partial i\backslash j} \sum_{x_k} \psi_{ik}(x_i, x_k)\nu^{(t)}_{k\to i}(x_k). \tag{2.15}$$

Useful quantities

BP equations do not only provide marginal distributions, they also allow us to compute other useful quantities.

- Joint distribution of a subset of variables

Let F_R be a subset of function nodes, V_R a subset of variable nodes adjacent to F_R. R is the induced subgraph that we assume to be connected. Let ∂R be the subset of function nodes which are adjacent to a variable node from V_R, but are not in F_R. For a function $a \in \partial R$, there exists a unique variable node i from V_R connected to it, we will denote it $i(a)$. Then the joint distribution of $\mathbf{x}_R$, the restriction of $\mathbf{x}$ to V_R variables, reads

$$P(\mathbf{x}_R) = \frac{1}{Z_R} \prod_{a \in F_R} \psi_a(\mathbf{x}_{\partial a}) \prod_{a \in \partial R} \hat{\nu}^*_{a \to i(a)}(x_{i(a)}). \tag{2.16}$$

- Joint probability distribution of all variables

We can also express the joint probability distribution of the complete set of variables in terms of marginal probabilities. We call $P_a(\mathbf{x}_{\partial a})$ the marginal probability distribution of all variables involved in function ψ_a. $P_i(x_i)$ is the marginal distribution of variable x_i. Then the joint distribution $P(\mathbf{x})$ on tree graphs can be written in terms of those marginals as

$$P(\mathbf{x}) = \prod_{a \in F} P_a(\mathbf{x}_{\partial a}) \prod_{i \in V} P_i(x_i)^{1-|\partial i|}. \tag{2.17}$$

This can be proven by induction on the number of factors.

- Internal energy

The physicist reader is of course interested in applying this formalism to physical systems, such as the Ising chain mentioned above in 2.1.2. In a physical system, the functions ψ_a read

$$\psi_a(\mathbf{x}_{\partial a}) = e^{-\beta E_a(\mathbf{x}_{\partial a})} \tag{2.18}$$

with β an inverse temperature, and E_a an energy function characterizing the constraint a. For simplicity, let us set β to 1. From there, we define the internal energy, which is the expectation of the total energy:

$$U[P] = -\sum_{\mathbf{x}} P(\mathbf{x}) \sum_{a \in F} \log \psi_a(\mathbf{x}_{\partial a}). \tag{2.19}$$

We will now use (2.16) and apply it to $F_R = \{a\}$ for each $a \in F$. In that case $V_R = \partial a$, and $\prod_{a \in \partial R} = \prod_{i \in \partial a} \prod_{b \in \partial i \setminus a}$, which gives

$$U[P] = -\sum_{a \in F} \sum_{\mathbf{x}_{\partial a}} P(\mathbf{x}_{\partial a}) \log \psi_a(\mathbf{x}_{\partial a}) \tag{2.20}$$

$$= -\sum_{a \in F} \sum_{\mathbf{x}_{\partial a}} \frac{1}{Z_a} \psi_a(\mathbf{x}_{\partial a}) \log \psi_a(\mathbf{x}_{\partial a}) \prod_{i \in \partial a} \prod_{b \in \partial i \setminus a} \hat{\nu}^*_{b \to i}(x_i) \tag{2.21}$$

$$U[P] = -\sum_{a \in F} \frac{1}{Z_a} \sum_{\mathbf{x}_{\partial a}} \psi_a(\mathbf{x}_{\partial a}) \log \psi_a(\mathbf{x}_{\partial a}) \prod_{i \in \partial a} \nu^*_{i \to a}(x_i). \tag{2.22}$$

where we have used (2.9a), and $Z_a = \sum_{\mathbf{x}_{\partial a}} \psi_a(\mathbf{x}_{\partial a}) \prod_{i \in \partial a} \nu^*_{i \to a}(x_i)$

- Entropy of $P(\mathbf{x})$

The entropy of distribution $P(\mathbf{x})$ is defined by

$$H[P] = -\sum_{\mathbf{x}} P(\mathbf{x}) \log P(\mathbf{x}). \tag{2.23}$$

From (2.17), it is straightforward that

$$H[P] = -\sum_{a\in F} P_a(\mathbf{x}_{\partial a}) \log P_a(\mathbf{x}_{\partial a}) - \sum_{i\in V}(1-|\partial i|)P_i(x_i)\log P_i(x_i). \tag{2.24}$$

- Free energy

We can finally consider the free energy $\Phi = -\frac{1}{\beta}\log \mathcal{Z}$, where $\mathcal{Z} = \sum_{\mathbf{x}} \prod_{a\in F} \psi_a(\mathbf{x}_{\partial a})$ is the partition function. The free energy relates to the internal energy and the entropy defined above, such that

$$\Phi[P] = U[P] - H[P] \tag{2.25}$$

$$\Phi[P] = \sum_{a\in F} P_a(\mathbf{x}_{\partial a}) \log \frac{P_a(\mathbf{x}_{\partial a})}{\psi_a(\mathbf{x}_{\partial a})} + \sum_{i\in V}(1-|\partial i|)P_i(x_i)\log P_i(x_i). \tag{2.26}$$

Bethe free energy

Relying on (2.16), we can rewrite the free energy in terms of the BP messages, that we group in one variable containing the $2|E|$ messages $\nu = \{\nu_{i\to a}, \hat{\nu}_{a\to i} \ | i \in V, a \in \partial i\}$,

$$\Phi[\nu] = -\sum_{a\in F}\Phi_a(\nu) - \sum_{i\in V}\Phi_i(\nu) + \sum_{(ia)\in E}\Phi_{ia}(\nu), \tag{2.27}$$

where

$$\Phi_a(\nu) = \log\left\{\sum_{\mathbf{x}_{\partial a}} \psi_a(\mathbf{x}_{\partial a}) \prod_{i\in\partial a} \nu_{i\ \to a}(x_i)\right\} \tag{2.28}$$

$$\Phi_i(\nu) = \log\left\{\sum_{x_i} \prod_{b\in\partial i} \hat{\nu}_{b\to i}(x_i)\right\} \tag{2.29}$$

$$\Phi_{(ia)}(\nu) = \log\left\{\sum_{x_i} \nu_{i\ \to a}(x_i)\hat{\nu}_{a\to i}(x_i)\right\}. \tag{2.30}$$

$\Phi[\nu]$ is called the *Bethe free energy* and it is exact on tree graphs. If ν^* is the ensemble of all marginals at the fixed point of BP equations, we thus have $\Phi = \Phi[\nu^*]$.

2.1.4 BP on loopy graphs

We have seen that belief propagation is an exact and remarkable tool for tree-like networks, and provides several types of information. We would like to have something similar on more complicated networks. However, the tree structure is heavily used in the calculations above. For instance for pairwise models, the BP equation (2.15) relies on the following idea: if you consider the modified graphical model where we remove the function ψ_{ij}, then variable x_j becomes uncorrelated with other variables $\{x_k, k \in \partial i\backslash j\}$. In other words, because of the tree structure, the only "link" x_j had with other neighbours of x_i was through ψ_{ij}. The marginal distribution for x_j in this model is thus the same as its marginal distribution in another model where we remove all ψ_{il} for $l \in \partial i$, i.e. where all neighbors of x_i are uncorrelated variables (see Fig. 2.2). This equality allows to simplify marginal computations and to decouple independent terms. Of course, it does not work for any network: the logic breaks down if x_j is connected to other variables $\{x_l, l \in \partial i\backslash j\}$ through another variable node than x_i.

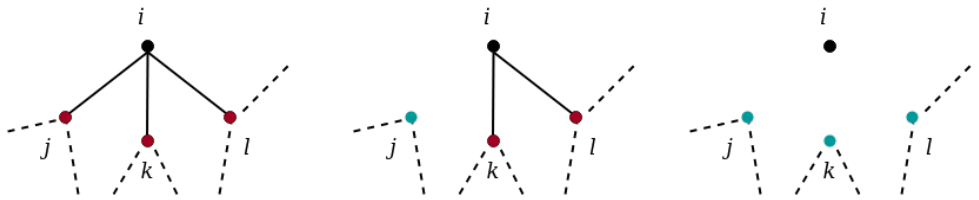


FIGURE 2.2: A pairwise tree model. Left: Complete graphical model focused on variable i and its neighbors. Middle: modified graphical model where we have removed function ψ_{ij}. Right: modified graphical model where we have removed all functions ψ_{il} for $l \in \partial i$. Due to the tree structure, the marginal distribution for variable j is the same in the middle and right models.

For general models, using BP means filling some voids. We would need to find a set of functions that play the roles of messages $\nu_{i\to a}$, $\hat{\nu}_{a\to i}$. We would like them to satisfy BP equations, which should also converge to an (existing? unique?) fixed point. We will need to define possible marginals, i.e. a collection of functions b_i over $\mathcal{X}$ for $i \in V$, and b_a over $\mathcal{X}^{|\partial a|}$ for $a \in F$. Since they need to mimic marginals, those functions should satisfy:

$$\sum_{x_i} b_i(x_i) = 1 \quad \forall i \in V \tag{2.31}$$

$$\sum_{\mathbf{x}_{\partial a}} b_a(\mathbf{x}_{\partial a}) = 1 \quad \forall a \in F \tag{2.32}$$

$$\sum_{\mathbf{x}_{\partial a \backslash i}} \mathbf{x}_{\partial a} = b_i(x_i) \quad \forall a \in F, \forall i \in \partial a \backslash i. \tag{2.33}$$

Note that any joint distribution $P(\mathbf{x})$ would provide such functions, however it is not true the other way round: a well-defined set of functions $\{b_i, b_a\}$ does not necessarily correspond to a distribution. For this reason, we say that we are using *beliefs*; the messages simply describe what could be actual marginals.

Another way to address BP on loopy graphs is through the variational approach, using the Bethe free energy defined in (2.27). We have seen that it is exact on trees, and our hope would be that it provides a good approximation of the actual free energy on more complicated graphs. The Bethe free energy can be defined as a function on a set of possible marginals $\{b_i, b_a\}$. It shares a close bond with BP equations, through the following property: the stationary points of the Bethe free energy, where the energy is finite, correspond to fixed points of BP equations. In fact, in tree graphs the Bethe free energy is convex, hence it has a unique stationary point which provides the unique fixed point of BP equations, but this unicity is not true in general.

In practice, BP can be very effective even on loopy graphs. In particular, note that BP is a local algorithm: outgoing messages on a given node are updated as functions of incoming messages received at the previous iteration. BP will bear fruit if the graph is *locally tree-like* in the asymptotic limit (for a large enough graph, starting from any node and for any finite distance d, the part of the factor graph which is at distance at most d from the center node is a tree). Another case of graphs where BP provides remarkable results is the case of densely-connected graphs (which are not locally tree-like!), where correlations are small enough, so that local loops in the network have negligible contribution. In the following, we will see that BP equations in such settings can sometimes be simplified, in connection with statistical physics methods already used in Chapter 1. We will deal with a more general class of *message passing* algorithms. Similar to BP, they can be defined on a factor graph, and also involve messages (often marginals) on its directed edges, as well as a local update rule of the messages at the nodes.

2.2 Theoretical background on linear regression

2.2.1 Introduction to linear regression

Let us cast belief propagation aside for a moment, and introduce the problem of linear regression. Say that you manage an online library, and would like to issue targeted book recommendations to your clients. You already have a fair amount of clients, and you ask them to rate each book they read. You also ask them for some information when they create an account: their age, address, profession, and a few other elements, in total N. Based on this data, you would like to know whether a particular book would be a good recommendation for a new client or not. Putting all the ratings for this book from M previous users in a vector $\mathbf{y} \in \mathbb{R}^M$, and the information about them in a matrix $\mathbf{F} \in \mathbb{R}^{M \times N}$ (one line per user, one column per information), you would like to find a vector $x \in \mathbb{R}^N$ such that

$$\mathbf{y} = \mathbf{F}\mathbf{x}. \tag{2.34}$$

If we can find such a vector $\mathbf{x}$, it would contain coefficients to express the rating of the book as a weighted sum of the various pieces of information on the reader. If you get a new client, you can then simply add a line to the matrix $\mathbf{F}$, and obtain an estimate of this book's rating for that client. Hence you would have an idea of this client's susceptibility to enjoy the book. At first glance,

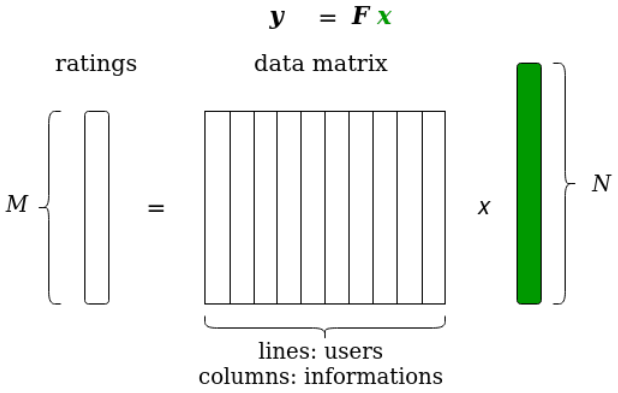


FIGURE 2.3: Linear regression to obtain ratings as a linear combination of elements from the data matrix lines.

this does not seem like a very efficient method. Indeed, having only a few pieces of information on each user might not be enough to actually explain their rating of the book. However, we now live in the *big data* era: we are overwhelmed by information, and we should design techniques and algorithms that take into account the large amount of accessible data to better extract information from it. In particular, the online library manager now does not only know 5 or 6 elements about each user: a simple research on google might reveal a social media account, and a real avalanche of data. Our linear regression problem thus becomes relevant in the asymptotic regime that we are used to: M, N are very large. Linear problems might seem too simple and somewhat out-of-date, however even they hold a fair share of mystery and delicate questions. In particular, when $M < N$, the system is undetermined: there are more unknowns than measurements, and the regression becomes a challenging task. These problems appear in many fields: signal processing where we call it *compressed sensing*, experimental physics, biology... Seminal work on techniques to solve the problem computationally can be found in [44, 30]. An example of brilliant and exciting application is magnetic resonance imaging (MRI) [95, 96]. The same set-up used to evaluate book ratings could be exploited to store personal medical information, and explain which factors influence how likely a person is to contract a disease.

2.2.2 Sparse linear regression

Why sparse?

Coming back to library management, say that you now hacked into your users' social media accounts[1], and your data matrix has an enormous number of columns. Of course, most of them are irrelevant, hence the vector $\mathbf{x}$ needs to be *sparse*: most of its components are null. Allowing

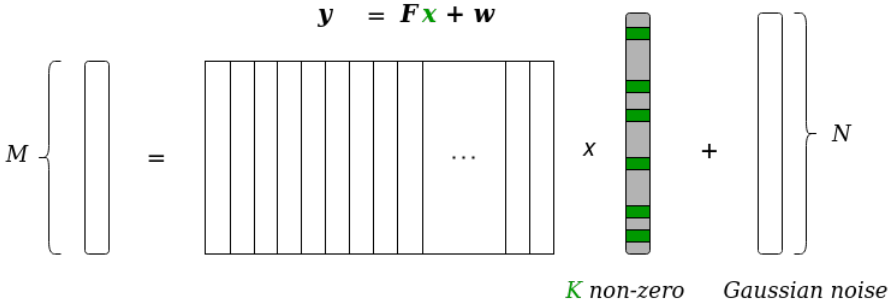


FIGURE 2.4: Sparse linear regression: only K out of N elements of $\mathbf{x}$ are non zero, which selects K columns from the data matrix $\mathbf{F}$.

for some noise, our problem is now phrased this way: we know $\mathbf{y} \in \mathbb{R}^M$ and a data matrix (also called the sensing matrix) $\mathbf{F} \in \mathbb{R}^{M \times N}$, and we assume that $\mathbf{y}$ is generated through

$$\mathbf{y} = \mathbf{F}\mathbf{x_0} + \mathbf{w} \tag{2.35}$$

with $\mathbf{w} \sim \mathcal{N}(0, \Delta_0)^N$ a Gaussian noise. We would like to find an estimate $\mathbf{x} \in \mathbb{R}^N$, which reconstructs $\mathbf{x_0}$ and has only K non-zero components. We fall into the teacher-student scenario, and the problem is hence characterized by two ratios: $\alpha \equiv M/N$, and $\rho \equiv K/N$. We are interested in the asymptotic regime: $K, M, N \to \infty$, with α, ρ of order 1.

Depending on the context of the problem, matrix $\mathbf{F}$ can have very different structures. Of course, we would like to have a theory as general as possible and we might be interested in some specific forms. However, we need to accept a trade-off between generalization and specialization, and to start with easier types of matrices to analyze, in the hope of gaining more understanding of such questions. In the following, we will take $\mathbf{F}$ a random matrix, sampled from a given distribution. Several questions arise:

- When do we have enough measurements to reconstruct the signal? We first rule out the region where we have less measurements than unknowns: information-theoretically, it is impossible to find $\mathbf{x}$ if $\alpha < \rho$.
- In the region $\alpha > \rho$, where can we estimate $\mathbf{x}$, and how can we do it with an algorithm in polynomial time?

Bayes-optimal versus ℓ_1 reconstruction

We measure the performance of an estimator through the mean squared error with respect to the ground truth. Two theoretical settings can be considered:

i) The teacher-student Bayes-optimal scenario, in which we know that $\mathbf{x_0}$ comes from a given distribution p_{x_0}, i.e. we know the statistical property of the signal that we are trying to estimate.

[1] not recommended

ii) In a more general mismatched setting, we do not know anything about $\mathbf{x_0}$, and we only try to find an estimate which is as sparse as possible, to reduce the number of significant parameters.

Our ideal target would be to find

$$\hat{\mathbf{x}}_\mathbf{0} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2, \text{ such that } \hat{\mathbf{x}}_\mathbf{0} \text{ is the most sparse possible.} \tag{2.36}$$

We would like to minimize the so-called ℓ_0 *norm* (which is not a norm!), which counts the number of non zero components of the estimator. However, there is no way to do so in less than exponential time, as we would have to try all possible positions of non zero elements. Hence we turn to the ℓ_1 *norm* $\|\cdot\|_1$ defined by $\|\mathbf{x}\|_1 = \sum_{i=1}^N x_i$. We thus define the estimator

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \|\mathbf{x}\|_1 \right\}. \tag{2.37}$$

We have incorporated the ℓ_1 norm inside the function to minimize, as a *penalty function*. This linear regression with ℓ_1 norm is called the *LASSO* (which stands for Least absolute shrinkage and selection operator). This approach was suggested in a line of work from [29, 28]. The ℓ_1 norm is a convex function, so estimator $\mathbf{x}$ can be found using a convex optimization algorithm. But why would we pick the ℓ_1 norm instead of, say, the usual ℓ_2 norm which is also convex? The answer lies in the geometry of those norms. We usually say that the ℓ_1 penalty favors sparsity: minimizing it will tend to select solutions which are actually sparse, more than other norms, as shown in Fig. 2.5. A refined analysis on sparse estimators and their interpretation as maximum a posteriori (MAP) estimators can be found in [66].

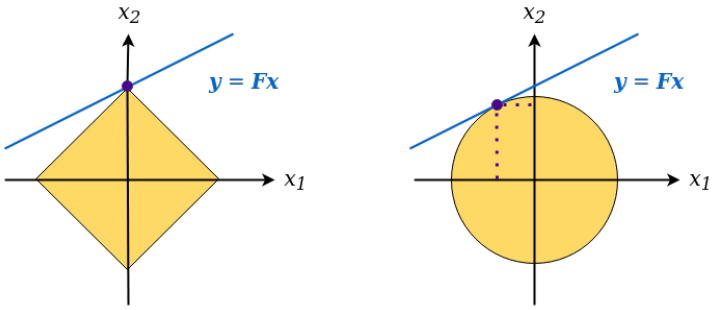


FIGURE 2.5: Linear regression with on $\mathbb{R}^2$. The blue line is the space of solutions to $\mathbf{y} = \mathbf{F}\mathbf{x}$, and we select the point $\hat{\mathbf{x}}$ of coordinates $\hat{x}_1, \hat{x}_2$ in purple. Left: The yellow square is the sphere of vectors with the smallest ℓ_1 norm that still intersects the space of solutions. The estimator satisfies $\hat{x}_1 = 0$. Right: The yellow circle is the sphere of vectors with the smallest ℓ_2 norm that intersects the space of solutions. $\hat{\mathbf{x}}$ has two non zero components.

Two types of matrices

In the rest of the chapter, we will be interested in several types of random matrices. For now, we start with two of them:

- *Gaussian i.i.d. matrices*

$\mathbf{F} \in \mathbb{R}^{M\times N}$ is Gaussian i.i.d. when its elements are independently and identically distributed according to $\mathcal{N}(0, \frac{1}{N})$ the Gaussian distribution of mean zero and variance $\frac{1}{N}$.

- *Rotationally invariant matrices*

This is a much larger class of matrices, that can be defined through the singular value decomposition (SVD) of $\mathbf{F}$, which yields

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V} \tag{2.38}$$

where $\mathbf{U} \in \mathbb{R}^{M\times M}$ and $\mathbf{V} \in \mathbb{R}^{N\times N}$ are orthogonal matrices. $\mathbf{\Sigma} \in \mathbb{R}^{M\times N}$ contains the singular values of $\mathbf{F}$, i.e. the square roots of the eigenvalues of $\mathbf{F}^T\mathbf{F}$. $\mathbf{F}$ is rotationally invariant if the matrices $\mathbf{U}$ and $\mathbf{V}$ have been generated from the Haar measure, i.e. randomly sampled from the space of rotations. This definition might seem a bit mysterious, however note that such matrices can be generated: simply by sampling the orthogonal matrices and choosing the desired spectrum inside $\mathbf{\Sigma}$. Gaussian i.i.d. matrices are included in this class of rotationally invariant matrices, in this case the singular values of $\mathbf{F}$ follow the Marchenko-Pastur law [169]. We will also be interested in *right rotationally invariant matrices*, where only $\mathbf{V}$ needs to be Haar-generated, independently from $\mathbf{U}$ and $\mathbf{\Sigma}$. Rotationally invariant data matrices allow for correlation between data samples, they are thus a significant improvement with respect to i.i.d. matrices.

Replica analysis for i.i.d. and right rotationally invariant matrices

Let us see how to analyze the problem of linear regression in the inference framework introduced in 1.2. We turn to a probabilistic reconstruction analysis, as explained in [88, 87]. We allow for noise $\mathbf{w}$ again, we can later take its zero limit. To obtain an estimate of the signal, we want to sample from the posterior probability distribution, obtained through Bayes' theorem

$$P(\mathbf{x}|\mathbf{F},\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x},\mathbf{F})P(\mathbf{x},\mathbf{F})}{P(\mathbf{y},\mathbf{F})}. \tag{2.39}$$

Let us stop here for a second. In this setting, we assume that we know that $\mathbf{y}$ is generated as a linear product of a random matrix of known distribution and a ground truth signal, plus a Gaussian noise, but we do not necessarily know the true noise variance Δ_0, so we will try to approximate it with a noise of variance Δ. We said before that $P(\mathbf{x},\mathbf{F})$ was the prior distribution, that encapsules our knowledge of $\mathbf{x}$, or what we believe about its distribution. Let us see how to incorporate this in our case.

i) In the Bayes-optimal setting,

we know the signal true probability distribution $p_{x_0}(\mathbf{x_0})$, and we use it as prior distribution $P(\mathbf{x},\mathbf{F})$. The properties of the true signal $\mathbf{x_0}$ will then be described by maximimizing the posterior distribution, which will be done through a replica calculation.

ii) In the mismatched case,

$P(\mathbf{x},\mathbf{F})$ can help in enforcing constraints about $\mathbf{x}$. To generalize our approach to any separable penalty function f, we want to characterize the estimator

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\left\{\frac{1}{2}\|\mathbf{y}-\mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x})\right\}. \tag{2.40}$$

For the LASSO estimation, we take $f = \|\cdot\|_1$. To enforce the penalty role of f, we impose

$$P(\mathbf{x}|\mathbf{F}) \cong e^{-\frac{f(\mathbf{x})}{\Delta}}, \tag{2.41}$$

and this choice will become clear in a few lines. Note that the prior distribution is independent from the matrix $\mathbf{F}$, so we can simply denote it $P(\mathbf{F})$. We rename $P(\mathbf{y},\mathbf{F}) = \mathcal{Z}$ the partition

function, that we obtain by plugging in the likelihood

$$P(\mathbf{y}|\mathbf{x},\mathbf{F}) = \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)^2} \tag{2.42}$$

such that

$$P(\mathbf{x}|\mathbf{y},\mathbf{F}) \cong \prod_{i=1}^{N} P(x_i) \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)^2}. \tag{2.43}$$

We are interested the partition function which reads

$$\mathcal{Z}(\mathbf{y},\mathbf{F}) = \int \prod_{i=1}^{N} dx_i \prod_{i=1}^{N} P(x_i) \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)^2}. \tag{2.44}$$

where $P(x_i) = p_{x_0}(x_i)$ in the Bayes-optimal case, or $P(x_i) = e^{-\frac{f(x_i)}{\Delta}}$ for the mismatched setting: in that case, we can factor out $1/\Delta$ in the exponential terms in front of the quantity $\frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x})$. Taking the noise $\Delta \to 0$ in the student model will successfully make the integral concentrate on the desired estimator $\hat{\mathbf{x}}$.

To proceed with the calculation, we want to compute Φ the free energy averaged on the randomness of the model, i.e. $\mathbf{F}, \mathbf{x_0}, \mathbf{w}$, which can be done through the replica trick

$$\Phi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}[\log \mathcal{Z}] = \lim_{N\to\infty} \frac{1}{N} \lim_{n\to 0} \frac{\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}[\mathcal{Z}^n] - 1}{n}. \tag{2.45}$$

Introducing n replicas of the system, we want the replicated partition function

$$\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(\mathcal{Z}^n) = \int \prod_{i,a} \mathrm{d}x_i^a \prod_{i,a} P(x_i^a) \prod_{\mu} \mathbb{E}_{\mathbf{F},\mathbf{x_0},\mathbf{w}} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\sum_{a=1}^{n}\left(\sum_{i=1}^{N} F_{\mu i} x_{0,i} + w_\mu - \sum_{i=1}^{N} F_{\mu i} x_i^a\right)^2}. \tag{2.46}$$

The replica calculation unfolds in terms of the following order parameters for all $a = 1, .., n$:

$$m_a = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_{0,i} \tag{2.47}$$

$$Q_a = \frac{1}{N}\sum_{i=1}^{N} (x_i^a)^2 \tag{2.48}$$

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_i^b. \tag{2.49}$$

They measure the overlaps between replicas and the ground-truth, the self-overlap of the replicas, and overlaps between different replicas. The computation involves an average over the distribution of the data matrix $\mathbf{F}$, which is a delicate point. In fact, we know how to compute it in two cases:

- When $\mathbf{F}$ is (Gaussian) i.i.d.

The replica formula (the Tanaka formula [157]) for this case is well-known and has been postulated in different situations [168, 87, 88, 186]. Generic methods to prove replica formulas have been proposed based on the Guerra interpolation technique [67]. This heuristic replica result has been recently rigorously proven in a series of papers [13, 134]. In a more recent proof [14], it has been shown, again, that the formula is not specific to Gaussian i.i.d. matrices, but that

any matrix with i.i.d elements of unit variance and zero mean leads to the same exact result for mean squared error.

- When $\mathbf{F}$ is right rotationally invariant

This case includes the Gaussian i.i.d. one, but is more general. It has been computed in [79] and proving it for a convex penalty is at the heart of another work, detailed in the next chapter. We consider that $\mathbf{C} \equiv \mathbf{F}^T\mathbf{F}$ has a well-defined eigenvalue distribution p_λ with compact support. We can define its minimum λ_{min} and maximum λ_{max}. To write the free energy, we need some transforms which come from random matrix theory, that are associated with p_λ. The *Stieltjes transform* of $\mathbf{C}$ reads

$$\mathcal{S}_{\mathbf{C}}(z) = \int_{\lambda_{min}}^{\lambda_{max}} \mathrm{d}\lambda \frac{p_\lambda(\lambda)}{\lambda - z} = \mathbb{E}\left[\frac{1}{\lambda - z}\right] \tag{2.50}$$

and is correctly defined outside of p_λ's support. The corresponding *R-transform* is

$$\mathcal{R}_{\mathbf{C}}(x) = \mathcal{S}_{\mathbf{C}}^{-1}(-x) - \frac{1}{x}. \tag{2.51}$$

The detail of the computation is given in appendix B. We take the replica symmetrix (RS) ansatz, which is necessarily the correct one when the penalty function is convex, since our minimization problem admits only one solution. RS also holds for the teacher-student Bayes-optimal scenario, thanks to the Nishimori identity. We can thus remove the subscripts of the order parameters which share a unique value. The replica formula for the average free energy finally yields

$$\begin{aligned}\Phi(Q, q, m, \hat{Q}, \hat{q}, \hat{m}) = \mathcal{G}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) + \left(-\frac{\mathbb{E}[x_0^2] - 2m + q}{\Delta} + \frac{\Delta_0(Q-q)}{\Delta^2}\right)\mathcal{G}'_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) \\ + \frac{Q\hat{Q}}{2} - m\hat{m} + \frac{q\hat{q}}{2} + \int \mathrm{d}x_0\, p_{x_0}(x_0) \int \mathrm{D}z \log\left\{\int \mathrm{d}x P(x) e^{-\frac{\hat{Q}+\hat{q}}{2}x^2 + \hat{m}xx_0 + z\sqrt{\hat{q}}x}\right\}\end{aligned} \tag{2.52}$$

where $\mathcal{G}_{\mathbf{C}}$ is defined with respect to p_λ as

$$\mathcal{G}_{\mathbf{C}}(x) = \frac{1}{2}\mathrm{Sup}_\Lambda\left\{-\int \mathrm{d}\lambda p_\lambda(\lambda) \log|\Lambda - \lambda| + \Lambda x\right\} - \frac{1}{2}\log|x| - \frac{1}{2}. \tag{2.53}$$

Note that in the domain of definition of $\mathcal{R}_{\mathbf{C}}$, we have $\mathcal{G}'_{\mathbf{C}}(x) = \frac{1}{2}\mathcal{R}_{\mathbf{C}}(x)$. For simplicity, we will use this relation in the following results. However, remember that it only holds in a specific regime, and the general (and somewhat less aesthetic) correct expression would be to simply write things in terms of $\mathcal{G}_{\mathbf{C}}$ and its derivative. We can now minimize the free energy with respect to its parameters, which provides 6 saddle-point equations. We can rewrite them as equations on two parameters, which characterize the mean squared error with respect to the ground-truth E, and variance V of the estimator:

$$E = q - 2m + \mathbb{E}[x_0^2] \tag{2.54}$$

$$V = Q - q. \tag{2.55}$$

Defining the two functions

$$f_a(A, B) = \frac{\int \mathrm{d}x\, xP(x)e^{-\frac{(x-B)^2}{2A}}}{\int \mathrm{d}x P(x) e^{-\frac{(x-B)^2}{2A}}} \qquad f_v(A, B) = A\frac{\partial f_a(A, B)}{\partial B}, \tag{2.56}$$

the replica equations on (E, V) read

$$E = \mathbb{E}\left[\left\{ f_a\left(\frac{\Delta}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}, x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\sqrt{(E - \frac{\Delta_0}{\Delta}V)\mathcal{R}'_{\mathbf{C}}(-\frac{V}{\Delta}) + \Delta_0\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\right) - x_0\right\}^2\right] \tag{2.57a}$$

$$V = \mathbb{E}\left[f_v\left(\frac{\Delta}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}, x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\sqrt{(E - \frac{\Delta_0}{\Delta}V)\mathcal{R}'_{\mathbf{C}}(-\frac{V}{\Delta}) + \Delta_0\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\right)\right] \tag{2.57b}$$

with the expectation taken on $x_0 \sim p_{x_0}(x_0)$ and $z \sim \mathcal{N}(0, 1)$. To numerically solve the equations, we simply need to iterate them by computing the left-hand-side at iteration $t + 1$ as a function of the right-hand-side taken from iteration t. In the limit $\Delta \to 0$, these equations can be written in terms of a rescaled variance and error and naturally yield proximal operators that are widely used in convex optimization. We will spend some time discussing them in the following chapter. For now, all we need to know is that the replica method allows to obtain a set of equations on the mean squared error of the estimator with respect to the ground truth signal, for right rotationally invariant matrices. Those equations can be solved by initializing them properly, then iterating them. We can thus draw a phase diagram depending on parameters α, ρ: for each point we can solve the equations and get a theoretical value of the error made by the estimator. However, we need to proceed with care: indeed it is not always clear whether equations (2.57) have only one solution, and finding one could simply correspond to a local minimum of the free energy instead of a global one.

Theoretical phase transitions

We first focus on data matrices which are i.i.d. which is simply a particular case of the replica analysis. We need to specify a prior distribution $p_{x_0}(\mathbf{x_0})$, in the following we will assume that the ground truth comes from a separable Gauss-Bernoulli distribution, such that for sparsity $\rho \in [0; 1]$, for a scalar x:

$$p_{x_0}(x) = (1 - \rho)\delta(x) + \rho\phi_0(x) \tag{2.58}$$

with ϕ_0 a known distribution, classically a normal one. For a fixed ρ, we know that in the region $\alpha < \rho$ we cannot reconstruct the signal. The line $\alpha = \rho$ is called the *information-theoretical (IT)* transition. As we increase α, we get more and more measurements, and we expect the reconstruction error to decrease. In fact, the theory predicts a *phase transition*: the error is non-zero, then past a given threshold $\alpha(\rho)$, the error goes to zero and the described estimator achieves perfect reconstruction.

i) In the Bayes-optimal case,

the phase diagram is somewhat subtle. Recall that we are trying to recover $\mathbf{x_0}$, which is described by the global minimum of the free energy. However, say that we initialize our equations (2.57) on (E, V) starting from no information at the signal, i.e. with $m = 0$ (no overlap with the ground-truth). Then, there is a regime where the iterations will converge to a local minimum of the free energy, and will not be able to reach the global minimum. The phase diagram is divided into three parts. Below the IT threshold, it is impossible to recover the ground truth. Directly above the IT threshold, the free energy has a local minimum which absorbs our iterations if we start with no knowledge of the signal. The error associated with this local minimum is not zero. If we increase the number of measurements, the free energy will at some point have only one global minimum, then our iterations will converge to it, and the error will become zero since we are describing the ground truth: it is the *easy* phase. This transition is marked by a phase transition that we call the *Bayesian hard phase line*, or the *spinodal* (a term borrowed from thermodynamics which indicates a phase transition). The region between the IT threshold and the spinodal is called the *hard phase*. The existence of this phase might be seen as irrelevant,

but we will see that it is very important for the algorithmic treatment of our problem, and that understanding the behavior of the free energy is informative.

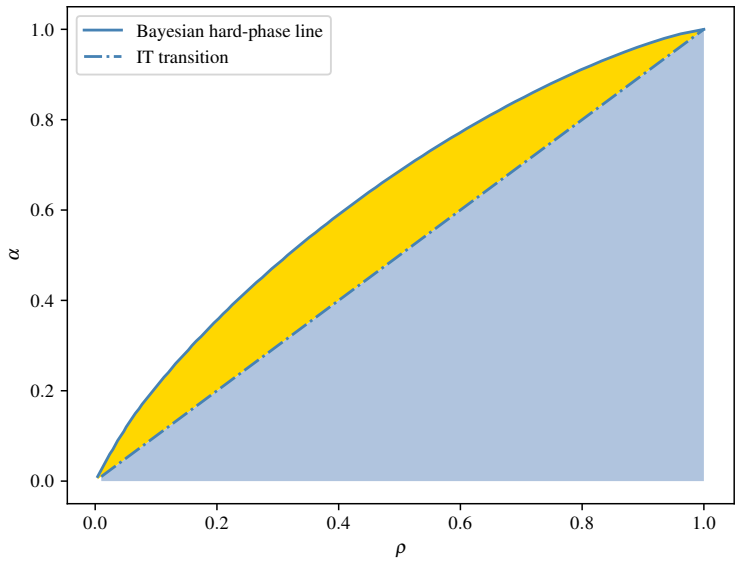


FIGURE 2.6: Phase diagram for i.i.d. matrices in the (α, ρ) space. Below the dashed line of the IT transition (blue region), it is impossible to recover the signal. Between the IT line and the Bayesian hard phase solid line lies the hard phase (yellow region), where a naive iteration of the replica equations reaches a local minimum of the free energy with a non zero error. Above the Bayesian hard phase line, the free energy has a unique global minimum.

ii) In the ℓ_1 reconstruction case,

the free energy of the replica calculation has only one global minimum which describes the estimator $\hat{\mathbf{x}}$. Iterating the equations on (E, V) safely describe the error achieved by this estimator with respect to the ground-truth. There is a phase transition marked by the *Donoho–Tanner line* [46]: above this line, $\hat{\mathbf{x}}$ matches $\mathbf{x_0}$, but it fails to do so below. We cannot do better than this line with ℓ_1 reconstruction, which is therefore sub-optimal. The Donoho–Tanner and Bayesian hard phase lines are compared in Fig. 2.7.

2.2.3 Approximate message passing algorithms

The phase diagrams for i.i.d. matrices provide some insight on the properties of the estimator $\hat{\mathbf{x}}$ in the ℓ_1 reconstruction, and of replica free energy for Bayes-optimal reconstruction. We now turn to the practical aspect of our problem: how can we reconstruct the signal $\mathbf{x_0}$ with an algorithm in polynomial time? We will use message passing algorithms, a statistical physics inspired variant of belief propagation, where local beliefs are approximated by Gaussian distributions. For our setting in particular, we turn to approximate message passing (AMP) [45] that applies to Gaussian i.i.d. matrices, and Vector approximate message passing (VAMP) [132] which applies to rotationally invariant matrices. We choose these algorithms because they come with a rich physical interpretation, as will be seen further on. To provide a general grasp on those algorithms' logic, we will show a glimpse of AMP's derivation starting from BP. A detailed derivation can be found in [87]. Let us start by writing the BP equations associated to the

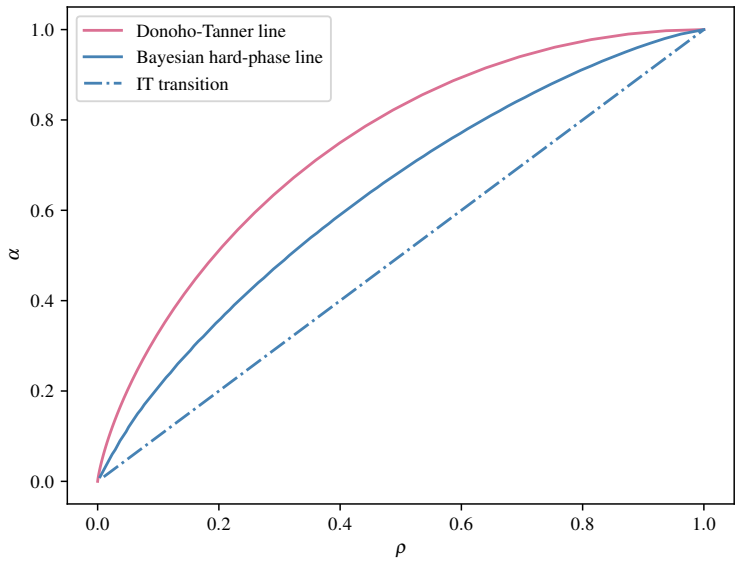


FIGURE 2.7: Theoretical phase transitions for noiseless linear regression for Gaussian i.i.d. matrices in the (α, ρ) phase diagram. The purple Donoho–Tanner is for ℓ_1 reconstruction, the blue Bayesian hard phase line is for Bayes-optimal reconstruction of a Gauss-Bernoulli signal. Above the line: the mean squared error of the estimator is null and the signal is perfectly estimated. Below the line: the error obtained by solving the replica equations grow (initializing with no knowledge of the signal). The Bayes-optimal spinodal is below the Donoho–Tanner line, indeed the Bayes-optimal setting allows recovering the signal with fewer measurements than ℓ_1 reconstruction, as it exploits more information.

probability distribution (2.43). The associated factor graph is given in Fig. 2.8. There are in total MN edges, thus the BP equations will involve $2MN$ messages, that we call $m_{i\to\mu}, m_{\mu\to i}$, and read:

$$m_{\mu\to i}(x_i) = \frac{1}{Z_{\mu\to i}} \int \prod_{j\neq i} \mathrm{d}x_j e^{-\frac{1}{2\Delta}(\sum_{j\neq i} F_{\mu j}x_j + F_{\mu i}x_i - y_\mu)^2} \prod_{j\neq i} m_{j\to\mu}(x_j) \tag{2.59a}$$

$$m_{i\to\mu}(x_i) = \frac{1}{Z_{i\to\mu}} P(x_i) \prod_{\gamma\neq\mu} m_{\gamma\to i}(x_i) \tag{2.59b}$$

where $Z_{\mu\to i}$, $Z_{i\to\mu}$ are normalizations to ensure that messages are probability distributions. Recall that $P(x_i)$ here is the prior distribution.

We would like to simplify these $2MN$ equations, which are two numerous and difficult to handle, since we look at the asymptotic limit of large N. There are two stages of simplification of these equations: first exploiting the fact that $F_{\mu i}$ elements are of order $1/\sqrt{N}$, and further on using the fact they are Gaussian. The result is a set of *TAP equations*, named for Thouless, Anderson and Palmer who first derived them to write a mean-field theory for the Sherrington-Kirkpatrick model [163], introduced in Chapter 1. A possible derivation is the *cavity method* exploited in [106]. The recipe of the cavity method usually includes these ingredients: introducing a cavity by removing a variable, then focusing on resulting marginal distributions, and making use of the resulting absent (for a tree) or weak correlations between remaining variables to simplify some terms (in particular invoking a central limit theorem on local fields to treat them as Gaussian variables). Such a cavity approach was used in 1.3. We will here only give the gist ideas of the derivation of TAP equations starting from BP equations. First of all we need

to define the mean and variance of message $m_{i\to\mu}$:

$$a_{i\to\mu} = \int \mathrm{d}x_i \; x_i m_{i\to\mu}(x_i) \tag{2.60a}$$

$$v_{i\to\mu} = \int \mathrm{d}x_i \; x_i^2 m_{i\to\mu}(x_i) - a_{i\to\mu}^2. \tag{2.60b}$$

Then, defining the probability distribution

$$m_i(x) \cong P(x_i) \prod_\gamma m_{\gamma\to i}(x_i) \tag{2.61}$$

where we have "completed" $m_{i\to\mu}$ by the marginal $m_{\mu\to i}$, we also define its mean and variance

$$a_i = \int \mathrm{d}x_i \; x_i m_i(x_i) \tag{2.62a}$$

$$v_i = \int \mathrm{d}x_i \; x_i^2 m_i(x_i) - a_i^2. \tag{2.62b}$$

The key point is to write $a_{i\to\mu}$ as a_i, plus a correction term which is linear in $F_{\mu i}$ to keep elements of order $1/\sqrt{N}$, but discard those of smaller order. This additional correction is absolutely crucial, and is called the *Onsager term*. For the SK model, the Onsager term translates the difference between the average of the local fields in the complete model, and their average in the model induced by creating a cavity. It sits in the magnetization approximation as an extra correction term compared to traditional mean-field approaches[2]. Defining $w_\mu = \sum_i F_{\mu i} a_{i\to\mu}$, we can write the TAP equations for Gaussian i.i.d matrices. Adding the appropriate time indices, the TAP equations turn into *approximate message passing (AMP)* equations which read:

$$V^{(t+1)} = \frac{1}{N}\sum_i v_i^{(t)} \tag{2.63}$$

$$\omega_\mu^{(t+1)} = \sum_i F_{\mu i} a_i^{(t)} - \frac{(y_\mu - \omega_\mu^{(t)})}{\Delta + V^{(t)}} \left[\frac{1}{N}\sum_i v_i^{(t)}\right] \tag{2.64}$$

$$\Sigma^{(t+1)} = \frac{\Delta + V^{(t+1)}}{\alpha} \tag{2.65}$$

$$R_i^{(t+1)} = a_i^{(t)} + \sum_\mu F_{\mu i} \frac{(y_\mu - \omega_\mu^{(t+1)})}{\alpha} \tag{2.66}$$

$$a_i^{(t+1)} = f_a\left(\Sigma^{(t+1)}, R_i^{(t+1)}\right) \tag{2.67}$$

$$v_i^{(t+1)} = f_v\left(\Sigma^{(t+1)}, R_i^{(t+1)}\right) \tag{2.68}$$

where f_a, f_v are the functions defined in (2.56) and depend on the prior distribution $P(x)$. For instance, if the signal model taken for inference is a Gauss-Bernoulli $P(x) = (1-\rho)\delta(x) + \rho\phi(x)$ with some known distribution ϕ; a reasonable initialization is

$$a_i^{(0)} = \rho \int \mathrm{d}x \; x\phi(x) \tag{2.69a}$$

$$v_i^{(0)} = \rho \int \mathrm{d}x \; x^2\phi(x) - (a_i^{(0)})^2 \tag{2.69b}$$

$$\omega_\mu^{(0)} = y_\mu. \tag{2.69c}$$

[2]Note that the first appearance of the Onsager term involves three Nobel prize holders: Onsager himself, Thouless and Anderson... It is definitely a beautiful piece of physics.

Let us note a few facts:
– the TAP equations correspond to the fixed point of AMP. If we derive them starting from BP equations (that we know how to properly iterate), the time indices seem obvious. However, starting from the physics cavity method only provides the fixed point, and adding the appropriate time indices can be tricky. In fact, finding the right time indices to guarantee convergence for the original SK model TAP equations was quite a challenge, and it was only fully understood recently.
– BP equations involved a number of variables of order MN, while for AMP the number is of the order $M + N$: one loop of AMP is much faster than one loop of iterating BP messages.

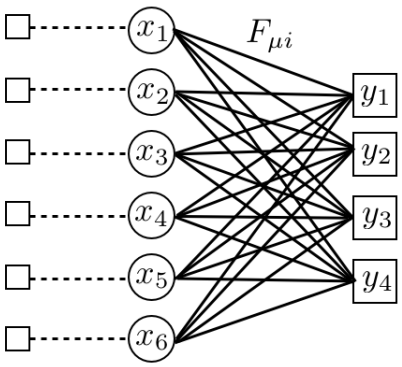


FIGURE 2.8: Factor graph for linear reconstruction. The circles are variable nodes $\{x_i\}$, $i = 1, ..., N$. The squares are function nodes, to the left they describe the priors $P(x_i)$, and to the right the measurements y_μ, $\mu = 1, ..., M$ which are obtained by multiplying the signal elements through matrix elements $F_{\mu i}$.

We could reduce even further AMP equations by removing intermediate variables. A particular advantage of these equations is that we can rewrite them on meaningful quantities that characterize statistical properties of the involved variables. The first one is the variable $V^{(t)} = \frac{1}{N}\sum_i v_i^{(t)}$, the average variance of local beliefs. We can define another quantity as $E^{(t)} = \frac{1}{N}\sum_i (a_i^{(t)} - x_{0,i})^2$ which computes the error between the ground truth $\mathbf{x_0}$, and the belief on the estimator average at a given iteration of the algorithm. Without going into detail, by simply analyzing the variable R_i, we find out that in the asymptotic limit, $(R_i - x_{0,i})$ is a Gaussian variable of mean 0 and variance $\alpha(E + \Delta_0)$. Exploiting this fact, we can pack AMP into equations on E and V, that we call *density evolution* equations:

$$V^{(t+1)} = \mathbb{E}\left[f_v\left(\frac{\Delta + V^{(t)}}{\alpha}, x_0 + z\sqrt{\frac{E^{(t)} + \Delta}{\alpha}} \right) \right] \tag{2.70a}$$

$$E^{(t+1)} = \mathbb{E}\left[\left\{ f_a\left(\frac{\Delta + V^{(t)}}{\alpha}, x_0 + z\sqrt{\frac{E^{(t)} + \Delta}{\alpha}} \right) - x_0 \right\}^2 \right] \tag{2.70b}$$

where the expectation is on $x_0 \sim p_{x_0}$, and $z \sim \mathcal{N}(0, 1)$. These equations should seem familiar: in fact they are exactly the same as those we obtained in (2.57) with the replica method, in the particular case of a Gaussian i.i.d. matrix $\mathbf{F}$. This "coincidence" is incredible: we have designed an algorithm (AMP), hoping that it will converge, and we obtain a set of equations on the error and variance achieved by its estimator at each iteration. But those equations, at their fixed point, are none other than the error and variance predicted by the replica method by

minimizing the free energy associated with the problem of maximizing the posterior distribution (2.43). This delightful correspondence will allow to draw many parallels between message passing algorithms, which are already closely linked to the cavity method and the Bethe free energy, and replica calculations.

Many variants of message passing algorithms exist. In this chapter, we are mainly interested in AMP [45] and Vector AMP [132], which tackle Gaussian i.i.d and rotationally invariant matrices for linear regression. While AMP is based on a loopy factor graph with scalar-valued nodes, VAMP proceeds on a non-loopy graph with vector-valued nodes. It also displays an Onsager term. A study of VAMP's convergence will be at the heart of the next chapter. For now, we underline the fact that both algorithms converge very quickly (around a dozen iterations).

2.3 Universal transitions in noiseless compressed sensing

This section is adapted from [2].

2.3.1 Equivalence between right rotationally invariant and Gaussian i.i.d. matrices

We have seen in the previous part the theoretical phase transitions for ℓ_1 reconstruction for Gaussian i.i.d. matrices; and for Bayes-optimal estimation the existence of a hard phase. We would now like to do numerical simulations to see how different classes of matrices relate to these transitions; using message passing algorithms as solvers. A priori, we do not know what those transitions will become for different matrices: there could be no phase transition, no hard phase, or the transition line could be above or below the one for Gaussian i.i.d. matrices. A first easy comparison can be established between rotationally invariant matrices and Gaussian ones. We start from a right rotationally invariant matrix decomposed as $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$, with an arbitrary rotation matrix $\mathbf{U}$ and singular values on $\mathbf{\Sigma}$'s diagonal, but where the matrix $\mathbf{V}$ has been randomly (and independently of $\mathbf{\Sigma}$ and $\mathbf{U}$) generated from the Haar measure. We focus the noiseless setting: we wish to find $\boldsymbol{x}$ such that

$$\boldsymbol{y} = \mathbf{F}\boldsymbol{x} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}\boldsymbol{x}. \tag{2.71}$$

We will do some lego-playing with those matrices to modify the writing of the problem. If $M \leq N$, then $\mathbf{\Sigma}$ is written as $\mathbf{\Sigma} = \left[\begin{array}{c|c} \tilde{\mathbf{\Sigma}} & 0 \end{array}\right]$ and we define

$$\mathbf{\Sigma}^{\text{inv}} = \left[\begin{array}{c} \tilde{\mathbf{\Sigma}}^{-1} \\ \hline 0 \end{array}\right] \text{ such that } \mathbf{\Sigma}^{\text{inv}}\mathbf{\Sigma} = \left[\begin{array}{c|c} \mathbf{I_M} & 0 \\ \hline 0 & 0 \end{array}\right].$$

Multiplying (2.71) on both sides by $\mathbf{U}^T$, and then by $\mathbf{\Sigma}^{\text{inv}}$; one reaches

$$\tilde{\boldsymbol{y}} = \mathbf{\Sigma}^{\mathbf{inv}}\mathbf{U}^T\boldsymbol{y} = \tilde{\mathbf{V}}\boldsymbol{x} \tag{2.72}$$

where $\tilde{\mathbf{V}}$ is an $M \times N$ matrix composed of the first M lines of $\mathbf{V}$. We have assumed without loss of generality that the singular values are non-zero, otherwise we have to keep the first r lines of $\mathbf{V}$ with r its rank.
If instead $M > N$, $\mathbf{\Sigma}$ is written as

$$\Sigma = \left[\begin{array}{c} \tilde{\mathbf{\Sigma}} \\ \hline 0 \end{array}\right]$$

and we define $\mathbf{\Sigma^{inv}} = \left[\begin{array}{c|c} \tilde{\mathbf{\Sigma}}^{-1} & 0 \end{array} \right]$ such that $\mathbf{\Sigma^{inv}\Sigma} = \mathbf{I_N}$. Multiplying (2.71) by $\mathbf{U}^T$ then $\mathbf{\Sigma^{inv}}$, we obtain a similar equation

$$\tilde{\boldsymbol{y}} = \mathbf{\Sigma^{inv}}\mathbf{U}^T\boldsymbol{y} = \mathbf{V}\boldsymbol{x}. \tag{2.73}$$

In both cases, we thus see that the problem has been transformed – in a constructive way – into a standard linear system with the sensing matrix $\tilde{\mathbf{V}}$ when $M \leq N$ being a (sub-sampled) random rotation one, or sensing matrix $\mathbf{V}$ when $M > N$. This shows that all rotationally invariant matrices, which satisfy $\mathbf{U}$ and $\mathbf{\Sigma}$'s independence on $\mathbf{V}$, can be transformed the same way and are in the same universality class as far as noiseless linear recovery is concerned, i.e. they will display the same phase transitions.

Since Gaussian i.i.d. matrices belong to the ensemble of random rotationally invariant matrices (in this case $\mathbf{\Sigma}$ follows the Marcenko-Pastur law [169]) this means that all the information theoretic rigorous results (such as phase transitions and mean squared error values) with zero noise for random Gaussian i.i.d. matrices applies verbatim to all rotationally invariant ensemble, (as long as the SVD matrices $\mathbf{U}$ and $\mathbf{\Sigma}$ are independent of $\mathbf{V}$). This is a very strong universality, that applies to the phase transitions of the three lines discussed in section 2.2.2: the Donoho–Tanner for ℓ_1 reconstruction, the Bayes-optimal hard phase line, and the mean squared error with respect to the ground-truth at each point of the (α, ρ) phase diagram. However, note that the above construction depends crucially on the fact that we consider here noiseless measurements. It would not work if an additional Gaussian noise were added in (2.71): in this case, the transformation would make the i.i.d. Gaussian noise a correlated one. Indeed, the replica formula for noisy measurements underlines that the mean squared error depends on the precise set of matrices in noisy reconstruction [153, 168] (this formula is not yet fully rigorous, but see [16] for a proof in a restricted setting).

The equivalence of transitions for noiseless measurements can also be seen thanks to the replica calculation, by the following hand-waving argument[3]. Note that the universality of the Donoho–Tanner was already hinted to in [80]. We will start from the replica equations on (E, V) for rotationally invariant matrices (2.57), and see how they become the same as their sisters for Gaussian i.i.d. matrices (2.70) in the limit $\Delta \to 0$. Taking the case $M < N$, the main point is to look at the R-transform $\mathcal{R}_{\mathbf{C}}$ of $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ in the appropriate limit, i.e. when its argument $-V/\Delta$ goes to $-\infty$. The matrix $\mathbf{C}$ has $M - N$ zero eigenvalues, hence $p_\lambda(\lambda)$ has a delta-peak in 0 of weight $1 - \alpha$ and the Stieltjes $\mathcal{S}_{\mathbf{C}}$ diverges to the left of 0. $\mathcal{S}_{\mathbf{C}}$ is a bijection from $]-\infty, 0[$ into $]0, +\infty[$. In particular, we easily obtain its equivalent in 0^- from its definition:

$$\mathcal{S}_{\mathbf{C}}(x) \underset{x\to 0^-}{\sim} \frac{1-\alpha}{x}. \tag{2.74}$$

which goes to $+\infty$. Then

$$\mathcal{S}^{-1}_{\mathbf{C}}(-x) \underset{x\to -\infty}{\sim} \frac{1-\alpha}{x}, \tag{2.75}$$

$$\mathcal{R}_{\mathbf{C}}(x) \underset{x\to -\infty}{\sim} -\frac{\alpha}{x}, \tag{2.76}$$

$$\mathcal{R}'_{\mathbf{C}}(x) \underset{x\to -\infty}{\sim} \frac{\alpha}{x^2}. \tag{2.77}$$

Using those equivalents for $x = -V/\Delta$ inside (2.57), we get the same equations as (2.70) in the zero noise limit. Hence, the expression of the error E and variance V is independent of the distribution of non-zero eigenvalues of $\mathbf{C}$, i.e. the singular values of $\mathbf{F}$: the transitions are the same for any right rotationally invariant matrix.

[3] to differentiate from a wand-waving argument, which is pure magic!

2.3.2 Universal transitions for structured matrices

Some types of structured matrices

We will now go beyond rotationally invariant matrices, and consider the following types of matrices.

- *Discrete cosine transform (DCT) matrices*

An $N \times N$ discrete cosine transform (DCT) matrix $\mathbf{Y}$ is defined by:

$$Y_{jk} = \sqrt{\frac{2}{N}} \epsilon_k \cos\left(\frac{\pi(2j+1)k}{2N}\right), \tag{2.78}$$

where $j, k \in [\![0, N-1]\!]$, $\epsilon_0 = 1/\sqrt{2}$, $\epsilon_i = 1$ for $i = 1, ..., N-1$. To obtain $\mathbf{F} \in \mathbb{R}^{M \times N}$ if $M < N$, we randomly pick M rows out of N from $\mathbf{Y}$; and if $M > N$ we add $N - M$ randomly picked rows from $\mathbf{Y}$ to complete it. The DCT is a Fourier-like transform [5], but it provides a decomposition of a signal on a basis of cosine functions of varying magnitudes and frequencies, with real coefficients instead of complex ones. Decomposing an image thanks to the DCT usually concentrates most of the visually significant information in just a few coefficients. The DCT is thus widely used in image compression applications [179], and other types of data compression. Note that the DCT matrix is not rotationally invariant: if we perform its SVD decomposition, we can see that the resulting orthogonal matrices have very specific structure in terms of sinusoids.

- *Hadamard matrices*

A natural variant of the DCT is given by Hadamard matrices. $\mathbf{H}$ is an $N \times N$ Hadamard matrix if its entries are ± 1 and its rows are pairwise orthogonal, i.e. $\mathbf{H}\mathbf{H}^T = N\mathbf{I_N}$. For every integer k, there exists a Hadamard matrix $\mathbf{H_k}$ of size 2^k. These can be created with Sylvester's construction: Let H be a Hadamard matrix of order N. Then the partitioned matrix

$$\begin{bmatrix} \mathbf{H} & \mathbf{H} \\ \mathbf{H} & -\mathbf{H} \end{bmatrix}$$

is a Hadamard matrix of order $2N$. Again, Hadamard matrices are not rotationally invariant. In [137], AMP is used as a capacity-achieving decoder for sparse superposition codes, and decoding complexity is shown to be significantly reduced by using Hadamard data matrices.

- *Random features maps*

Finally, we define random features (RF) maps as encountered in nonlinear regression problems. In such settings, a random features matrix

$$\mathbf{F} = h(\mathbf{W}\mathbf{X}) \tag{2.79}$$

is obtained from the raw data matrix $\mathbf{X}$ by means of a random projection matrix $\mathbf{W}$ and a pointwise nonlinear activation h. Kernel regression models, nonlinear in the original data $\mathbf{X}$, can then be approximately but efficiently solved by the linear estimation problem (2.71), with an appropriate choice for h and the $\mathbf{W}$-distribution [128]. Such matrices, that can be seen as the output of a neuron with random weights, have been investigated in particular in the context of neural networks [127, 93]. Indeed, in neural networks configurations with random weights play an important role as they define the initial loss landscape. They are also fundamental in the random kitchen sinks algorithm in machine learning [128]. In what follows, we will use random features matrices where both $\mathbf{W}$ and $\mathbf{X}$ are random Gaussian i.i.d. matrices. The function h is taken successively to be a sign step function, a hyperbolic tangent, and a rectified linear unit (ReLu) defined as $\mathrm{ReLu}(x) = 0$ if $x < 0$, x otherwise.

Numerical results

A first amusing test is to apply the routine from section 2.3.1 on a fabricated problem. We first generate a right-rotationally invariant matrix and a ground-truth vector, then the noiseless measurement vector $\mathbf{y}$. After transforming the problem to reach vector $\tilde{\mathbf{y}}$, we apply AMP on $\tilde{\mathbf{y}}$ and the associated sensing matrix, which allows to indeed reconstruct the ground-truth signal in the easy region of the phase diagram for Gaussian matrices.

Moving on, we now want apply VAMP out of its comfort zone with structured matrices described above. In fact, we could pick an algorithm out of different options [153, 27], in particular, using the general expectation-propagation (EP) [109, 121] scheme. A variation of EP called OAMP specially adapted to rotation matrices is developed in [97]. However, we enjoy using VAMP because it provenly follows in the asymptotic limit state evolution equations that describe the error achieved by its estimator at each iteration, and which (we will elaborate on this in our next chapter) correspond to the replica equations (2.57) [153, 168, 16]. Therefore, VAMP comes with a theoretical description of the mean squared error. Note that we have no guarantee that VAMP will converge for structured matrices, however it turns out that for many cases (adding some damping on its iterations, which slows down convergence but helps in setting the algorithm on a converging path rather than straying off after one over-enthusiastic step) it does converge. To perform simulations, we generate a synthetic problem with a Gauss-Bernoulli distributed ground-truth, then the random matrix of our choice, and we multiply both to obtain the noiseless measurement vector. We run VAMP and when it converges, we can compute the error of the estimator with respect to the ground-truth, and compare it to the one we get for rotationally invariant matrices (that we also know analytically). For ℓ_1 recovery, it has been shown empirically that the Donoho–Tanner transition seems to hold for a wider range of random matrix ensembles, see e.g. [43, 114]. Another line of work showed that the convex ℓ_1 reconstruction problem can be treated through conic geometry, and the success probability of signal recovery only depends on a geometric number characterizing a subcone (statistical dimension or Gaussian width) [32, 6]. Our experiments include both the ℓ_1 reconstruction case, and the Bayes-optimal estimation.

- Bayes-optimal reconstruction

To generate Figure 2.9, we ran VAMP 50 times on 50×50 points spanning the (α,ρ)-space with a generated DCT matrix, and computed the average mean squared error (MSE) between the signal $\mathbf{x_0}$ and the reconstructed configuration. The MSE is represented with a color bar (white means perfect reconstruction). We observe a phase transition in the Bayes-optimal that matches the theoretical Bayesian hard phase line. For all other structured matrices mentioned above, we obtain the same color diagram: each time the phase transition traces the same line. We also compared the error obtained by VAMP for different matrices. In figure 2.10, we plot the MSE averaged on 20 executions of VAMP for three values of fixed ρ and α ranging between 0 and 1. We get the same error in reconstruction for all matrices, following the MSE for Gaussian i.i.d. matrix for $\rho = 0.25$, 0.5 and 0.75.

- ℓ_1 recovery

The same protocols are applied to the ℓ_1 reconstruction. Averaging on 20 executions of VAMP (or 50 for small α where finite-size effects are more important), we recover again in Figure 2.11 a phase transition matching the theoretical Donoho–Tanner line for Gaussian i.i.d. matrices [43]. Figure 2.12 is obtained the same way as 2.10.

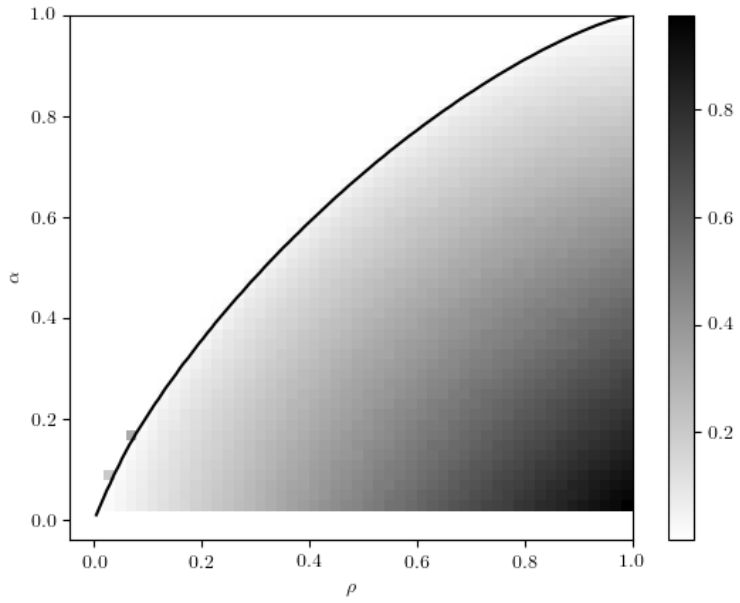


FIGURE 2.9: Phase diagram for a DCT matrix (width $N = 1000$) in the Bayes-optimal case. The averaged MSE on 50 executions of VAMP is represented by a color-code, displaying a phase transition that matches the theoretical Bayesian hard phase line for Gaussian i.i.d. matrices (black line). Some finite-size effects can be seen.

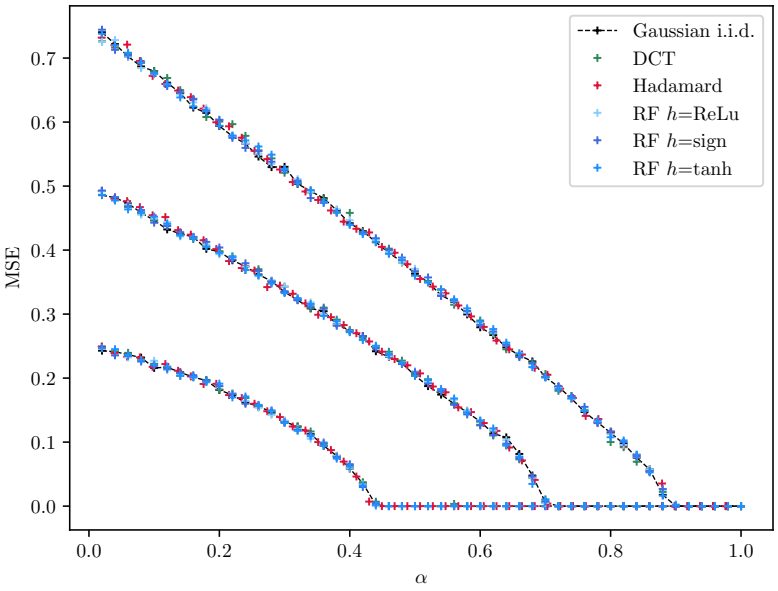


FIGURE 2.10: Mean squared error for $\rho = 0.25$, 0.5 and 0.75 (bottom to up curves) in the Bayes-optimal case averaged on 20 executions of VAMP for Gaussian i.i.d, DCT, Hadamard, random features matrices $\mathbf{F} = h(\mathbf{WX})$ with $h = \text{ReLu}$, $h = \text{sign}$, $h = \tanh$ ($\mathbf{W}$ and $\mathbf{X}$ are Gaussian i.i.d of size $\alpha N \times N$ and $N \times N$) . The width is $N = 2000$ for almost all matrices, except the Hadamard ($N = 2048$).

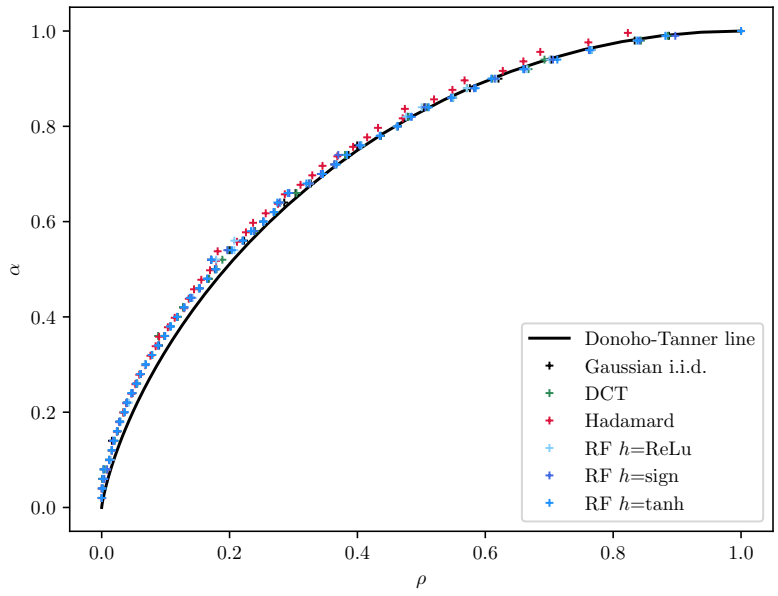


FIGURE 2.11: Phase diagram in the ℓ_1 reconstruction case obtained by averaging on 20 to 50 executions on VAMP. The dots indicate the phase transitions for Gaussian i.i.d., DCT (width $N = 2000$), Hadamard matrices ($N = 4096$); and random feature matrices $\mathbf{F} = h(\mathbf{WX})$ with $h = \text{ReLu}$, $h = \text{sign}$, $h = \tanh$ ($\mathbf{W}$ and $\mathbf{X}$ are Gaussian i.i.d. of size $\alpha N \times N$ and $N \times N$ with $N = 2000$). They match the theoretical Donoho–Tanner transition for Gaussian i.i.d. matrices (black line).

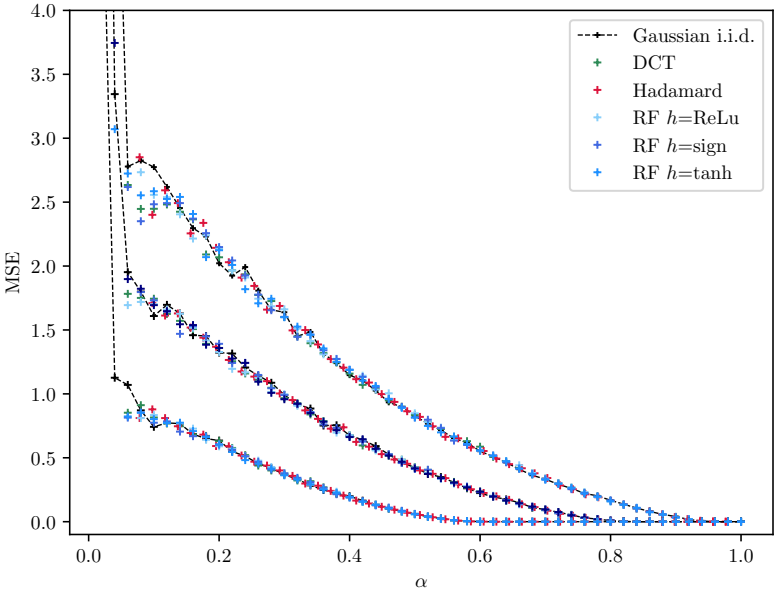


FIGURE 2.12: Mean squared error for $\rho = 0.25$, 0.5 and 0.75 (bottom to up curves) in the ℓ_1 reconstruction case averaged on 20 executions of VAMP for Gaussian i.i.d, DCT, Hadamard, random features matrices $\mathbf{F} = h(\mathbf{WX})$ with $h = \text{ReLu}$, $h = \text{sign}$, $h = \tanh$ ($\mathbf{W}$ and $\mathbf{X}$ are Gaussian i.i.d of size $\alpha N \times N$ and $N \times N$). The width is $N = 2000$ for all matrices except the Hadamard ($N = 2048$).

2.3.3 Discussion and limits of the universality

We have seen that the universality in noiseless compressed sensing is not limited to the ℓ_1-type reconstruction as in [43, 114], but extends to other quantities and estimators, such as the hard phase line in Bayesian reconstruction, and the mean squared error. Besides, it does not only include right rotationally invariant matrices, but empirically extends to Fourier-type matrices and to the random features maps currently studied in machine learning. It seems that the density evolution equations which predict the error of the estimator achieved by VAMP also apply to those matrices, even though they are not proved in this setting. However, all matrices do not share the same properties. Let us have a look at two examples of structured matrices that do not seem to follow these universal phase transitions.

- Haar wavelet matrices

Haar wavelet matrices can be defined recursively by:

$$\mathbf{W_2} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \text{ and } \mathbf{W_{2k}} = \begin{bmatrix} \mathbf{W_k} \otimes [1, -1] \\ \mathbf{I_k} \otimes [1, 1] \end{bmatrix}$$

where $\mathbf{I_k}$ is the identity matrix of size k and $\otimes$ is the Kronecker product. Those matrices are used, in particular, for the reknown Daubechies wavelet decomposition [38]. In the easy phase of compressed sensing, both in the Bayes-optimal setting and the ℓ_1 recovery case, where VAMP applied to i.i.d. matrices (as well as Hadamard, DCT, random features matrices) perfectly reconstructs the signal; it fails to do so when applied to a Haar wavelet matrix. VAMP will then converge to a fixed point with a non-zero MSE, as seen in Figures 2.13 and 2.14. In fact, VAMP seems to always fail in reconstructing the signal for a Haar wavelet matrix: the mean squared error converges to a finite quantity, but never to zero. Of course, this approach is algorithm-dependent, but it means that VAMP is not necessarily appropriate to work with those matrices, and that the theoretical insight we gain from its density evolution equations does not apply empirically to some types of structured matrices. To sum up, we can roughly say thet we do not observe the same phase transitions for VAMP applied to a Haar wavelet matrix.

- Gaussian correlated matrices

Let $\mathbf{T}(c)$ be the Toeplitz matrix defined as $\mathbf{T}(c)_{ab} = c^{|a-b|}$. As in [129], we consider structured matrices which satisfy the following property: if $\mathbf{F}$ is a $M \times N$ matrix, its elements have covariance

$$\mathbb{E}[F_{ia}F_{jb}] = \frac{1}{M} C_{ij} D_{ab} \tag{2.80}$$

where all $D_{aa} = 1$. Such a matrix can be obtained, for instance, by multiplying a $M \times N$ Gaussian i.i.d. matrix $\mathbf{G}$ by a $N \times N$ Toeplitz matrix $\mathbf{T}(\sqrt{c})$. In our simulations, we thus used matrices

$$\mathbf{F}(c) = \frac{1}{\sqrt{M}} \mathbf{G}\mathbf{T}(\sqrt{c}) \tag{2.81}$$

for different values of c. Running VAMP in the Bayes-optimal case with parameters (α, ρ) in the easy phase of compressed sensing, we find that it converges and perfectly reconstructs the signal for c small enough ($c = 0.15$), but fails to converge and has a diverging MSE when c is larger ($c = 0.8$), as seen in Figure 2.13. In the ℓ_1 recovery setting, still in the easy phase above the Donoho–Tanner line, VAMP fails to converge to a fixed reconstructed vector $\hat{\boldsymbol{x}}$ both for c very small ($c = 0.001$) or large ($c = 0.8$). However, the MSE stays very close to a small non-zero value, which can be seen in Figure 2.14. After a large number of iterations, VAMP keeps returning a vector very close to the original signal, but does not manage to reconstruct it. The final MSE's approximate value also depends on c: the larger the correlations are, the larger

the MSE is. In [129], the authors study such correlated matrices in the very sparse regime when α is close to zero, and show that the theoretical phase transition for ℓ_1 recovery depends on c.

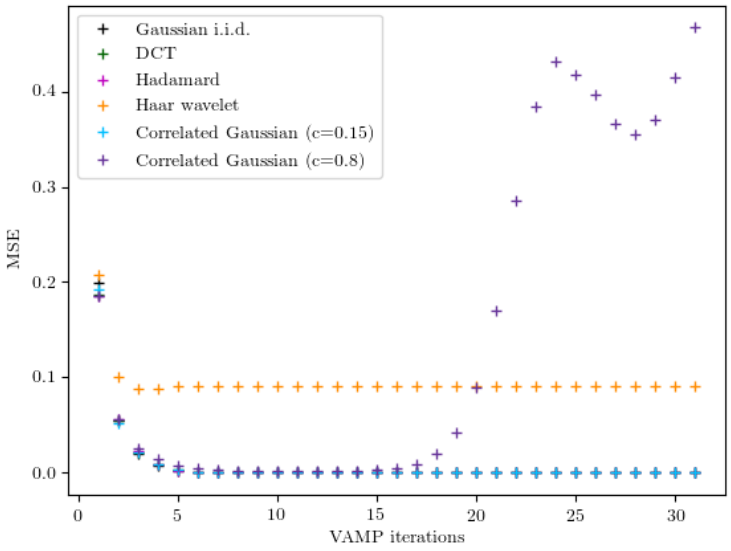


FIGURE 2.13: Mean squared error at each iteration of VAMP in a Bayes-optimal setting, for $\rho = 0.3$, $\alpha = 0.7$ (in the easy phase of compressed sensing, i.e. above the Bayesian hard phase line) . VAMP is applied to a Gaussian i.i.d, a DCT, correlated Gaussian (width $N = 2000$); Hadamard and Haar wavelet matrices (width $N = 2048$). The MSE for the Haar wavelet matrix converges to a finite value but does not go to zero as for the other matrices. The MSE for a Gaussian correlated matrix converges for small correlation $c = 0.15$ and diverges for larger correlation $c = 0.8$.

Summary of Chapter 2 We considered the problem of noiseless compressed sensing, both in the Bayes-optimal setting and in the ℓ_1 recovery case. We have clear theoretical understanding of phase transitions (Bayesian hard phase and Donoho–Tanner lines) for i.i.d. matrices with mean zero and variance 1 (thanks to a proven replica formula), and heuristic understanding for right rotationally invariant matrices (which will be made rigorous in the next chapter). Through simple arguments, we have shown that right rotationally invariant matrices share the same phase transitions as i.i.d. ones. Besides, we observe through simulations that this universality of transitions and of error values apply to a much larger class of matrices, including DCT, Hadamard and random features matrices. Our simulations are made using message passing algorithms which are a variant of belief propagation, and come with a rigorous theoretical description in the asymptotic limit through their state evolution equations, that correspond to the replica equations on (E, V). However, VAMP applied on Haar wavelet matrices and Gaussian correlated matrices does not display the same phase transitions. It would be interesting to find a good criterion to identify which matrices satisfy this universality and which do not; and what happens in a generalized model including non-linearities.

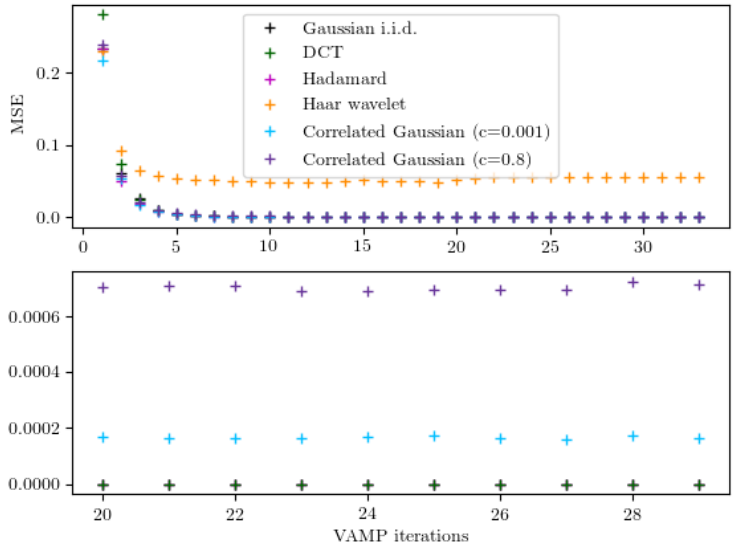


FIGURE 2.14: Mean squared error at each iteration of VAMP in the case of ℓ_1 recovery, for $\rho = 0.3$, $\alpha = 0.8$ (in the easy phase of compressed sensing, above the Donoho–Tanner line). VAMP is applied to a Gaussian i.i.d, a DCT, correlated Gaussian (width $N = 2000$); Hadamard and Haar wavelet matrices (width $N = 2048$). The MSE for the Haar wavelet matrix converges to a finite value but does not go to zero. The MSE for Gaussian correlated matrices does not effectively converge, but stays very close to a small non-zero value, as seen in the zoomed-in second subplot.

Chapter 3

Asymptotic errors for convex penalized linear regression beyond Gaussian matrices

3.1 Definition of the problem

In this section, we will focus on the regression problem with convex penalty introduced in 2.2.2, in the asymptotic regime. We are interested in the standard quadratic minimization problem, given input space $\mathcal{X} \subset \mathbb{R}^N$ with M samples:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathcal{X}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x}) \right\} \tag{3.1}$$

where $\mathbf{F} \in \mathbb{R}^{M\times N}$ is a known data matrix. f is a convex and separable regularization function. For instance, this setting includes the LASSO [165] by taking the ℓ_1 norm as penalty, ridge regression [101] by taking the squared ℓ_2 norm, or elastic nets [187] which imply a linear combination of the two. We assume the vector $\mathbf{y}$ has been obtained according to a noisy linear process as

$$\mathbf{y} = \mathbf{F}\mathbf{x}_0 + \mathbf{w} \tag{3.2}$$

where all elements from the vector $\mathbf{x}_0 \in \mathbb{R}^N$ are identically and independently distributed (i.i.d.) according to an arbitrary given distribution $p_{x_0}(\cdot)$, and $\mathbf{w} \in \mathbb{R}^M$ is an i.i.d. Gaussian white noise of zero mean and variance Δ_0, independent of $\mathbf{F}$ and $\mathbf{x}_0$. Our aim is to provide expressions for the mean squared error on the recovery of $\mathbf{x}_0$, which are asymptotically exact. The mean squared error is defined as:

$$\text{MSE} = \frac{1}{N}\mathbb{E}\left[\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2\right]. \tag{3.3}$$

We consider random matrices $\mathbf{F}$ with fixed aspect ratio $\alpha \equiv M/N$ as $M, N \to \infty$.

In a pioneering paper, [20] considered this case for Gaussian i.i.d. matrices and provided a rigorous derivation of an explicit formula for the asymptotic mean squared error of the LASSO estimator. Our goal here is to go beyond the Gaussian case, hence we will look at rotationally invariant matrices, previously defined in 2.2.2. Note that this setting, although specific, enjoys a long standing tradition in signal processing [131], statistical physics [69], random matrix theory [68] and communications theory [169]. We will adopt the statistical physics point of view, in particular through the replica method approach. It allows to give typical-case results, that represent a replacing approach to the worst-case analysis [113]. This chapter is adapted from [61].

Main assumptions Since we want to obtain a rigorous statement, we have to be somewhat careful with the mathematical assumptions involved, that we state now to lay proper foundation

for our work:

- f is proper, closed, convex and separable.
- The empirical distributions of the underlying truth $\mathbf{x}_0$ and singular values of the rotationally invariant sensing matrix respectively converge with second order moments, as defined in appendix C.1, to given distributions p_{x_0} and p_λ.
- The distribution p_λ is non all-zero and has compact support.
- We consider the limit $M, N \to \infty$ with fixed ratio $M/N = \alpha$.

3.2 Statistical physics result: the replica formula

As presented in 2.2.2 and detailed in appendix B, the statistical physics replica method allows to analyze this teacher-student setting, and incorporates the regularization f through the prior distribution term. f being convex, the minimization problem has a single solution, and the replica-symmetric ansatz is necessarily the correct one, since it allows for a unique minimum of the free energy. The replica method allows to compute the free energy of the associated inference problem [79], and by minimizing it, provides properties of the estimator $\hat{\mathbf{x}}$ though scalar parameters, such as its squared norm or its overlap with the ground truth. Replica equations can also be recast in terms of the sought-after mean squared error, and heuristically predict that it corresponds to E which solves the fixed point equations:

$$V = \mathbb{E}\left[\frac{1}{\mathcal{R}_{\mathbf{C}}(-V)}\mathrm{Prox}'_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V)}\right)\right] \tag{3.4a}$$

$$E = \mathbb{E}\left[\left\{\mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V)}\right) - x_0\right\}^2\right]. \tag{3.4b}$$

where $\mathbf{C} = \mathbf{F}^T\mathbf{F}$, $\mathcal{R}_{\mathbf{C}}$ is the R-transform with respect to p_λ, and expectations are over $z \sim \mathcal{N}(0,1)$ and $x_0 \sim p_{x_0}$. Prox is the proximal operator defined as:

$$\forall \gamma \in \mathbb{R}^+, x, y \in \mathbb{R} \quad \mathrm{Prox}_{\gamma f}(y) \equiv \underset{x\in\mathbb{R}}{\arg\min}\left\{f(x) + \frac{1}{2\gamma}(x-y)^2\right\}. \tag{3.5}$$

3.3 Vector approximate passing and its state evolution

To prove the replica formula characterizing the error, we resort to Vector approximate message passing (VAMP) [132] and we will build on results established mostly in [132] and [52]. VAMP belongs to the class of message passing algorithms and is also linked with the expectation-propagation strategy [110], as well as other algorithms [31, 97]. However, we will see that it has the significant trait of providing rigorously derived state evolution equations. Let us start by writing the VAMP equations that correspond to our problem. [65] explains that penalized least squares regression can be seen as a maximum a posteriori (MAP) estimation, but also has other equally acceptable Bayesian interpretations. We will here simply resort to the MAP formulation of VAMP.

3.3.1 MAP formulation of Vector approximate message passing

Choose initial $A_1^{(0)}$ *and isotropically distributed* $\mathbf{B}_1^{(0)}$

$$\hat{\mathbf{x}}_1^{(t)} = \mathrm{Prox}_{f/A_1^{(t)}}\left(\frac{\mathbf{B}_1^{(t)}}{A_1^{(t)}}\right) \qquad \hat{\mathbf{x}}_2^{(t)} = (\mathbf{F}^T\mathbf{F} + A_2^{(t)}\mathrm{Id})^{-1}(\mathbf{F}^T\mathbf{y} + \mathbf{B}_2^{(t)}) \tag{3.6a}$$

$$V_1^{(t)} = \frac{1}{A_1^{(t)}}\left\langle \mathrm{Prox}'_{f/A_1^{(t)}}\left(\frac{\mathbf{B}_1^{(t)}}{A_1^{(t)}}\right)\right\rangle \qquad V_2^{(t)} = \frac{1}{N}\mathrm{Tr}\left[(\mathbf{F}^T\mathbf{F} + A_2^{(t)}\mathrm{Id})^{-1}\right] \tag{3.6b}$$

$$A_2^{(t)} = \frac{1}{V_1^{(t)}} - A_1^{(t)} \qquad A_1^{(t+1)} = \frac{1}{V_2^{(t)}} - A_2^{(t)} \tag{3.6c}$$

$$\mathbf{B}_2^{(t)} = \frac{\hat{\mathbf{x}}_1^{(t)}}{V_1^{(t)}} - \mathbf{B}_1^{(t)} \qquad \mathbf{B}_1^{(t+1)} = \frac{\hat{\mathbf{x}}_2^{(t)}}{V_2^{(t)}} - \mathbf{B}_2^{(t)} \tag{3.6d}$$

where $\langle\cdot\rangle$ is an element-wise averaging operator $\langle\mathbf{x}\rangle = \frac{1}{N}\sum_{i=1}^N x_i$, and the vector valued proximal operator is defined as :

$$\forall\gamma\in\mathbb{R}^+, \mathbf{x},\mathbf{y}\in\mathcal{X} \quad \mathrm{Prox}_{\gamma f}(\mathbf{y}) \equiv \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\left\{f(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{y}\|_2^2\right\}. \tag{3.7}$$

Proximal operators are well-known objects in convex optimization, and we will strongly rely on their properties. In particular, $\mathrm{Prox}_{\gamma f}$ can be evaluated even when f is non-differentiable. In the MAP formulation of VAMP, the proximals play the role of *denoiser functions*. In this work, we slightly abuse notations by noting the vector-valued proximal (which is separable) and the induced element-wise scalar-valued proximal in the same way, since it is easy to see which one is used depending on the argument. Note that the second equation in (3.6a) can also be written in terms of the proximal of $\mathbf{x}\mapsto\|\mathbf{y}-\mathbf{F}\mathbf{x}\|_2^2$ as

$$\hat{\mathbf{x}}_2^{(t)} = \mathrm{Prox}_{\frac{1}{2A_2^{(t)}}\|\mathbf{y}-\mathbf{F}\cdot\|_2^2}\left(\frac{\mathbf{B}_2^{(t)}}{A_2^{(t)}}\right).$$

This proximal formulation is convenient to draw parallels with proximal descent algorithms. The latter enjoy a long lasting success in machine learning and signal processing [36] because of their stability, simplicity to implement and solid theoretical anchoring, notably from a monotone operator theory point of view [18]. An example of popular algorithm for solving composite convex optimization problems of the form $\arg\min_{\mathbf{x}}\{f(\mathbf{x})+g(\mathbf{x})\}$ is the Douglas-Rachford splitting method [122], which roughly amounts to successively applying the proximal of f and the one of g. It is shown in [52], a connection pursued in [100], that VAMP is similar to a Douglas-Rachford descent with parameters that adapt to the local curvature of the cost function.

Let us take a closer look at VAMP's equations. At each iteration, it returns two estimators, $\hat{\mathbf{x}}_1^{(t)}$ and $\hat{\mathbf{x}}_2^{(t)}$. $V_1^{(t)}, V_2^{(t)}$ are their respective variances, and they are linked through parameters $A_1^{(t)}, A_2^{(t)}$. $\mathbf{B_1}^{(t)}, \mathbf{B_2}^{(t)}$ are intermediate vector-variables. Rescaling them by $A_1^{(t)}, A_2^{(t)}$ and applying the proximal operators of $f/A_1^{(t)}$ and of $\|\mathbf{y}-\mathbf{F}\cdot\|_2^2/(2A_2^{(t)})$ yields $\hat{\mathbf{x}}_1^{(t+1)}$ and $\hat{\mathbf{x}}_2^{(t+1)}$.

3.3.2 Equality of $\hat{\mathbf{x}}$ and VAMP's fixed point

Sub-differential and proximal relation We start by defining a useful operator for convex analysis. The *sub-differential* of $f:\mathbb{R}\to\mathbb{R}$, denoted ∂f, is the valued-set operator

$$\partial f : x \mapsto \{u\in\mathbb{R} | \forall y\in\mathbb{R}, \langle x-y, u\rangle + f(x) \leq f(y)\}. \tag{3.8}$$

Graphically, the sub-differential in x gives back all the slopes of affine functions that coincide with f in x, but only touch the curve of f in one point. If f is differentiable in x, these affine functions reduce to only one, which is the tangent in x, thus $\partial f(x) = f'(x)$. The advantage of the sub-differential is that it is well-defined for convex functions and can also be evaluated even if they are non-differentiable. Taking the example of the absolute value $f_1 = |x|$, we find that

$$\partial f_1(x) = \begin{cases} -1 & \text{if } x < 0 \\ ]-1, 1[& \text{if } x = 0 \\ -1 & \text{if } x > 0. \end{cases} \tag{3.9}$$

A beautiful relation links proximal operator and sub-differential: in fact the proximal operator of a convex function f is the resolvent of its sub-differential [18], i.e.

$$\text{Prox}_{\gamma f} = (\text{Id} + \gamma \partial f)^{-1}. \tag{3.10}$$

For the ℓ_1 norm in particular, it is straightforward to recover the well-known *soft thresholding* function as its proximal, from this identity and the expression of the sub-differential of the absolute value above.

Optimality condition If we consider the definition of the desired estimator $\hat{\mathbf{x}}$ (3.1), we can rewrite it as a first-order condition:

$$\mathbf{F}^T(\mathbf{F}^T\mathbf{y} - \hat{\mathbf{x}}) = \partial f(\hat{\mathbf{x}}). \tag{3.11}$$

Now say that we are looking at a converging trajectory of VAMP, such that it reaches its fixed point. We can look at the returned estimator $\hat{\mathbf{x}}_\mathbf{1} = \hat{\mathbf{x}}_\mathbf{2}$ and see if it matches the desired $\hat{\mathbf{x}}$. In particular, we would like $\hat{\mathbf{x}}_\mathbf{1}$ to satisfy the first order condition.

Fixed point analysis Looking at the fixed point of VAMP, we replace the proximal by the resolvent of ∂f in (3.6d), which provides:

$$\mathbf{B}_1 = A_1\hat{\mathbf{x}}_1 + \partial f(\hat{\mathbf{x}}_1) \qquad \mathbf{B}_2 = (\mathbf{F}^T\mathbf{F} + A_2\text{Id})\hat{\mathbf{x}}_2 - \mathbf{F}^T\mathbf{y}. \tag{3.12}$$

Inserting in (3.6d) yields

$$(\mathbf{F}^T\mathbf{F} + A_2\text{Id})\hat{\mathbf{x}}_2 - \mathbf{F}^T y = \frac{\hat{\mathbf{x}}_1}{V_1} - A_1\hat{\mathbf{x}}_2 - \partial f(\hat{\mathbf{x}}_1). \tag{3.13}$$

Since $V_1 = V_2$ and $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2$ at the fixed point, knowing from (3.6c) that $A_1 + A_2 = \frac{1}{V_1}$, we land on

$$\mathbf{F}^T(\mathbf{y} - \mathbf{F}\hat{\mathbf{x}}_\mathbf{1}) = \partial f(\hat{\mathbf{x}}_\mathbf{1}) \tag{3.14}$$

which is non other than the optimality condition of problem (3.1). Therefore, the fixed point of VAMP is indeed the solution estimator $\hat{\mathbf{x}}$.

3.3.3 State evolution of VAMP

The state evolution (SE) equations of VAMP are a set of equations that follow the algorithm, and provide the statistical distribution of the iterates. They are exact in the asymptotic limit, and a finite-size concentration inequality for AMP with Gaussian i.i.d. matrices can be found in [138]. In particular, SE builds on the fact that $\mathbf{B_1}$ and $\mathbf{B_2}$ behave as noisy Gaussian estimates

of $\mathbf{x}_0$ at each iteration:

$$\mathbf{B}_1^{(t)} = A_1^{(t)}(\mathbf{x}_0 + \mathbf{P}_1^{(t)}) \qquad \mathbf{B}_2^{(t)} = A_2^{(t)}(\mathbf{x}_0 + \mathbf{P}_2^{(t)}), \tag{3.15}$$

$$\text{where} \quad P_1^{(t)} \sim \mathcal{N}(0, \tau_1^{(t)}) \qquad P_2^{(t)} \sim \mathcal{N}(0, \tau_2^{(t)}). \tag{3.16}$$

The state evolution equations will involve the following parameters:

- $\tau_1^{(t)}$ and $\tau_2^{(t)}$ the variances of $P_1^{(t)}, P_2^{(t)}$.
- $V_1^{(t)}$ and $V_2^{(t)}$ the variances of the estimates $\hat{\mathbf{x}}_1^{(t)}$ and $\hat{\mathbf{x}}_2^{(t)}$.
- The mean squared errors of $\hat{\mathbf{x}}_1^{(t)}$ and $\hat{\mathbf{x}}_2^{(t)}$, given through functions $\mathcal{E}_1$ and $\mathcal{E}_2$, defined as

$$\mathcal{E}_1(A_1^{(t)}, \tau_1^{(t)}) = \mathbb{E}\left[\left(\text{Prox}_{f/A_1^{(t)}}(x_0 + P_1^{(t)}) - x_0\right)^2\right] \tag{3.17}$$

$$\mathcal{E}_2(A_2^{(t)}, \tau_2^{(t)}) = \mathbb{E}\left[\frac{\Delta_0 \lambda_{\mathbf{C}} + \tau_2^{(t)} A_2^{(t)2}}{(\lambda_{\mathbf{C}} + A_2^{(t)})^2}\right] \tag{3.18}$$

where expectations are taken on scalar variables $x_0 \sim p_{x_0}$, $P_1^{(t)} \sim \mathcal{N}(0, \tau_1^{(t)})$, and $\lambda_{\mathbf{C}} \sim p_\lambda$. $\mathcal{E}_2$ can also be written as:

$$\mathcal{E}_2 = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\|(\mathbf{F}^T\mathbf{F} + A_2^{(t)}\text{Id})^{-1}(\mathbf{F}^T y + \mathbf{B}_2^{(t)}) - \mathbf{x}_0\|_2^2\right] \tag{3.19}$$

where the expectation is with respect to $\mathbf{x_0}$ and $\mathbf{P_2}^{(t)}$. The state evolution equations then read

$$\alpha_1^{(t)} = \mathbb{E}\left[\text{Prox}'_{\frac{1}{A_1^{(t)}}f}(x_0 + P_1^{(t)})\right] \qquad V_1^{(t)} = \frac{\alpha_1^{(t)}}{A_1^{(t)}} \tag{3.20a}$$

$$A_2^{(t)} = \frac{1}{V_1^{(t)}} - A_1^{(t)} \qquad \tau_2^{(t)} = \frac{1}{(1-\alpha_1^{(t)})^2}\left[\mathcal{E}_1(A_1^{(t)}, \tau_1^{(t)}) - \alpha_1^{(t)2}\tau_1^{(t)}\right] \tag{3.20b}$$

$$\alpha_2^{(t)} = \mathbb{E}\left[\frac{A_2^{(t)}}{\lambda_{\mathbf{C}} + A_2^{(t)}}\right] \qquad V_2^{(t)} = \frac{\alpha_2^{(t)}}{A_2^{(t)}} \tag{3.20c}$$

$$A_1^{(t+1)} = \frac{1}{V_2^{(t)}} - A_2^{(t)} \qquad \tau_1^{(t+1)} = \frac{1}{(1-\alpha_2^{(t)})^2}\left[\mathcal{E}_2(A_2^{(t)}, \tau_2^{(t)}) - \alpha_2^{(t)2}\tau_2^{(t)}\right]. \tag{3.20d}$$

SE equations thus present an iterative scalar equivalent model which allows to track the asymptotic statistical properties of the iterates of VAMP. A series of groundbreaking papers initiated with [19] proved the exactness of these equations in the asymptotic limit, and extended the method to treat nonlinear problems [130] and handle rotationally invariant matrices [132].

If VAMP converges, the fixed point ensures that $\hat{\mathbf{x}}_\mathbf{1}$ is equal to $\hat{\mathbf{x}}_\mathbf{2}$, as well as their variances and errors. SE equations can be solved analytically if the teacher distribution $p_{x_0}(x_0)$ is known. In practice, all the averages are empirical and $\tau_1^{(0)}$ is initialized with the empirical variance of $\mathbf{B}_1^{(0)}$. Note that SE equations only hold if VAMP is properly initialized with an isotropically distributed vector $\mathbf{B}_1^{(0)}$, which empirically converges with second order moment. This is an important subtlety that we need to keep in mind if we want to make a point using the state evolution equations. Three additional assumptions on the denoiser functions (which in our case correspond to proximal functions) are required for the state evolution theorem in [132] to hold. These are automatically verified in the convex MAP case, as properties of the proximal mapping. This is reminded in appendix C.1.

3.3.4 Equivalence of state evolution and replica equations

We know lay out a first major argument : the fixed point of state evolution (3.20) provides the same set of equations as the replica density evolution equations (3.4). The error E and rescaled variance V from the replica formula correspond to the fixed point errors $\mathcal{E}_1 = \mathcal{E}_2$ and variances $V_1 = V_2$ from state evolution. This equivalence can be found easily by tinkering with the equations, and is detailed in appendix C.2.

3.4 A simplified algorithm: Oracle VAMP

3.4.1 Idea of the proof

We are now holding all the cards. We have an algorithm, VAMP, that tries to solve our minimization problem. If it converges, VAMP indeed goes to $\hat{\mathbf{x}}$, and its trajectory is rigorously described by state evolution equations. In particular, SE equations at their fixed point characterize the reached estimator $\hat{\mathbf{x}}$, and provide the sought-after mean squared error. They also coincide with the replica equations.

We are immediately tempted to close the case and conclude that replica equations do present the mean squared error of $\hat{\mathbf{x}}$. However, this would be a mistake at this point. We are still missing the flour that makes all ingredients stick together: everything said above relies on the fixed point of VAMP, but what if VAMP diverges? This is a classical caveat encountered when studying sequences: we can start by characterizing the limit, but we still have to prove that the sequence converges. In our case, if VAMP diverges, the state evolution equations will simply be describing an off-trailing series of estimators that have nothing to do with $\hat{\mathbf{x}}$. We thus have to show that there exists *any* converging trajectory of VAMP. Taking the fixed point itself as constant trajectory does not work either, because it is not a valid initialization. Indeed, SE equations only hold if initialized properly with $\mathbf{B_1}^{(0)}$ isotropically distributed. Therefore, our goal is the following: show that there exists a converging sequence of VAMP, for any initialization that is allowed by SE equations.

3.4.2 Definition of Oracle VAMP

Since we have the right to pick any proper initialization, we will craft a convenient one that simplifies VAMP's iterations. To do so, we choose A_1^0 and $\mathbf{B_1}^{(0)}$ as being the ones that we would get from the fixed point of state evolution. Therefore, parameters A_1, A_2, $V_1 = V_2$ which are assigned by SE equations will remain constant throughout VAMP's iterations. We call the resulting algorithm Oracle VAMP. It is of course completely unpractical, since it implies already solving SE equations before running VAMP, but it a valid theoretical construct and will serve our purpose. Oracle VAMP reads

$$\hat{\mathbf{x}}_1^{(t)} = \mathrm{Prox}_{\frac{1}{A_1}f}\left(\frac{\mathbf{B}_1^{(t)}}{A_1}\right) \qquad \hat{\mathbf{x}}_2^{(t)} = \mathrm{Prox}_{\frac{1}{2A_2}||\mathbf{y}-\mathbf{F}\cdot||_2^2}\left(\frac{\mathbf{B}_2^{(t)}}{A_2}\right) \tag{3.21a}$$

$$\mathbf{B}_2^{(t)} = \frac{\hat{\mathbf{x}}_1^{(t)}}{V_1} - \mathbf{B}_1^{(t)} \qquad \mathbf{B}_1^{(t+1)} = \frac{\hat{\mathbf{x}}_2^{(t)}}{V_1} - \mathbf{B}_2^{(t)}, \tag{3.21b}$$

where the coefficients A_1 and A_2 verify:

$$V_1 = \mathcal{S}_{\mathbf{C}}(-A_2) = \langle \mathrm{Prox}'_{\frac{1}{A_1}f}(\mathbf{x_0} + \mathbf{P_1})\rangle / A_1 \qquad A_1 + A_2 = \frac{1}{V_1} \tag{3.22}$$

where $\mathbf{P_1}$ is Gaussian with variance τ_1 prescribed by SE, and $\mathcal{S}_\mathbf{C}$ is the Stieltjes transform with respect to the spectral measure of matrix $\mathbf{C}$. Oracle VAMP can be made even more compact, by writing it as one iteration on vector $\mathbf{B_2}^{(t)}$:

$$\mathbf{B}_2^{(t+1)} = \mathcal{O}_1 \circ \mathcal{O}_2(\mathbf{B}_2^{(t)}) \tag{3.23}$$

$$\text{where} \quad \mathcal{O}_1 = \frac{1}{V_1}\text{Prox}_{f/A_1}\left(\frac{\cdot}{A_1}\right) - \text{Id}, \quad \text{and} \quad \mathcal{O}_2 = \left(\frac{1}{V}\text{Prox}_{\frac{1}{2A_2}||\mathbf{y}-\mathbf{F}\cdot||_2^2}\left(\frac{\cdot}{A_2}\right) - \text{Id}\right). \tag{3.24}$$

Notice, on the way, that this iteration would become the Peaceman-Rachford operator [125] if we had $A_1 = A_2 = \frac{1}{2V_1}$ (but such a prescription would render the state evolution equations invalid: Oracle VAMP and Peaceman-Rachford remain different algorithms).

A simple example: the squared ℓ_2 penalty In the ℓ_2 penalty case, Oracle VAMP (3.23) drastically simplifies and converges after one iteration. To see it, we compute the proximal operator of a ℓ_2 penalty with parameter λ_2, which is a constant function:

$$\text{Prox}_{\frac{\lambda_2}{2}||\cdot||_2^2} = \frac{1}{1+\lambda_2}. \tag{3.25}$$

Using (3.25) in the definition of V_1 from (3.24) as the average of the proximal derivative immediately shows that $\frac{1}{V_1}\text{Prox}_{f/A_1}(\frac{\cdot}{A_1}) = \text{Id}$, and operator $\mathcal{O}_1$ is null. Therefore $\mathbf{B}_2^{(t)}$ cancels itself at the first iteration of the algorithm, directly leading to the fixed point:

$$\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 = (\mathbf{F}^T\mathbf{F} + A_2\text{Id})^{-1}(\mathbf{F}^T\mathbf{y}). \tag{3.26}$$

With squared ℓ_2 regularization, the convergence of Oracle VAMP is immediate. We will now like to explicit general convergence bounds for this algorithm, in particular derive Lipschitz constants for operators $\mathcal{O}_1$ and $\mathcal{O}_2$. Our approach is similar to [64] for Peaceman/Douglas-Rachford splitting. These bounds will depend on the properties of the convex regularization function. In particular, we will need two constants to characterize its convexity, that we define now.

3.4.3 Strong convexity and smoothness of a convex function

Definition *(Strong convexity)* A proper closed function is σ-strongly convex with $\sigma > 0$ if $f - \frac{\sigma}{2}\|.\|^2$ is convex. If f is differentiable, the definition is equivalent to

$$f(x) \geqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{\sigma}{2}\|x-y\|^2 \tag{3.27}$$

for all $x, y \in \mathcal{X}$.

Definition *(Smoothness for convex functions)* A proper closed function f is β-smooth with $\beta > 0$ if $\frac{\beta}{2}\|.\|^2 - f$ is convex. If f is differentiable, the definition is equivalent to

$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{\beta}{2}\|x-y\|^2 \tag{3.28}$$

for all $x, y \in \mathcal{X}$.

A consequence of those definitions is the following second order condition: if f is twice differentiable, it is σ-strongly convex and β-smooth if and only if:

$$\sigma\text{Id} \preceq \mathcal{H}_f \preceq \beta\text{Id}. \tag{3.29}$$

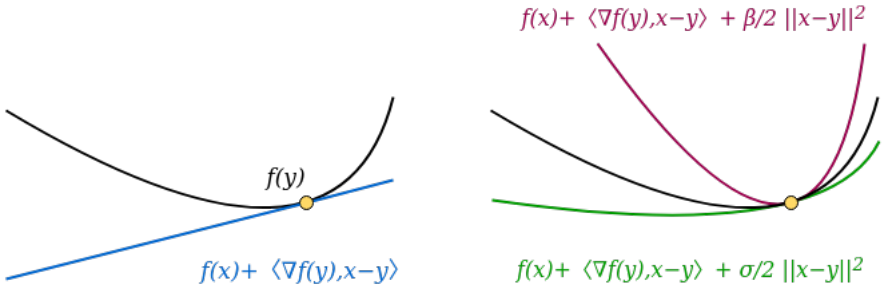


FIGURE 3.1: Left: a convex function f (black curve) with its usual definition of being above the blue line directed by its derivative in each point. Right: the function lies above the green curve, which adds a quadratic function of coefficient σ to the blue line, and the crimson line which adds a quadratic function of coefficient β. f is σ-strongly convex and β-smooth.

where $\mathcal{H}_f$ is its Hessian matrix; which means that the eigenvalues of $\mathcal{H}_f$ are between σ and β. Hence for any $\gamma > 0$, $f(\mathbf{x}) = \frac{\gamma}{2}\|\mathbf{x}\|_2^2$ is γ-smooth and γ-strongly convex. Note that any convex function is at least 0 strongly convex as it amounts to the usual convexity definition, but does not necessarily have a smoothness constant, as is the case for the non differentiable ℓ_1 norm, as shown in Fig. 3.2. However, we will be using theorems that include the case where there is no smoothness assumption, by setting the smoothness constant to $+\infty$.

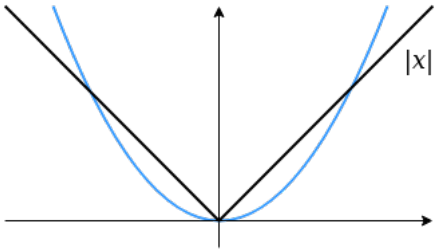


FIGURE 3.2: Absence of smoothness for the ℓ_1 norm: it will cross with any quadratic function.

3.4.4 Lipschitz constants of Oracle VAMP's operators

We will work with two sets of strong convexity and smoothness constants: (σ_1, β_1) associated with f (where β_1 is set to its $+\infty$ limit if there is no smoothness asssumption); and (σ_2, β_2) associated with $(\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{y}-\mathbf{F}\mathbf{x}\|_2^2)$. Since the eigenvalue distribution p_λ of $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ has compact support, we clearly have $(\sigma_2, \beta_2) = (\lambda_{min}(\mathbf{C}), \lambda_{max}(\mathbf{C}))$ the minimal and maximal value of the support. Using the properties of proximal operators and the fixed point of SE equations, we get the following upper bounds on the Lipschitz constant of the iteration (3.23), depending on the aspect ratio $\alpha = M/N$ and constants $(\sigma_{1,2}, \beta_{1,2})$. Calculations are detailed in appendix C.3.

Lipschitz constant of $\mathcal{O}_1$ — The Lipschitz constant $\mathcal{L}_1$ of the operator $\mathcal{O}_1$ in the cases where $0 < \sigma_1 < \beta_1$ or $0 < \sigma_1 = \beta_1$ respectively reads:

$$\mathcal{L}_1 = \max\left(\frac{|A_2 - \sigma_1|}{A_1 + \sigma_1}, \frac{|\beta_1 - A_2|}{A_1 + \beta_1}\right) \ , \ \mathcal{L}_1 = \sqrt{\left(\frac{(A_2^2 - A_1^2)}{(A_1 + \sigma_1)^2} + 1\right)} \tag{3.30}$$

Lipschitz constant of $\mathcal{O}_2$ — The Lipschitz constant $\mathcal{L}_2$ of the operator $\mathcal{O}_2$ reads

$$\mathcal{L}_2 = \max\left(\frac{|A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})|}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1|}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})}\right). \tag{3.31}$$

The case $0 < \sigma_1 < \beta_1, \alpha > 1$ yields the same constant as the one derived in [52], which studies a more general version of VAMP. Note that all those constants reduce to 1, if $A_1 = A_2$ is set, which is consistent with the 1-Lipschitz property of the Peaceman-Rachford operator [125].

3.5 Smoothed problem and its convergence

3.5.1 Definition of the modified problem

The ideal scenario to prove the convergence of Oracle VAMP would be to show that its iteration (3.23) is a contraction, which is true if the product of Lispchitz constants $\mathcal{L}_1\mathcal{L}_2$ is strictly smaller than 1. This inequality does not hold in general, but we will turn to a modified problem which will force convergence by shrinking down the Lipschitz constants. Basically, we add a squared ℓ_2 penalty with factor λ_2, that increases the strength of the regularization. We also replace f (which is potentially non-differentiable) by its twice differentiable approximation $\tilde{f}$ [89]. We now consider estimator

$$\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \tilde{f}(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2\right\}. \tag{3.32}$$

The new penalty function has become $h = \tilde{f} + \frac{\lambda_2}{2}\|\cdot\|_2^2$. We will show that for λ_2 large enough, Oracle VAMP applied to this modified problem becomes a contraction. To do this, we first derive two bounds on parameters A_1 and A_2. Again, we will make use of the strong convexity and smoothness constants (σ_h, β_h) of h, which depend on $(\tilde{\sigma}_1, \tilde{\beta}_1)$ the constants of $\tilde{f}$ as

$$\sigma_h = \tilde{\sigma}_1 + \lambda_2 \qquad \beta_h = \tilde{\beta}_1 + \lambda_2. \tag{3.33}$$

As before, the smoothness constants can be set to $+\infty$ in case there is no smoothness assumption on f.

3.5.2 Bounds on variance parameters A_1 and A_2

- Bound on A_1

From (3.6b), we immediately have $V_1 = \mathcal{S}_{\mathbf{C}}(-A_2)$, which results in:

$$\frac{1}{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_2} \leq V_1 \leq \frac{1}{\lambda_{min}(\mathbf{F}^T\mathbf{F}) - A_2}, \tag{3.34}$$

and using equality $A_1 + A_2 = \frac{1}{V_1}$ at the fixed point we obtain

$$\lambda_{min}(\mathbf{F}^T\mathbf{F}) \leq A_1 \leq \lambda_{max}(\mathbf{F}^T\mathbf{F}). \tag{3.35}$$

- Bound on A_2

We will proceed as for A_1, by writing V_1 as the Stieltjes transform associated with a bounded eigenvalue distribution, related to convexity properties of the penalty h. We recall the following equality on the proximal operator of h, for any $\gamma > 0$:

$$\mathrm{Prox}_{\gamma h}(\mathbf{x}) = (\mathrm{Id} + \gamma\partial h)^{-1}(\mathbf{x}). \tag{3.36}$$

h being twice differentiable and separable, the proximal is separable too and we get to the element-wise identity:

$$\text{Prox}'_{\gamma h}(x) = \frac{1}{1+\gamma f''(\text{Prox}_{\gamma h}(x))}. \tag{3.37}$$

Using V_1's definition from (3.6b) as a function of $\mathbf{B_1}$ taken at the fixed point yields

$$V_1 = \frac{1}{N}\text{Tr}\left[(A_1 I_N + \mathcal{H}_h(\text{Prox}_{\gamma h}(\mathbf{B_1}/A_1)))^{-1}\right] \tag{3.38}$$

$$= \frac{1}{N}\text{Tr}\left[(A_1 I_N + \mathcal{H}_h(\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}}))^{-1}\right] \tag{3.39}$$

$$V_1 = \mathcal{S}_{\mathcal{H}_h(\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}})}(-A_1) \tag{3.40}$$

since $\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}}$ is the desired estimator matched by VAMP's fixed point, and $\mathcal{S}_{\mathcal{H}_h(\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}})}$ is the Stieltjes transform associated with the eigenvalue distribution of $\mathcal{H}_h(\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}})$. We can bound the eigenvalues of $\mathcal{H}_h(\hat{\mathbf{x}}_{\boldsymbol{\lambda_2}})$: they are larger than $\tilde{\sigma}_h$ and smaller than $\tilde{\beta}_h$. Using this bound inside the definition of the Stieltjes transform provides

$$\frac{1}{\beta_h + A_1} \leq V_1 \leq \frac{1}{\sigma_h + A_1}, \tag{3.41}$$

then using equality $A_1 + A_2 = \frac{1}{V_1}$ at the fixed point, we get to

$$\sigma_h \leq A_2 \leq \beta_h. \tag{3.42}$$

3.5.3 A note on non-separable denoisers

For now, we have focused on the case of separable f and separable proximal functions, which play the role of denoiser functions in VAMP. In fact, in [53], the state evolution analysis of VAMP was extended to a large class of non-separable convex denoisers which verify a convergence property, called *convergence under Gaussian noise*. This result built upon previous work on convex, non-separable regularization in message passing algorithms in [23]. The state evolution equations are thus valid for this family of denoisers, which includes for instance group-based denoisers, convolutional denoisers, and convolutional neural nets. The Lipschitz constants derived in 3.4.4 still hold for non-separable denoisers, as the proof only depends on strong convexity and smoothness assumptions. Besides, according to [53], the VAMP iteration on V_1, V_2 in the non-separable setting are defined for $i = 1, 2$ by

$$V_i = \frac{1}{A_i}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial g_{i,k}(\mathbf{B_i}/A_i, \gamma)}{\partial B_{i,n}} \tag{3.43}$$

where $\mathbf{g}_i : \mathbb{R}^N \to \mathbb{R}^N$ is the proximal of the regularization for $k = 1$, and of the quadratic loss function for $k = 2$, and g_n its k-th component. This definition is exactly the normalized trace of the Jacobian matrix of the proximal, i.e. a more general case that includes the separable-case definition. We can replace (3.37) by its matricial equivalent. For h a twice-differentiable function and $\gamma > 0$, starting again from

$$\text{Prox}_{\gamma h}(\mathbf{x}) = (\text{Id} + \gamma \partial h)^{-1}(\mathbf{x}), \tag{3.44}$$

we apply $(\text{Id} + \gamma \nabla h)$ on both sides

$$\text{Prox}_{\gamma f}(\mathbf{x}) + \gamma \nabla f(\text{Prox}_{\gamma f}(\mathbf{x})) = \mathbf{x}, \tag{3.45}$$

then use the chain rule to differentiate the left-hand side with respect to $\mathbf{x}$, reaching

$$\mathcal{J}_{\text{Prox}_{\gamma h}}(\mathbf{x}) + \gamma \mathcal{H}_h(\text{Prox}_{\gamma f}(\mathbf{x}))\mathcal{J}_{\text{Prox}_{\gamma h}}(\mathbf{x}) = \text{Id} \tag{3.46}$$

where $\mathcal{J}$ denotes a Jacobian matrix and $\mathcal{H}$ a Hessian. Since h is a convex function, its Hessian is positive semi-definite, and, knowing that γ is strictly positive, the matrix $(Id+\gamma\mathcal{H}_h(\text{Prox}_{\gamma h}(\mathbf{x})))$ is invertible. We thus have :

$$\mathcal{J}_{\text{Prox}_{\gamma h}}(\mathbf{x}) = (\text{Id} + \gamma \mathcal{H}_h(\text{Prox}_{\gamma h}(\mathbf{x})))^{-1}. \tag{3.47}$$

Using the updated definition of V_1 (3.43), we reach (3.40) again. Therefore our bounds on A_1 and A_2 still hold for non-separable denoisers that satisfy the assumptions from [53].

3.5.4 Convergence bound on Oracle VAMP

The previous bounds on A_1 and A_2 serve to establish this fact: A_1 is bounded between quantities that only depend on $\mathbf{F}$, regardless of the penalty function. However, A_2 is bounded by constants that depend on the strong convexity (and possible smoothness) of the regularization. Tuning the strength of the added squared ℓ_2 regularization through parameter λ_2 also modifies A_2's range since $\sigma_h = \tilde{\sigma}_1 + \lambda_2$. In particular, if we look at the problem with λ_2 large, then A_2 will scale accordingly, while A_1 will remain stuck between the same finite bounds. Our goal now is to show that if λ_2 is large enough, Oracle VAMP applied on the corresponding problem becomes a contraction. We use the Lipschitz bounds (3.30) and (3.31), but applied to the modified problem with penalty h. We have a few cases to navigate:

i) $0 < \sigma_h < \beta_h$, β_h is finite and $\mathcal{L}_1 = \frac{A_2 - \sigma_h}{A_1 + \sigma_h}$

The complete Lispchitz constant $\mathcal{L} = \mathcal{L}_1\mathcal{L}_2$ yields

$$\mathcal{L} = \frac{A_2 - \sigma_h}{A_1 + \sigma_h} \max\left(\frac{A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})}\right) \tag{3.48}$$

and taking $\lambda_2 > \lambda_{max}(\mathbf{F}^T\mathbf{F}) - 2\lambda_{min}(\mathbf{F}^T\mathbf{F}) - \tilde{\sigma}_1$, combined to relations (3.33) and (3.35) guarantees $\mathcal{L} < 1$.

ii) $0 < \sigma_h < \beta_h$, β_h is finite and $\mathcal{L}_1 = \frac{\beta_h - A_2}{A_1 + \beta_h}$

In that case, clearly $\mathcal{L}_1 \leq 1$, and (3.35) tells us that

$$\mathcal{L}_2 \leq \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - \lambda_{min}(\mathbf{F}^T\mathbf{F})}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}. \tag{3.49}$$

Taking $\lambda_2 > \lambda_{max}(\mathbf{F}^T\mathbf{F}) - 2\lambda_{min}(\mathbf{F}^T\mathbf{F}) - \tilde{\sigma}_1$ guarantees again that $\mathcal{L}_2 < 1$, hence $\mathcal{L} < 1$.

iii) $0 < \sigma_h < \beta_h$, β_h is infinite (i.e. there is no smoothness assumption)

In that case, $\mathcal{L}_1 = \max\left(\frac{A_2 - \sigma_h}{A_1 + \sigma_h}, 1\right)$ and like before, $\lambda_2 > \lambda_{max}(\mathbf{F}^T\mathbf{F}) - 2\lambda_{min}(\mathbf{F}^T\mathbf{F}) - \tilde{\sigma}_1$ results in $\mathcal{L} < 1$.

iv) $0 < \sigma_h = \beta_h$

We then know from (3.42) that $\sigma_h = \beta_h = A_2$, and

$$\mathcal{L}_1 = \sqrt{\frac{2\sigma_h}{\sigma_h + A_1}}. \tag{3.50}$$

This function of σ_h increases on the real positive axis, and is thus bounded by $\sqrt{2}$. We define $\Lambda_2 = \sqrt{2}(\lambda_{max}(\mathbf{F}^T\mathbf{F}) - \lambda_{min}(\mathbf{F}^T\mathbf{F})) - \lambda_{min}(\mathbf{F}^T\mathbf{F}) - \tilde{\sigma}_1$, then picking $\lambda_2 > \Lambda_2$ ensures that $\mathcal{L} < 1$.

Finally, we see that Oracle VAMP applied for the modified problem (3.32) with λ_2 large enough (i.e. larger than a given constant Λ_2 that only depends on $\mathbf{F}$ and on the strong convexity constant of the analytic approximation $\tilde{f}$) reduces to a single contracting operator. Fig. 3.3 illustrates how Oracle VAMP, seen as one iteration on vector $\mathbf{B_2}$, becomes convergent when the LASSO penalty is strengthened by a ridge with regularization parameter λ_2 large enough. We conclude that for $\lambda_2 > \Lambda_2$, Oracle VAMP for the smoothed problem converges to its fixed point, which is the associated estimator $\hat{\mathbf{x}}_{\lambda_2}$.

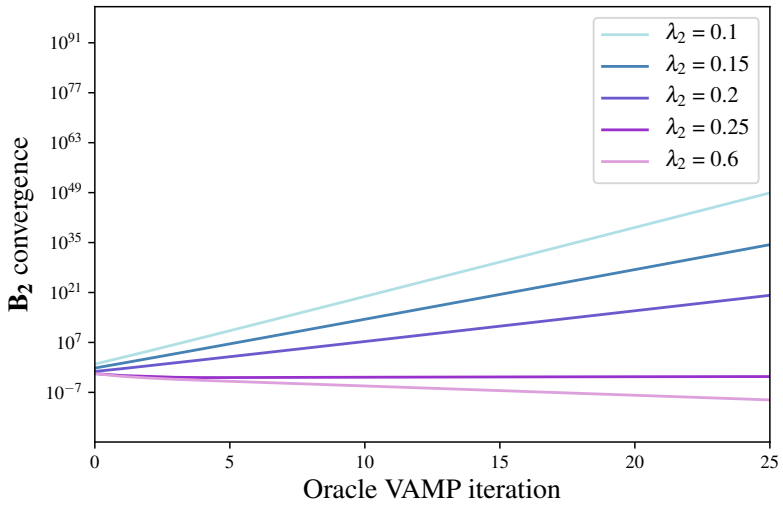


FIGURE 3.3: Convergence of vector $\mathbf{B_2}$ measured by $\frac{1}{N}\|\mathbf{B_2}^{(t+1)} - \mathbf{B_2}^{(t)}\|_2$ through successive Oracle VAMP iterations indexed by t. Oracle VAMP is ran on a Gaussian i.i.d matrix of size 100×1000 ($\alpha = 0.1$), for a Gauss-Bernoulli distributed ground truth with sparsity $\rho = 0.3$ with noise $\Delta_0 = 0.01$. The penalty function is the elastic net $f = \lambda_1\|\cdot\|_1 + \frac{\lambda_2}{2}\|\cdot\|_2^2$ with $\lambda_1 = 0.1$. The different lines correspond to different ridge regularization parameters λ_2. For λ_2 too small, Oracle VAMP clearly diverges; but convergence is enforced when λ_2 gets large enough. Oracle VAMP's initialization parameters are found by solving the state evolution equations for the elastic net problem, detailed in appendix C.4.

3.6 Analytic continuation and end of the proof

We have now proven the convergence of Oracle VAMP for the smoothed problem (3.32), provided the regularization is strong enough, thus completing the recipe to the proof explained in 3.4.1. In this regime, we can conclude two things: the replica formula for asymptotic error is correct, and the statistical distribution of $\hat{\mathbf{x}}_{\lambda_2}$ is given by VAMP's state evolution. Indeed, SE equations do not only give the scalar value of the error, but also statistical properties of the estimators. In particular, we know the distribution of $\mathbf{B}_1$ at the fixed point, which is related to variance τ_1. τ_1 can itself be written in terms of (E, V) the mean and variance of the estimator, as seen in appendix. Finally, for $\lambda_2 > \Lambda_2$, $\hat{\mathbf{x}}_{\lambda_2}$ has the following element-wise distribution:

$$\hat{x}_{\lambda_2} \sim \text{Prox}_{h/\mathcal{R}_\mathbf{C}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_\mathbf{C}(V)}\sqrt{(E - \Delta_0 V)\,\mathcal{R}'_\mathbf{C}(-V) + \Delta_0 \mathcal{R}_\mathbf{C}(-V)}\right) \tag{3.51}$$

where $z \sim \mathcal{N}(0, 1)$.

Nevertheless, we are not satisfied yet. We want to prove the expression of the error for any penalty function f, not only for a smoothed version of the problem with added regularization. The key here will be to perform an analytic continuation: the error is a scalar function that depends on parameter λ_2. If it is analytic in λ_2, and has a known expression for $\lambda_2 > \Lambda_2$, then we will be able to extend its expression to the rest of the domain.

Finite N case Let us first focus on the finite N regime, before taking the asymptotic limit $N \to \infty$. We invoke the optimality condition on the convex problem (3.32) which prescribes a solution $\mathbf{x}$ satsifying:

$$\left(\mathbf{F}^T\mathbf{F} + \lambda_2 \mathrm{Id} + \nabla \tilde{f}\right) \mathbf{x} = \mathbf{F}^T\mathbf{y}. \tag{3.52}$$

The left hand side is an analytic operator in λ_2 applied to $\mathbf{x}$. Using the analytic inverse function theorem [83], this clearly prescribes an analytic solution for $\mathbf{x}$ in λ_2. We then turn to the SE equations (3.20), which can also be written as (3.4). The contribution of λ_2 in these equations show up through the proximal operator of $h = \tilde{f} + \frac{\lambda_2}{2}\|\cdot\|_2^2$, and can be explicited thanks to the following expression of the proximal of a differentiable function with added squared ℓ_2 regularization:

$$\forall x \in \mathbb{R}, \quad \forall \gamma > 0 \quad \mathrm{Prox}_{\gamma(\tilde{f}+\frac{\lambda_2}{2}\|\cdot\|_2^2)}(x) = \left((1+\gamma\lambda_2)\mathrm{Id} + \gamma \tilde{f}'\right)^{-1}(x). \tag{3.53}$$

On the right-hand side, the function to invert is analytical in λ_2 with non-zero derivative, therefore by virtue of the analytical inverse function theorem [83], its inverse is also analytical in λ_2. Equations (3.4) are thus also analytical in λ_2. The implicit function theorem [84] ensures that the scalar quantities defined by those equations, including the mean squared error E at finite N, are analytic in λ_2. We can conclude using the analytic continuation property [83] that the replica equations, and all the scalar quantities from SE equations hold true whatever the value of λ_2. In particular, taking $\lambda_2 = 0$ provides the MSE of the modified problem with penalty $h = \tilde{f}$. This only differs from the original problem (3.1) by the use of a twice differentiable penalty function $\tilde{f}$. Going from the differentiable relaxation to the real problem is intuitive: it only relies on finding an appropriate sequence of twice differentiable functions $(f_k)_{k\in\mathbb{N}}$ converging towards f, as done in [48], and taking the limit $k \to \infty$ inside the MSE. At finite N, the loss achieved by estimator $\hat{\mathbf{x}}$ is thus given by the replica equations.

Note that our approach allows to continuate all scalar quantities, the MSE being among them, but we have no easy way to show that the Gaussian distribution of $\mathbf{B_1}/A_1$ predicted by state evolution for large λ_2 also extends to the original problem, therefore we have not shown that (3.53) holds for $\lambda_2 < \Lambda_2$.

Asymptotic limit and analyticity of the loss We now know the expression of the loss for all λ_2, for finite N. We are then tempted to take the asymptotic limit, but we encounter a subtle caveat. We have a sequence of real loss functions, all analytic in λ_2, described by the replica equations. However, taking their pointwise limit when $N \to \infty$ does not necessarily yield an analytic function in λ_2. If the asymptotic MSE in *not* analytic in λ_2, we have no clue about its expression for $\lambda_2 < \Lambda_2$, and our proof would only hold for λ_2 large enough when VAMP actually converges to estimator $\hat{\mathbf{x}}$. To properly conclude, we thus need one more assumption: *the MSE of estimator $\hat{\mathbf{x}}$ has to be analytic in λ_2*, in other words there should be no transition for $\lambda_2 < \Lambda_2$. This assumption is very reasonable, and proving it is left for future work, where one will need for instance to bound all derivatives of the loss [99].

3.7 Applications and numerical experiments

We want to compare the analytic expression of the error, obtained by the fixed point of SE (3.20) or equivalently the replica equations (3.4) with numerics on two typical problems. In both, the underlying truth vector $\mathbf{x_0}$ has i.i.d. elements sampled from a Gauss-Bernoulli distribution with sparsity parameter $\rho \in \mathbb{R}_+$, like in Chapter 2:

$$p_{x_0}(x_0) = (1-\rho)\delta(x_0) + \rho \frac{1}{\sqrt{2\pi}} \exp\left(-x_0^2/2\right), \tag{3.54}$$

and the training vector $\mathbf{y}$ is obtained as (3.2). Numerical experiments are performed using the Scikit-learn [126] implementation of the LASSO, which uses a coordinate descent method detailed in [55, 82].

3.7.1 Linear regression with row orthogonal matrices

In the first model, we use a LASSO regression by setting the penalty function to $f = \lambda_1 \|\cdot\|_1$, where λ_1 is the regularization parameter. We consider two types of matrices :

1. Gaussian i.i.d. matrices
2. Row orthogonal matrices, i.e. rotationally invariant matrices, with all singular values equal to 1. Such random matrices are very similar to subsampled Fourier and Hadamard matrices, and play a fundamental role in e.g. compressed sensing [167] and communication [69]. $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ then has the following eigenvalue distribution:

$$\lambda_{\mathbf{C}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha)\delta(1). \tag{3.55}$$

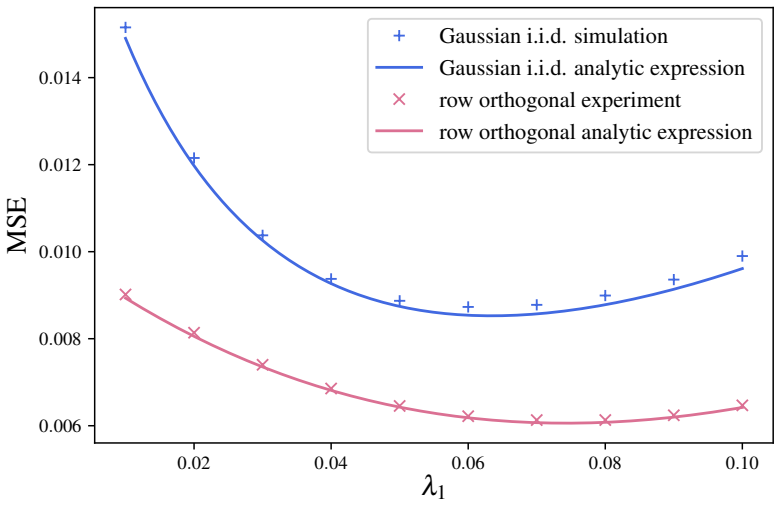


FIGURE 3.4: Mean squared error MSE $= \frac{1}{N}\|\mathbf{x_0} - \hat{\mathbf{x}}\|_2^2$ computed numerically (dots) for two types of matrices of size 200×100, for a Gauss-Bernoulli ground truth of sparsity $\rho = 0.3$ with noise $\Delta_0 = 0.3$, compared tp the analytical replica/state evolution asymptotic prediction (line). We use LASSO regression with penalty $f = \lambda_1\|\cdot\|_1$ and compute the MSE as function of λ_1. Numerical results match the analytic expression fairly well, despite the small size of the matrices.

In our simulations from Fig. 3.4, we take $M, N = 200, 100$ ($\alpha = 2$), $\Delta_0 = 0.01$ and $\rho = 0.3$. Each point is an average over 10^4 realizations. The error bars in this case are vanishingly small

($\approx 10^{-5}$). We see that an excellent agreement is obtained with the theoretical result, although the simulation matrices are rather small. Thus, the asymptotical setting is very practical, since it is already very well described with matrices with values of M, N of a few hundreds.

3.7.2 Overparametrization and double descent

In the second setup, we consider the effect of the aspect ratio $\alpha = M/N$ on the reconstruction performance of a sparse vector. We want to reproduce the double descent phenomenon that was observed and discussed recently in several papers [21, 71, 112, 103] in linear regression (but was already mentioned in [119]). In order to provide a minimal model of such a phenomenon, we follow the intuition proposed in [4]: to observe a divergence of the error, the eigenvalue distribution of $\mathbf{C}$ must have a singularity (but still be integrable) at $\lambda_1 = 0$ for $\alpha = 1$. Since we are using rotationally invariant matrices, we can design any spectrum with compact support that satisfies this criterion. We choose to sample the singular values of $\mathbf{F}$ from the uniform distribution $\mathcal{U}([(1-\alpha)^2, (1+\alpha)^2])$. We can compute the distribution of the eigenvalues of $\mathbf{C}$, in particular the non-zero ones are the squared singular values of $\mathbf{F}$. A little algebra yields:

$$\lambda_{\mathbf{C}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha) \left(\frac{1}{2((1+\alpha)^2 - (1-\alpha)^2)} \mathbb{I}_{\lambda_{\mathbf{C}} \in [(1-\alpha)^2, (1+\alpha)^2]} \frac{1}{\sqrt{\lambda_{\mathbf{C}}}} \right), \quad (3.56)$$

where $\mathbb{I}$ is the indicator function.

Our results are shown in Fig. 3.5 using $N = 250$, $\Delta_0 = 0.05$, for two values of the regularization parameter $\lambda_1 = 10^{-4}, 10^{-1}$. We recover the double descent with the very small regularization (light-colored curve). Note that the error peak could be moved to any point p on the x-axis by sampling singular values from the uniform distribution $\mathcal{U}([(p-\alpha)^2, (p+\alpha)^2])$. Multiple descents can also be obtained by adding several distributions of the form (3.56), with different shifts p. Augmenting the regularization to enforce a realistic LASSO penalty removes the error peak. As before, one observes striking agreement between the asymptotics and the simulation. Our formulas generalize here the results of [112] for any distribution of singular values.

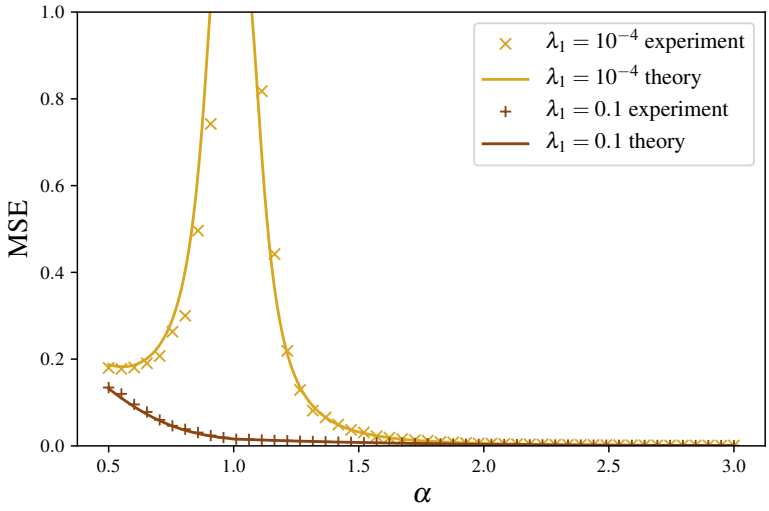


FIGURE 3.5: Mean squared error MSE $= \frac{1}{N}\|\mathbf{x_0} - \hat{\mathbf{x}}\|_2^2$ computed with LASSO regression $f = \lambda_1\|\cdot\|_1$ for two different regularization parameters. The matrices have width $N = 250$ and the MSE is plotted as a function of their aspect ratio α. The noise is $\Delta_0 = 0.05$. Each experiment point is an average over a hundred realizations. We use rotationally invariant matrices with singular values sampled from the uniform distribution on $[(1-\alpha)^2, (1+\alpha)^2]$. For $\lambda_1 = 10^{-4}$, we observe a peak around $\alpha = 1$ as predicted by the theory. Increasing the ℓ_1 regularization to $\lambda_1 = 0.1$ explicitly removes the peak to give a smooth curve.

Summary of Chapter 3 In this chapter, we have proved an exact asymptotic expression for the mean squared error of the estimator solving a penalized linear regression problem, using an analyticity assumption for the loss in the asymptotic limit. The data matrix is rotationally invariant, while the penalty function is convex, which includes the case of ridge regression or the LASSO, already discussed in Chapter 2 in the noiseless setting. The error is given by the replica formula which is derived in a heuristic way, but coincides with the rigorous state evolution fixed point of VAMP. Our method relies on proving the convergence of one carefully chosen instance of an oracle version of the algorithm, thanks to convex optimization tools. Numerical experiments show very good agreement between theory and simulation, even for small matrices ($N \approx$ a few hundreds).

Chapter 4

Asymptotic errors for teacher-student convex generalized linear models

4.1 Introduction of the problem

In Chapter 3, we dealt with linear regression with convex penalty, and relied on vector approximate message passing. We now step our game up, and would like to focus on generalized linear models, which add a non-linearity in the teacher vector. This chapter is adapted from [62]. The problem is defined as follows: we aim at reconstructing a given i.i.d. weight vector $\mathbf{x}_0 \in \mathbb{R}^N$ from outputs $\mathbf{y} \in \mathbb{R}^M$ generated using a training set $(\mathbf{f}_\mu)_{\mu=1,\dots,M}$ and the teacher rule:

$$\mathbf{y} = \phi(\mathbf{F}\mathbf{x}_0 + \boldsymbol{\omega_0}) \tag{4.1}$$

where ϕ is a proper, closed, convex and separable function, and $\boldsymbol{\omega_0} \sim \mathcal{N}(0, \Delta_0 \mathrm{Id})$ is an i.i.d. noise vector. We want to study the reconstruction performance of the generalized linear estimation method:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ g(\mathbf{F}\mathbf{x}, \mathbf{y}) + f(\mathbf{x}) \right\} \tag{4.2}$$

where g and f are proper, closed, convex and separable functions.

Let $\hat{\mathbf{x}}$ be the estimator of $\mathbf{x_0}$ defined in (4.2), and $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$ the estimator of $\mathbf{z_0} = \mathbf{F}\mathbf{x_0}$. Ground-truth vectors have norms $\rho_x \equiv \|\mathbf{x_0}\|_2^2/N$, $\rho_z \equiv \|\mathbf{z_0}\|_2^2/M$. We define the squared norms and the overlap of these estimators with the ground-truth:

$$m_x^* \equiv \lim_{N\to\infty} \frac{\hat{\mathbf{x}}^T\mathbf{x}_0}{N} \qquad m_z^* \equiv \lim_{M\to\infty} \frac{\hat{\mathbf{z}}^T\mathbf{z}_0}{M} \tag{4.3}$$

$$q_x^* \equiv \lim_{N\to\infty} \frac{\|\hat{\mathbf{x}}\|_2^2}{N} \qquad q_z^* \equiv \lim_{N\to\infty} \frac{\|\hat{\mathbf{z}}\|_2^2}{N} \tag{4.4}$$

We want an analytic expression of these quantities. With the knowledge of the asymptotic overlap m_x^*, and squared norms q_x^*, ρ_x, most quantities of interest can be estimated. For instance, the quadratic reconstruction error is obtained from its definition as $\mathrm{E} = \rho_x + q_x^* - 2m_x^*$, while the angle between the ground-truth vector and the estimator is $\theta = \arccos(m_x^*/(\sqrt{\rho_x q_x^*}))$. One can also estimate the generalization error for new random Gaussian samples [49], or compute similar errors for the denoising of $\mathbf{z_0}$.

The study of asymptotic (i.e. large-dimensional) reconstruction performance of generalized linear estimation in the teacher-student setting has been the subject of a significant body of work over the past few decades [145, 178, 49, 20, 48, 42, 184], and is currently witnessing a

renewal of interest especially for the case of Gaussian i.i.d. matrices, see e.g. [151, 71, 103]. We go beyond this setting by taking $\mathbf{F}$ rotationally invariant.

The simplest case of the present question, when both f and g are quadratic functions, can be mapped to a random matrix theory problem and solved rigorously, as in e.g. [71]. Handling non-linearity is, however, more challenging. A long history of research tackles this difficulty in the asymptotic limit, in particular using the replica method. In the case of Gaussian data, where the matrix $\mathbf{F}$ is Gaussian i.i.d., the asymptotic performance of the LASSO was rigorously derived in [19], and the existence of the logistic estimator discussed in [151]. A set of papers managed to extend this study to a large set of convex losses g, using the so-called Gordon comparison theorem [164]. As for rotationally invariant matrices, and for any convex and separable loss g and regularization f, a heuristic replica formula has been derived by Yoshiyuki Kabashima, providing a sharp analytical formula for the performance of reconstruction of the signal $\mathbf{x_0}$ [78].

Main assumptions We consider the minimization problem (4.2) with f and g proper closed, convex and separable functions, and $\mathbf{F} \in \mathbb{R}^{M\times N}$ a rotationally invariant matrix. We assume that the empirical distributions of the underlying truth $\mathbf{x}_0$ and eigenvalues of $\mathbf{F}^T\mathbf{F}$ respectively converge with second order moments, as defined in appendix C.1, to given distributions p_{x_0} and p_λ. We also assume that the distribution p_λ is non all-zero and has compact support. We focus on the limit $M, N \to \infty$ with fixed ratio $\alpha = M/N$.

4.2 Statistical physics result: the replica formula

4.2.1 Replica free energy

The scalar quantities that we want to characterize are given by the extremization of the replica-symmetric free energy associated to our problem, which is derived by Takashi and Kabashima[1] in [152], and reads:

$$\Phi = -\underset{m_x,\chi_x,q_x,m_z,\chi_z,q_z}{\text{extr}}\{g_{\text{F}} + g_{\text{G}} - g_{\text{S}}\}, \tag{4.5}$$

$$\begin{aligned}
g_{\text{F}} = &\underset{\hat{m}_{1x},\hat{\chi}_{1x},\hat{Q}_{1x},\hat{m}_{1z},\hat{\chi}_{1z},\hat{Q}_{1z}}{\text{extr}} \left\{\frac{1}{2}q_x\hat{Q}_{1x} - \frac{1}{2}\chi_x\hat{\chi}_{1x} - \hat{m}_{1x}m_x - \alpha\hat{m}_{1z}m_z + \frac{\alpha}{2}\left(q_z\hat{Q}_{1z} - \chi_z\hat{\chi}_{1z}\right)\right.\\
&\left. +\mathbb{E}\left[\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x})\right] + \alpha\mathbb{E}\left[\phi_z(\hat{m}_{1z},\hat{Q}_{1z},\hat{\chi}_{1z};z_0,\xi_{1z})\right]\right\},\\
g_{\text{G}} = &\underset{\hat{m}_{2x},\hat{\chi}_{2x},\hat{Q}_{2x},\hat{m}_{2z},\hat{\chi}_{2z},\hat{Q}_{2z}}{\text{extr}} \left\{\frac{1}{2}q_x\hat{Q}_{2x} - \frac{1}{2}\chi_x\hat{\chi}_{2x} - m_x\hat{m}_{2x} - \alpha m_z\hat{m}_{2z} + \frac{\alpha}{2}\left(q_z\hat{Q}_{2z} - \chi_z\hat{\chi}_{2z}\right)\right.\\
&\left. -\frac{1}{2}\left(\mathbb{E}\left[\log(\hat{Q}_{2x}+\lambda\hat{Q}_{2z})\right] - \mathbb{E}\left[\frac{\hat{\chi}_{2x}+\lambda\hat{\chi}_{2z}}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right] - \mathbb{E}\left[\frac{\rho_x(\hat{m}_{2x}+\lambda\hat{m}_{2z})^2}{(\hat{Q}_{2x}+\lambda\hat{Q}_{2z})}\right]\right)\right\},\\
g_{\text{S}} = &\frac{1}{2}\left(\frac{q_x}{\chi_x} - \frac{m_x^2}{\rho_x\chi_x}\right) + \frac{\alpha}{2}\left(\frac{q_z}{\chi_z} - \frac{m_z^2}{\rho_z\chi_z}\right),
\end{aligned}$$

[1] In fact, the first version of this formula was derived in 2008 in [78], but it is slightly misleading. In this first paper, the author performs an Itzykson-Zuber-Harish-Chandra integral which yields a function dependent on the eigenvalue distribution p_λ. However, this function can saturate, as seen for function $\mathcal{G}_{\mathbf{C}}$ in the replica calculation in appendix B. If saturation occurs, the result depends specifically on the larger eigenvalue of $\mathbf{F}^T\mathbf{F}$, and this eigenvalue has non-negligible deviation in the limit $N \to \infty$. In fact, [78] writes an annealed average on $\mathbf{F}$ by assuming that it is equal to the quenched average, but this is not necessarily true. This slight mishap happens because delta-functions are written as Fourier transforms inside the computation. Howevere in [152], the delta-functions as Gaussian functions with vanishigly small variances, and it allows to safely conclude the computation. We thus use the replica formula [152], but still credit the original and elegant derivation from [78].

where ϕ_x and ϕ_z are the potential functions

$$\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) = \lim_{\beta\to\infty} \frac{1}{\beta} \log \int e^{-\frac{\beta\hat{Q}_{1x}}{2}x^2 + \beta(\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - \beta f(x)} dx, \tag{4.6}$$

$$\phi_z(\hat{m}_{1z}, \hat{Q}_{1z}, \hat{\chi}_{1z}; z_0, \chi_{1z}) = \lim_{\beta\to\infty} \frac{1}{\beta} \log \int e^{-\frac{\beta\hat{Q}_{1z}}{2}z^2 + \beta(\hat{m}_{1z}z_0 + \sqrt{\hat{\chi}_{1z}}\xi_{1z})z - \beta g(y,z)} dz. \tag{4.7}$$

In this $\beta \to \infty$ limit (the so-called zero temperature limit in statistical physics), potentials ϕ_x and ϕ_z correspond to maximum a posteriori (MAP) estimation. Note that they are closely related to the Moreau envelopes [122, 18] of f and g. The Moreau envelope of a proper, closed and convex function f represents a smoothed convex function which shares the same minimizers as f, as shown on Fig. 4.1. It is defined for $\gamma \geq 0$ and $z \in \mathbb{R}$ as

$$\mathcal{M}_{\gamma f}(z) = \inf_x \left\{ f(x) + \frac{1}{2\gamma}\|x - z\|_2^2 \right\}. \tag{4.8}$$

ϕ_x and ϕ_z can also be written

$$\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) = \frac{\hat{Q}_{1x}}{2}X^2 - \mathcal{M}_{f/\hat{Q}_{1x}}(X) \tag{4.9}$$

$$\phi_z(\hat{m}_{1z}, \hat{Q}_{1z}, \hat{\chi}_{1z}; z_0, \xi_{1z}) = \frac{\hat{Q}_{1z}}{2}Z^2 - \mathcal{M}_{g(y,\cdot)/\hat{Q}_{1z}}(Z) \tag{4.10}$$

$$\text{where } X = \frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}} \quad \text{and} \quad Z = \frac{\hat{m}_{1z}z_0 + \sqrt{\hat{\chi}_{1z}}\xi_{1z}}{\hat{Q}_{1z}}.$$

This parallel with Moreau envelopes shows that the replica result extends the framework of [164], where the reconstruction performance of Gaussian generalized linear models is characterized using expected Moreau envelopes.

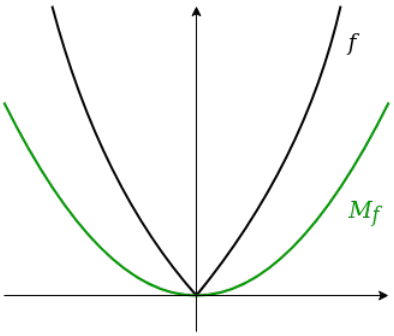


FIGURE 4.1: Moreau envelope $\mathcal{M}_f$ of a convex, non-differentiable function f.

4.2.2 Sketch of proof

We would like to prove the validity of the replica formula to characterize reconstruction performance. We will proceed as we did in Chapter 3, with the help of an algorithm that has an analytical description. This time, we use *multi-layer generalized vector approximate message passing* (MLVAMP) [54], the big brother of VAMP that applies to our setting in its two-layer version. [54] provides a rigorous proof of *state evolution equations* for MLVAMP, if they are properly initialized. Our proof involves the following steps:

i) We show that the state evolution fixed point of MLVAMP's state evolution matches the replica result,

ii) We verify that the sequence's fixed point reaches the estimator $\hat{\mathbf{x}}$,

iii) We determine the conditions for this sequence to be provably convergent.

4.3 MLVAMP and its state evolution

4.3.1 MAP formulation of two-layer VAMP

For our minimization problem, we need to solve the maximum a posteriori (MAP) version of MLVAMP. It is similar to a multilayer proximal descent method, and can be viewed as successively solving an alternating direction method of multipliers (ADMM) [25] step for each layer of the model, with an additional LMMSE (linear minimum mean squared error) step on each layer. As pointed out in [52], the main difference between MLVAMP and standard convex optimization methods are the implicit prescription of descent step sizes and prefactors at each iteration through variance parameters that are computed at each iteration. The latter adapt to the local curvature of denoiser functions.

We start by giving the full iterations of the MLVAMP algorithm from [54] applied to a 2-layer network. For a given operator $T : \mathbb{R}^d \to \mathbb{R}^d$, the brackets $\langle T(\mathbf{x})\rangle = \frac{1}{d}\sum_{i=1}^{d} T(\mathbf{x})_i$ denote element-wise averaging operations, where d is M or N in our case. For a given matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, the brackets amount to $\langle \mathbf{M}\rangle = \frac{1}{d}\mathrm{Tr}(\mathbf{M})$.

Initialize $\mathbf{h}_{1x}^{(0)}, \mathbf{h}_{2z}^{(0)}$ *isotropically*

Forward pass – denoising ($L=0$)	*Forward pass - LMMSE* ($L=1$)	
$\hat{\mathbf{x}}_1^{(t)} = g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$	$\hat{\mathbf{z}}_2^{(t)} = g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$	(4.11a)
$\chi_{1x}^{(t)} = \frac{1}{\hat{Q}_{1x}^{(t)}}\left\langle \frac{\partial g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})}{\partial \mathbf{h}_{1x}^{(t)}} \right\rangle$	$\chi_{2z}^{(t)} = \frac{1}{\hat{Q}_{2z}^{(t)}}\left\langle \frac{\partial g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})}{\partial \mathbf{h}_{2z}^{(t)}} \right\rangle$	(4.11b)
$\hat{Q}_{2x}^{(t)} = 1/\chi_{1x}^{(t)} - \hat{Q}_{1x}^{(t)}$	$\hat{Q}_{1z}^{(t)} = 1/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}$	(4.11c)
$\mathbf{h}_{2x}^{(t)} = (\hat{\mathbf{x}}_1^{(t)}/\chi_{1x}^{(t)} - \hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)})/\hat{Q}_{2x}^{(t)}$	$\mathbf{h}_{1z}^{(t)} = (\hat{\mathbf{z}}_2^{(t)}/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)})/\hat{Q}_{1z}^{(t)}$	(4.11d)

Backward pass – denoising ($L=1$)	*Backward pass - LMMSE* ($L=0$)	
$\hat{\mathbf{z}}_1^{(t)} = g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)})$	$\hat{\mathbf{x}}_2^{(t+1)} = g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)})$	(4.11e)
$\chi_{1z}^{(t)} = \frac{1}{\hat{Q}_{1z}^{(t)}}\left\langle \frac{\partial g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)})}{\partial \mathbf{h}_{1z}^{(t)}} \right\rangle$	$\chi_{2x}^{(t+1)} = \frac{1}{\hat{Q}_{2x}^{(t)}}\left\langle \frac{\partial g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)})}{\partial \mathbf{h}_{2x}^{(t)}} \right\rangle$	(4.11f)
$\hat{Q}_{2z}^{(t+1)} = 1/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}$	$\hat{Q}_{1x}^{(t+1)} = 1/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}$	(4.11g)
$\mathbf{h}_{2z}^{(t+1)} = (\hat{\mathbf{z}}_1^{(t)}/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)})/\hat{Q}_{2z}^{(t+1)}$	$\mathbf{h}_{1x}^{(t+1)} = (\hat{\mathbf{x}}_2^{(t+1)}/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)})/\hat{Q}_{1x}^{(t+1)}$.	(4.11h)

Denoiser functions g_{1x} and g_{1z} can be written as proximal operators in the MAP setting:

$$g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) = \underset{\mathbf{x}\in\mathbb{R}^{\mathbf{N}}}{\arg\min}\left\{ f(\mathbf{x}) + \frac{\hat{Q}_{1x}^{(t)}}{2}\|\mathbf{x} - \mathbf{h}_{1x}^{(t)}\|_2^2 \right\} = \mathrm{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)}) \tag{4.12}$$

$$g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) = \underset{\mathbf{z}\in\mathbb{R}^{M}}{\arg\min}\left\{ g(\mathbf{y}, \mathbf{z}) + \frac{\hat{Q}_{1z}^{(t)}}{2}\|\mathbf{z} - \mathbf{h}_{1z}^{(t)}\|_2^2 \right\} = \mathrm{Prox}_{g(.,\mathbf{y})/\hat{Q}_{1z}^{(t)}}(\mathbf{h}_{1z}^{(t)}). \tag{4.13}$$

To shorten notations, the scalar element-wise version of these separable proximal operators will be called η_f and $\eta_{g(y,\cdot)}$. LMMSE denoisers g_{2z} and g_{2x} in the MAP setting read (see [132]):

$$g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)}) = \mathbf{F}(\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{h}_{2z}^{(t)}) \tag{4.14}$$

$$g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}) = (\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{h}_{2z}^{(t+1)}). \tag{4.15}$$

At each iteration, MLVAMP returns two estimators for $\hat{\mathbf{x}}$ and two estimators for $\hat{\mathbf{z}}$, which are the hat vectors $\hat{\mathbf{x}}_\mathbf{1}$, $\hat{\mathbf{x}}_\mathbf{2}$, $\hat{\mathbf{z}}_\mathbf{1}$ and $\hat{\mathbf{z}}_\mathbf{2}$. The variance parameters are the $\hat{Q}_{1x}, \hat{Q}_{2x}, \hat{Q}_{1z}, \hat{Q}_{2z}$ and $\chi_{1x}, \chi_{2x}, \chi_{1z}, \chi_{2z}$ (they play the same role as A_1, A_2 and V_1, V_2 did for VAMP). Finally $\mathbf{h_{1x}}, \mathbf{h_{2x}}, \mathbf{h_{1z}}, \mathbf{h_{2z}}$ are intermediate vectors, and applying them denoiser functions with adapted variance parameters yields the hat estimators. We are sticking to notations from [152], which differ from the ones used in [132] and [54], but we will provide a dictionary in appendix D.2.

4.3.2 Equality of $\hat{\mathbf{x}}$ and MLVAMP's fixed point

We would like the fixed point of MLVAMP to satisfy the following first-order optimality condition

$$\partial f(\hat{\mathbf{x}}) + \mathbf{F}^T\partial g(\mathbf{F}\hat{\mathbf{x}}) = 0, \tag{4.16}$$

where $g(\cdot)$ designates $g(\cdot, \mathbf{y})$. This condition characterizes the unique minimizer of the convex problem (4.2). Writing the fixed point of the scalar parameters of the iterations (4.11), we get the following prescriptions on the scalar quantities:

$$\frac{1}{\chi_x} \equiv \frac{1}{\chi_{1x}} = \frac{1}{\chi_{2x}} = \hat{Q}_{1x} + \hat{Q}_{2x} \qquad \frac{1}{\chi_z} \equiv \frac{1}{\chi_{1z}} = \frac{1}{\chi_{2z}} = \hat{Q}_{1z} + \hat{Q}_{2z} \tag{4.17}$$

$$\hat{Q}_{1x}\chi_{1x} + \hat{Q}_{2x}\chi_{2x} = 1 \qquad \hat{Q}_{1z}\chi_{1z} + \hat{Q}_{2z}\chi_{2z} = 1. \tag{4.18}$$

Replacing $\mathbf{h}_{1x}$'s expression inside $\mathbf{h}_{2x}$ reads

$$\mathbf{h}_{2x} = \left(\frac{\hat{\mathbf{x}}_1}{\chi_x} - \hat{Q}_{1x}\mathbf{h}_{1x}\right)/\hat{Q}_{2x} = \left(\frac{\hat{\mathbf{x}}_1}{\chi_x} - \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - \hat{Q}_{2x}\mathbf{h}_{2x}\right)\right)/\hat{Q}_{2x} \tag{4.19}$$

and using (4.17) we get $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2$, and a similar reasoning gives $\hat{\mathbf{z}}_1 = \hat{\mathbf{z}}_2$. From (4.14) and (4.15), we clearly find $\hat{\mathbf{z}}_2 = \mathbf{F}\hat{\mathbf{x}}_\mathbf{2}$. Inverting the proximal operators in (4.12) and (4.13) yields

$$\hat{\mathbf{x}}_1 + \frac{1}{\hat{Q}_{1x}}\partial f(\hat{\mathbf{x}}_1) = \mathbf{h}_{1x} \tag{4.20}$$

$$\hat{\mathbf{z}}_1 + \frac{1}{\hat{Q}_{1z}}\partial g(\hat{\mathbf{z}}_1) = \mathbf{h}_{1z}. \tag{4.21}$$

We then take the MLVAMP equation on $\mathbf{h}_{1x}$, we write

$$\mathbf{h}_{1x} = \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - \hat{Q}_{2x}\mathbf{h}_{2x}\right)/\hat{Q}_{1x} \tag{4.22}$$

$$= \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - (\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})\hat{\mathbf{x}}_2 + \hat{Q}_{2z}\mathbf{F}^T\mathbf{h}_{2z}\right)/\hat{Q}_{1x} \tag{4.23}$$

$$= -\left(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\left(1 - \frac{1}{\chi_x\hat{Q}_{2x}}\right)\mathrm{Id}\right)\frac{\hat{\mathbf{x}}_2}{\hat{Q}_{2x}} + \mathbf{F}^T\left(\hat{Q}_{1z}\left(\frac{1}{\chi_z\hat{Q}_{1z}} - 1\right)\hat{\mathbf{z}}_1 - \partial\mathbf{g}(\hat{\mathbf{z}}_1)\right) \tag{4.24}$$

which is equal to the left-hand term in (4.20). Using this equality, as well as $\hat{\mathbf{z}}_1 = \mathbf{F}\hat{\mathbf{x}}_\mathbf{1}$ and relations (4.17) and (4.18) yields

$$\partial f(\hat{\mathbf{x}}_1) + \mathbf{F}^T\partial g(\mathbf{F}\hat{\mathbf{x}}_\mathbf{1}) = 0. \tag{4.25}$$

Hence, the fixed point of MLVAMP satisfies the optimality condition (4.16) and is indeed the desired MAP estimator: $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 = \hat{\mathbf{x}}$. This shows point ii of our sketch of proof.

4.3.3 State evolution of MLVAMP and its fixed point

We have two versions of state evolution equations in our hands. The first one comes from [54], and is proven in the asymptotic limit, but somewhat difficult to handle. We refer to it as (SE1). The second version is derived in [152] directly from MLVAMP's equations stated in (4.11), and in the same notations that we have adopted. We call this one (SE2). To stick with our notations, we will prefer the equations from (SE2), however they suffer from a drawback: they are derived with a more physical approach relying on a non-proven assumption, but we want to invoke SE equations that are exact for the needs of our proof. We thus have two ways of proceeding: we could directly show that (SE1) and (SE2) are exactly the same set of equations at their fixed point. However, this turns out to be a hassle (although we explain how to do it in appendix D.2.3), so we turn to an easier trick: we could simply show that the starting assumption of (SE2) is implied by the rigorous formulation of (SE1), which means that (SE2) becomes rigorous too, and we can use it to our hearts' content. This is done in appendix D.2.1.

Gaussian property assumption
[152] starts by assuming that $\mathbf{h_{1x}}, \mathbf{h_{1z}}, \mathbf{h_{2x}}, \mathbf{h_{2z}}$ behave as Gaussian estimates:

$$\hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)} - \hat{m}_{1x}^{(t)}\mathbf{x_0} \overset{\mathrm{d}}{=} \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)} \tag{4.26a}$$

$$\mathbf{V}^T(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)}\mathbf{x_0}) \overset{\mathrm{d}}{=} \sqrt{\hat{\chi}_{2x}^{(t)}}\xi_{2x}^{(t)} \tag{4.26b}$$

$$\mathbf{U}^T(\hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)} - \hat{m}_{1z}^{(t)}\mathbf{z_0}) \overset{\mathrm{d}}{=} \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)} \tag{4.26c}$$

$$\hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)} - \hat{m}_{2z}^{(t)}\mathbf{z_0} \overset{\mathrm{d}}{=} \sqrt{\hat{\chi}_{2z}^{(t)}}\xi_{2z}^{(t)} \tag{4.26d}$$

where $\overset{\mathrm{d}}{=}$ denotes equality of empirical distributions. $\mathbf{U}$ and $\mathbf{V}$ come from the singular value decomposition $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ and are Haar-sampled; $\xi_{1x}^{(t)}, \xi_{2x}^{(t)}, \xi_{1z}^{(t)}, \xi_{2z}^{(t)}$ are normal Gaussian vectors, independent from $\mathbf{x_0}, \mathbf{z_0}, \mathbf{V}^T\mathbf{x_0}$ and $\mathbf{U}^T\mathbf{z_0}$. We show that this assumption is exact in the asymptotic limit in appendix D.2.1.

Recall that parameters $\hat{Q}_{1x}^{(t)}, \hat{Q}_{1z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)}$ are defined through MLVAMP's iterations (4.11). Parameters $\hat{m}_{1x}^{(t)}, \hat{m}_{1z}^{(t)}, \hat{m}_{2x}^{(t)}, \hat{m}_{2z}^{(t)}$ and $\hat{\chi}_{1x}^{(t)}, \hat{\chi}_{1z}^{(t)}, \hat{\chi}_{2x}^{(t)}, \hat{\chi}_{2z}^{(t)}$ will be prescribed through SE equations. Other useful variables are the overlaps and squared norms of estimators, for $k \in \{1, 2\}$:

$$m_{kx}^{(t)} = \frac{\mathbf{x}_0^\top \hat{\mathbf{x}}_k^{(t)}}{N} \qquad q_{kx}^{(t)} = \frac{\|\hat{\mathbf{x}}_k^{(t)}\|_2^2}{N}$$
$$m_{kz}^{(t)} = \frac{\mathbf{z}_0^\top \hat{\mathbf{z}}_k^{(t)}}{M} \qquad q_{kz}^{(t)} = \frac{\|\hat{\mathbf{z}}_k^{(t)}\|_2^2}{M}.$$

The state evolution equations are scalar and describe the evolution of these 16 quantities throughout MLVAMP's iterations. This might seem very complicated to handle, but these quantities carry physical meaning which makes them friendlier to the scared-off reader, and SE equations turn out to be perfectly compatible with numerical implementation. Note that they involve the scalar proximals η_f and $\eta_{g(y,\cdot)}$. Expectations are taken with respect to the random variables $x_0 \sim p_{x_0}$, $z_0 \sim \mathcal{N}(0, \sqrt{\rho_z})$, $y \sim \phi(z_0 + \omega_0)$, $\xi_{1x}, \xi_{1z} \sim \mathcal{N}(0, 1)$, and eigenvalues $\lambda \sim p_\lambda$.

Starting from assumptions (4.26), and following the derivation of [152] adapted to the iteration order from (4.11), the scalar state evolution equations read:

Initialize $\hat{Q}_{1x}^{(0)}, \hat{Q}_{2z}^{(0)}, \hat{m}_{1x}^{(0)}, \hat{m}_{2z}^{(0)}, \hat{\chi}_{1x}^{(0)}, \hat{\chi}_{2z}^{(0)} > 0.$

$$m_{1x}^{(t)} = \mathbb{E}\left[x_0 \eta_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)} x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{4.27a}$$

$$\chi_{1x}^{(t)} = \frac{1}{\hat{Q}_{1x}^{(t)}}\mathbb{E}\left[\eta'_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)} x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{4.27b}$$

$$q_{1x}^{(t)} = \mathbb{E}\left[\eta^2_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)} x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{4.27c}$$

$$\hat{Q}_{2x}^{(t)} = \frac{1}{\chi_{1x}^{(t)}} - \hat{Q}_{1x}^{(t)} \tag{4.27d}$$

$$\hat{m}_{2x}^{(t)} = \frac{m_{1x}^{(t)}}{\rho_x \chi_{1x}^{(t)}} - \hat{m}_{1x}^{(t)} \tag{4.27e}$$

$$\hat{\chi}_{2x}^{(t)} = \frac{q_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \frac{(m_{1x}^{(t)})^2}{\rho_x (\chi_{1x}^{(t)})^2} - \hat{\chi}_{1x}^{(t)} \tag{4.27f}$$

$$m_{2z}^{(t)} = \frac{\rho_x}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t)})}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)}}\right] \tag{4.27g}$$

$$\chi_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)}}\right] \tag{4.27h}$$

$$q_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{\chi}_{2x}^{(t)} + \lambda \hat{\chi}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)})^2}\right] + \frac{\rho_x}{\alpha}\mathbb{E}_\lambda\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t)})^2}\right] \tag{4.27i}$$

$$\hat{Q}_{1z}^{(t)} = \frac{1}{\chi_{2z}^{(t)}} - \hat{Q}_{2z}^{(t)} \tag{4.27j}$$

$$\hat{m}_{1z}^{(t)} = \frac{m_{2z}^{(t)}}{\rho_z \chi_{2z}^{(t)}} - \hat{m}_{2z}^{(t)} \tag{4.27k}$$

$$\hat{\chi}_{1z}^{(t)} = \frac{q_{2z}^{(t)}}{(\chi_{2z}^{(t)})^2} - \frac{(m_{2z}^{(t)})^2}{\rho_z (\chi_{2z}^{(t)})^2} - \hat{\chi}_{2z}^{(t)} \tag{4.27l}$$

$$m_{1z}^{(t)} = \mathbb{E}\left[z_0 \eta_{g(y,.)/\hat{Q}_{1z}^{(t)}}\left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\right)\right] \tag{4.27m}$$

$$\chi_{1z}^{(t)} = \frac{1}{\hat{Q}_{1z}^{(t)}}\mathbb{E}\left[\eta'_{g(y,.)/\hat{Q}_{1z}^{(t)}}\left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\right)\right] \tag{4.27n}$$

$$q_{1z}^{(t)} = \mathbb{E}\left[\eta^2_{g(y,.)/\hat{Q}_{1z}^{(t)}}\left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\right)\right] \tag{4.27o}$$

$$\hat{Q}_{2z}^{(t+1)} = \frac{1}{\chi_{1z}^{(t)}} - \hat{Q}_{1z}^{(t)} \tag{4.27p}$$

$$\hat{m}_{2z}^{(t+1)} = \frac{m_{1z}^{(t)}}{\rho_z \chi_{1z}^{(t)}} - \hat{m}_{1z}^{(t)} \tag{4.27q}$$

$$\hat{\chi}_{2z}^{(t+1)} = \frac{q_{1z}^{(t)}}{(\chi_{1z}^{(t)})^2} - \frac{(m_{1z}^{(t)})^2}{\rho_z (\chi_{1z}^{(t)})^2} - \hat{\chi}_{1z}^{(t)} \tag{4.27r}$$

$$m_{2x}^{(t+1)} = \rho_x \mathbb{E}\left[\frac{\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t+1)}}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)}}\right] \tag{4.27s}$$

$$\chi_{2x}^{(t+1)} = \mathbb{E}\left[\frac{1}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)}}\right] \tag{4.27t}$$

$$q_{2x}^{(t+1)} = \mathbb{E}\left[\frac{\hat{\chi}_{2x}^{(t)} + \lambda \hat{\chi}_{2z}^{(t+1)}}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)})^2}\right] + \rho_x \mathbb{E}\left[\frac{(\hat{m}_{2x}^{(t+1)} + \lambda \hat{m}_{2z}^{(t+1)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)})^2}\right] \tag{4.27u}$$

$$\hat{Q}_{1x}^{(t+1)} = \frac{1}{\chi_{2x}^{(t+1)}} - \hat{Q}_{2x}^{(t)} \tag{4.27v}$$

$$\hat{m}_{1x}^{(t+1)} = \frac{m_{2x}^{(t+1)}}{\rho_x \chi_{2x}^{(t+1)}} - \hat{m}_{2x}^{(t)} \tag{4.27w}$$

$$\hat{\chi}_{1x}^{(t+1)} = \frac{q_{2x}^{(t+1)}}{(\chi_{2x}^{(t+1)})^2} - \frac{(m_{2x}^{(t+1)})^2}{\rho_x (\chi_{2x}^{(t+1)})^2} - \hat{\chi}_{2x}^{(t)}. \tag{4.27x}$$

The fixed point of those state evolution equations exactly coincides with the extremization conditions of the replica free energy (4.5), which shows point i of our sketch of proof.

4.4 Oracle MLVAMP as a dynamical system

4.4.1 Definition of Oracle MLVAMP

We move on to the hardest part of the proof, i.e. point iii. Just like we did in 3.4, we will introduce an oracle version of MLVAMP. Indeed, we only need to show that there exists one instance of the algorithm that converges to its fixed point, and we have the freedom to choose initialization, as long as it satisfies the criteria that keep state evolution equations valid. In the oracle algorithm, second-order parameters, i.e. the implicit step-sizes and prefactors of the denoisers, are prescribed from the fixed point of the state evolution equations. In our notations, these parameters correspond to $\hat{Q}_{1x}, \hat{Q}_{1z}, \hat{Q}_{2x}, \hat{Q}_{2z}, \chi_x, \chi_z$. The Oracle-MLVAMP iterations then read:

Initialize $\mathbf{h}_{1x}^{(0)} \mathbf{h}_{2z}^{(0)}$ isotropically, prescribe $\hat{Q}_{1x}, \hat{Q}_{1z}, \hat{Q}_{2x}, \hat{Q}_{2z}, \chi_x, \chi_z$.

Forward pass – denoising	*Forward pass – LMMSE*	
$\hat{\mathbf{x}}_1^{(t)} = \mathrm{Prox}_{f/\hat{Q}_{1x}}(\mathbf{h}_{1x}^{(t)})$	$\hat{\mathbf{z}}_2^{(t)} = \mathbf{F}(\hat{Q}_{2x}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})^{-1}(\hat{Q}_{2x}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}\mathbf{F}^T\mathbf{h}_{2z}^{(t)})$	(4.28a)
$\mathbf{h}_{2x}^{(t)} = (\hat{\mathbf{x}}_1^{(t)}/\chi_x - \hat{Q}_{1x}\mathbf{h}_{1x}^{(t)})/\hat{Q}_{2x}$	$\mathbf{h}_{1z}^{(t)} = (\hat{\mathbf{z}}_2^{(t)}/\chi_z - \hat{Q}_{2z}\mathbf{h}_{2z}^{(t)})/\hat{Q}_{1z}$	(4.28b)
Backward pass – denoising	*Backward pass – LMMSE*	
$\hat{\mathbf{z}}_1^{(t)} = \mathrm{Prox}_{g(.,y)/\hat{Q}_{1z}}(\mathbf{h}_{1z}^{(t)})$	$\hat{\mathbf{x}}_2^{(t+1)} = (\hat{Q}_{2x}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2z}\mathrm{Id})^{-1}(\hat{Q}_{2x}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}\mathbf{F}^T\mathbf{h}_{2z}^{(t+1)})$	(4.28c)
$\mathbf{h}_{2z}^{(t+1)} = (\hat{\mathbf{z}}_1^{(t)}/\chi_z - \hat{Q}_{1z}\mathbf{h}_{1z}^{(t)})/\hat{Q}_{2z}$	$\mathbf{h}_{1x}^{(t+1)} = (\hat{\mathbf{x}}_2^{(t+1)}/\chi_x - \hat{Q}_{2x}\mathbf{h}_{2x}^{(t)})/\hat{Q}_{1x}$.	(4.28d)

At each iteration, Oracle-MLVAMP returns two sets of estimators $(\hat{\mathbf{x}}_1^{(t)}, \hat{\mathbf{x}}_2^{(t)})$ and $(\hat{\mathbf{z}}_1^{(t)}, \hat{\mathbf{z}}_2^{(t)})$ which respectively aim at reconstructing the minimizer $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. At the fixed point, we have $\hat{\mathbf{x}}_1^{(t)} = \hat{\mathbf{x}}_2^{(t)}$ and $\hat{\mathbf{z}}_1^{(t)} = \hat{\mathbf{z}}_2^{(t)}$. The constant parameters verify the fixed point relations (4.17)

and (4.18). Of course, the fixed point of Oracle MLVAMP coincides with the fixed point of MLVAMP, i.e. the minimizer $\hat{\mathbf{x}}$.

4.4.2 Compact form of Oracle MLVAMP

We want to write Oracle MLVAMP in a compact way, to exhibit the main operators and try to characterize their Lipschitz constants. When studying Oracle VAMP, it could be reduced to the composition of two operators (3.23). For MLVAMP, it will be of course more complicated. We write it as two equations on vectors $\mathbf{h}_{1x}^{(t)}, \mathbf{h}_{2z}^{(t)}$:

$$\begin{aligned}
&\text{Initialize } \mathbf{h}_{1x}^{(0)}, \mathbf{h}_{2z}^{(0)} \\
&\mathbf{h}_{1x}^{(t+1)} = \mathbf{W_1}\tilde{\mathcal{O}}_\mathbf{1}\mathbf{h}_{1x}^{(t)} + \mathbf{W_2}\tilde{\mathcal{O}}_\mathbf{2}(\mathbf{W_3}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4\tilde{\mathcal{O}}_1\mathbf{h}_{1x}^{(t)}) && (4.29)\\
&\mathbf{h}_{2z}^{(t+1)} = \tilde{\mathcal{O}}_\mathbf{2}(\mathbf{W_3}\mathbf{h}_{2z}^{(t)} + \mathbf{W_4}\tilde{\mathcal{O}}_\mathbf{1}\mathbf{h}_{1x}^{(t)}) && (4.30)
\end{aligned}$$

where

$$\mathbf{W_1} = \frac{\hat{Q}_{2x}}{\hat{Q}_{1x}}\left(\frac{1}{\chi_x}(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})^{-1} - \mathrm{Id}\right) \qquad \mathbf{W_2} = \frac{\hat{Q}_{2z}}{\chi_x\hat{Q}_{1x}}(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})^{-1}\mathbf{F}^T \quad (4.31a)$$

$$\mathbf{W_3} = \frac{\hat{Q}_{2z}}{\hat{Q}_{1z}}\left(\frac{1}{\chi_z}\mathbf{F}(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})^{-1}\mathbf{F}^T - \mathrm{Id}\right) \qquad \mathbf{W_4} = \frac{\hat{Q}_{2x}}{\hat{Q}_{1z}\chi_z}\mathbf{F}(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})^{-1} \quad (4.31b)$$

$$\tilde{\mathbf{O}}_\mathbf{1} = \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\left(\frac{1}{\chi_x\hat{Q}_{1x}}\mathrm{Prox}_{f/\hat{Q}_{1x}}(\cdot) - \mathrm{Id}\right) \qquad \tilde{\mathbf{O}}_\mathbf{2} = \frac{\hat{Q}_{1z}}{\hat{Q}_{1z}}\left(\frac{1}{\chi_z\hat{Q}_{1z}}\mathrm{Prox}_{g(.,y)/\hat{Q}_{1z}}(\cdot) - \mathrm{Id}\right). \quad (4.31c)$$

4.4.3 Recast of Oracle VAMP as a linear system

This system is somewhat hard to handle and includes non-linear operators. We will follow the approach pioneered in [92], where the main idea is to recast any non-linear dynamical system as a linear one. In the linear system, convergence will be naturally characterized by a matrix norm. The linear recast works this way: for a given non-linearity $\tilde{\mathcal{O}}$ applied to an iterate $\mathbf{h}$, we define the variable $\mathbf{u} = \tilde{\mathcal{O}}(\mathbf{h})$ and rewrite the initial algorithm in terms of this trivial transform. Any property of $\tilde{\mathcal{O}}$ is then summarized in a constraint matrix linking $\mathbf{h}$ and $\mathbf{u}$. For example, if $\tilde{\mathcal{O}}$ has Lipschitz constant ω, then for all t:

$$\|\mathbf{u}_{t+1} - \mathbf{u}_t\|_2^2 \leqslant \omega^2\|\mathbf{h}_{t+1} - \mathbf{h}_t\|_2^2, \tag{4.32}$$

which can be rewritten in matrix form:

$$\begin{bmatrix}\mathbf{h}_{t+1} - \mathbf{h}_t \\ \mathbf{u}_{t+1} - \mathbf{u}_t\end{bmatrix}^T \begin{bmatrix}\omega^2\mathbf{I}_{\mathbf{d_h}} & 0 \\ 0 & -\mathbf{I}_{\mathbf{d_u}}\end{bmatrix} \begin{bmatrix}\mathbf{h}_{t+1} - \mathbf{h}_t \\ \mathbf{u}_{t+1} - \mathbf{u}_t\end{bmatrix} \geqslant 0 \tag{4.33}$$

where $\mathbf{I}_{\mathbf{d_h}}, \mathbf{I}_{\mathbf{d_u}}$ are the identity matrices with dimensions of $\mathbf{u}, \mathbf{h}$, i.e. M or N in our case. The matrix encapsules the Lipschitz property of the non-linear operator. Any cocoercivity property (verified by proximal operators) can also be rewritten in matrix form but yields non block diagonal constraint matrices. For our proof, we will derive the Lipschitz constants of non-linear operators instead of focusing on their cocoercivity, as it turns out simpler and sufficient to prove our point. After obtaining the linear formulation of our system, the condition of convergence is given by the main theorem from [92] (adapted to ADMM in [117]) through a linear matrix inequality.

Let us write the recast of Oracle MLVAMP as a linear system. We define the variables:

$$\mathbf{u}_0^{(t)} = \tilde{\mathcal{O}}_1(\mathbf{h}_{1x}^{(t)}), \qquad \mathbf{v}^{(t)} = \mathbf{W}_3\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4\mathbf{u}_0^{(t)}, \quad \text{and} \quad \mathbf{u}_1^{(t)} = \tilde{\mathcal{O}}_2(\mathbf{v}^{(t)}) \tag{4.34}$$

$$\text{such that} \quad \mathbf{h}_{2z}^{(t+1)} = \mathbf{u}_1^{(t)}, \qquad \mathbf{h}_{1x}^{(t+1)} = \mathbf{W}_1\mathbf{u}_0^{(t)} + \mathbf{W}_2\mathbf{u}_1^{(t)}. \tag{4.35}$$

where $\mathbf{u}_0, \mathbf{h}_{1x} \in \mathbb{R}^N$, $\mathbf{v}, \mathbf{u}_1, \mathbf{h}_{2z} \in \mathbb{R}^M$. We then write

$$\mathbf{h}^{(t)} = \begin{bmatrix} \mathbf{h}_{2z}^{(t)} \\ \mathbf{h}_{1x}^{(t)} \end{bmatrix}, \quad \mathbf{u}^{(t)} = \begin{bmatrix} \mathbf{u}_1^{(t)} \\ \mathbf{u}_0^{(t)} \end{bmatrix}, \quad \mathbf{z}_0^{(t)} = \begin{bmatrix} \mathbf{h}_{1x}^{(t)} \\ \mathbf{u}_0^{(t)} \end{bmatrix}, \quad \mathbf{z}_1^{(t)} = \begin{bmatrix} \mathbf{v}^{(t)} \\ \mathbf{u}_1^{(t)} \end{bmatrix}.$$

This leads to the following linear dynamical system recast of (4.29):

$$\mathbf{h}^{(t+1)} = \mathbf{A}\mathbf{h}^{(t)} + \mathbf{B}\mathbf{u}^{(t)} \tag{4.36}$$

$$\mathbf{z}_1^{(t)} = \mathbf{C}_1\mathbf{h}^{(t)} + \mathbf{D}_1\mathbf{u}^{(t)} \tag{4.37}$$

$$\mathbf{z}_2^{(t)} = \mathbf{C}_2\mathbf{h}^{(t)} + \mathbf{D}_2\mathbf{u}^{(t)} \tag{4.38}$$

where

$$\mathbf{A} = \mathbf{0}_{(\mathbf{M}+\mathbf{N})\times(\mathbf{M}+\mathbf{N})} \qquad \mathbf{B} = \begin{bmatrix} \mathbf{I_M} & \mathbf{0_{M\times N}} \\ \mathbf{W}_2 & \mathbf{W}_1 \end{bmatrix} \tag{4.39}$$

$$\mathbf{C}_1 = \begin{bmatrix} \mathbf{0_{N\times M}} & \mathbf{I_N} \\ \mathbf{0_{N\times M}} & \mathbf{0_{N\times N}} \end{bmatrix} \qquad \mathbf{D}_1 = \begin{bmatrix} \mathbf{0_{N\times M}} & \mathbf{0_{N\times N}} \\ \mathbf{0_{N\times M}} & \mathbf{I_N} \end{bmatrix} \tag{4.40}$$

$$\mathbf{C}_2 = \begin{bmatrix} \mathbf{W}_3 & \mathbf{0_{M\times N}} \\ \mathbf{0_{M\times M}} & \mathbf{0_{M\times N}} \end{bmatrix} \qquad \mathbf{D}_2 = \begin{bmatrix} \mathbf{0_{M\times M}} & \mathbf{W}_4 \\ \mathbf{I_M} & \mathbf{0_{M\times N}} \end{bmatrix}. \tag{4.41}$$

4.4.4 Lipschitz constants and constraint matrices

The next step is to impose the properties of the non-linearities $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$ through constraint matrices. The Lipschitz constants ω_1, ω_2 of $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$ can be determined using properties of proximal operators [64] and are directly linked to the strong convexity and smoothness of the cost function and regularization. Let (σ_1, β_1) and (σ_2, β_2) the the strong convexity and smoothness constants (that can be taken infinite if there is no smoothness assumption) of f and $g(y, \cdot)$. An upper bound on the Lipschitz constants then reads:

$$\omega_1 = \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\sqrt{1 + \frac{\hat{Q}_{2x}^2 - \hat{Q}_{1x}^2}{(\hat{Q}_{1x} + \sigma_1)^2}} \qquad \omega_2 = \frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\sqrt{1 + \frac{\hat{Q}_{2z}^2 - \hat{Q}_{1z}^2}{(\hat{Q}_{1z} + \sigma_2)^2}}. \tag{4.42}$$

From there, we define the constraints matrices

$$\mathbf{M}_1 = \begin{bmatrix} \omega_1^2 & 0 \\ 0 & -1 \end{bmatrix} \otimes \mathbf{I_N} \quad \mathbf{M}_2 = \begin{bmatrix} \omega_2^2 & 0 \\ 0 & -1 \end{bmatrix} \otimes \mathbf{I_M} \tag{4.43}$$

where $\otimes$ denotes the Kronecker product. To study the convergence of our new linear system, we resort to Theorem 4 from [92], that provides the following condition.

Convergence condition (Linear matrix inequality)
Consider the following linear matrix inequality with $\tau \in [0, 1]$:

$$0 \succeq \begin{bmatrix} \mathbf{A}^T\mathbf{P}\mathbf{A} - \tau^2\mathbf{P} & \mathbf{A}^T\mathbf{P}\mathbf{B} \\ \mathbf{B}^T\mathbf{P}\mathbf{A} & \mathbf{B}^T\mathbf{P}\mathbf{B} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 & \mathbf{D}_1 \\ \mathbf{C}_2 & \mathbf{D}_2 \end{bmatrix}^T \begin{bmatrix} \beta_1\mathbf{M}_1 & 0 \\ 0 & \beta_2\mathbf{M}_2 \end{bmatrix} \begin{bmatrix} \mathbf{C}_1 & \mathbf{D}_1 \\ \mathbf{C}_2 & \mathbf{D}_2 \end{bmatrix}. \tag{4.44}$$

If (4.44) is feasible for some $\mathbf{P} \succ 0$ and $\beta_1, \beta_2 \geqslant 0$, then for any initialization $\mathbf{h}^{(0)}$, $\mathbf{h}^{(t)}$ converges to a vector $\mathbf{h}^*$ and

$$\forall t, \qquad \|\mathbf{h}^{(t)} - \mathbf{h}^*\|_2 \leqslant \sqrt{\kappa(\mathbf{P})}\tau^t \|\mathbf{h}^{(0)} - \mathbf{h}^*\|_2 \tag{4.45}$$

where $\kappa(\mathbf{P})$ is the condition number of $\mathbf{P}$.

What does this inequality tell us? Matrices $\mathbf{A}, \mathbf{B}, \mathbf{C_1}, \mathbf{C_2}, \mathbf{M_1}, \mathbf{M_2}$ come from the definition of our system. Then for any $\tau \in [0, 1]$, we would like to find any $\mathbf{P} \succ 0$ and $\beta_1, \beta_2 \geqslant 0$, such that inequality (4.44) holds. If we do, then τ provides a convergence rate of our system. The convergence speed is also conditioned by a number $\kappa(\mathbf{P})$. Our job now will be to write the matrix to the right-hand side of (4.44), by assuming a given form of $\mathbf{P}$, and seeing if we can indeed find $\mathbf{P}$, τ, β_1 and β_2 to satisfy the inequality. To do this, we need to characterize some bounds on variance parameters.

4.4.5 Bounds on variance parameters

From the fixed point of the MLVAMP iterations, the definition of the averaging operators, and the form of the Jacobian of the proximal operator already used in (3.47), we proceed exactly like we did for VAMP in 3.5.2 to obtain the following relations on the $\hat{Q}$ parameters involving the Hessian matrices of f and g taken at the fixed point, denoted $\mathcal{H}_f$ and $\mathcal{H}_g$:

$$\frac{1}{\hat{Q}_{2x} + \hat{Q}_{1x}} = \frac{1}{N}\mathrm{Tr}\left[(\mathcal{H}_f + \hat{Q}_{1x}\mathrm{Id})^{-1}\right] = \frac{1}{N}\mathrm{Tr}\left[(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})^{-1}\right] \tag{4.46}$$

$$\frac{1}{\hat{Q}_{1z} + \hat{Q}_{2z}} = \frac{1}{N}\mathrm{Tr}\left[(\mathcal{H}_g + \hat{Q}_{1z}\mathrm{Id})^{-1}\right] = \frac{1}{M}\mathrm{Tr}\left[\mathbf{F}\mathbf{F}^T(\hat{Q}_{2z}\mathbf{F}\mathbf{F}^T + \hat{Q}_{2x}\mathrm{Id})^{-1}\right]. \tag{4.47}$$

From there, we obtain the following inequalities:

$$\sigma_1 \leqslant \hat{Q}_{2x} \leqslant \beta_1 \tag{4.48}$$

$$\sigma_2 \leqslant \hat{Q}_{2z} \leqslant \beta_2 \tag{4.49}$$

$$\hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F}) \leqslant \hat{Q}_{1x} \leqslant \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F}) \tag{4.50}$$

$$\frac{\hat{Q}_{2x}}{\lambda_{max}(\mathbf{F}\mathbf{F}^T)} \leqslant \hat{Q}_{1z} \leqslant \frac{\hat{Q}_{2x}}{\lambda_{min}(\mathbf{F}\mathbf{F}^T)}. \tag{4.51}$$

Note that $\lambda_{min}(\mathbf{F}\mathbf{F}^T)$ can be equal to 0, the right-hand side of the last inequality would then be uninformative. $\lambda_{max}(\mathbf{F}^T\mathbf{F}) = \lambda_{max}(\mathbf{F}\mathbf{F}^T)$ are strictly positive, since the spectrum is assumed to be non-trivial.

4.5 Smoothed problem and end of the proof

4.5.1 Convergence of the smoothed problem

In Chapter 3, we introduced a smooth problem by adding a ridge penalty to the convex regularization function, to enforce strong convexity. We follow the same idea here, but we add two ridge penalties: one that completes f, and one that completes g. We replace f and g by their twice-differentiable approximation [89, 10], but keep the same names for simplicity. The smoothed setting consists in solving the modified minimization problem, for $\lambda_2, \tilde{\lambda}_2 \geq 0$:

$$\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{\tilde{g}(\mathbf{F}\mathbf{x}, \mathbf{y}) + \tilde{f}(\mathbf{x})\right\} \tag{4.52}$$

where $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$ and $\tilde{g}(\mathbf{Fx}, \mathbf{y}) = g(\mathbf{Fx}, \mathbf{y}) + \frac{\tilde{\lambda}_2}{2}\|\mathbf{Fx}\|_2^2$. Note that $(\tilde{\sigma}_1, \tilde{\beta}_1)$ and $(\tilde{\sigma}_2, \tilde{\beta}_2)$ the strong convexity (and possibly infinite) smoothness constants of $\tilde{f}$ and $\tilde{g}$ verify:

$$\tilde{\sigma}_1 = \sigma_1 + \lambda_2 \qquad \tilde{\beta}_1 = \beta_1 + \lambda_2 \tag{4.53}$$

$$\tilde{\sigma}_2 = \sigma_2 + \tilde{\lambda}_2 \qquad \tilde{\beta}_2 = \beta_2 + \tilde{\lambda}_2. \tag{4.54}$$

For the smoothed problem (4.52), constants (σ_1, β_1) and (σ_2, β_2) are directly augmented by $\tilde{\lambda}_2, \lambda_2$, which allows us to control the scaling of $\hat{Q}_{2x}$ and $\hat{Q}_{2z}$ through (4.48) and (4.49). The rest of the convergence proof is then based on successive application of Schur's lemma [76] on the linear matrix inequality (4.44) by translating the convergence conditions into simpler inequalities; which can be verified by choosing the appropriate $\tilde{\lambda}_2, \lambda_2, \beta_1, \beta_2$. The computation needs to be meticulously done and is detailed in appendix D.4. Defining vector $\mathbf{h}^{(t)} = \left[\mathbf{h}_{2z}^{(t)}, \mathbf{h}_{1x}^{(t)}\right]^T$, our result formally reads:

$$\begin{aligned}
&\forall\ \tilde{\lambda}_2 > 0,\ \exists\ \Lambda_2 \ \text{ such that } \ \forall \lambda_2 \geqslant \Lambda_2 : \\
&\exists\ c \in]0, \lambda_2[\ \text{ such that } \ \|\mathbf{h}^{(t)} - \mathbf{h}^*\|_2^2 \leqslant \left(\frac{c}{\lambda_2}\right)^t \|\mathbf{h}^{(0)} - \mathbf{h}^*\|_2^2 \\
&\text{which implies } \quad \lim_{t\to\infty} \|\mathbf{h}^{(t)} - \mathbf{h}^*\|_2^2 = 0
\end{aligned} \tag{4.55}$$

where $\mathbf{h}^*$ is the fixed point value of $\mathbf{h}^{(t)}$ for Oracle MLVAMP on the smoothed problem. Basically, we show that for any ridge penalty added to g, Oracle MLVAMP converges if the ridge penalty added to f is strong enough. Note that proving convergence of gradient-based descent methods for sufficiently strongly-convex functions is a coherent result from an optimization point of view. This is corroborated by the symbolic convergence rates derived for ADMM in [117], where a sufficiently strongly convex function is also considered. We have now proven point iii, but only for the smoothed problem in a given regime of added ridge penalty.

4.5.2 Analytic continuation

We want to use an analytic continuation on scalar quantities that are obtained by the state evolution equations, which include the overlaps used to characterize our reconstruction error. First, note that the optimality condition of the smoothed problem (4.52) reads

$$\partial f(\hat{\mathbf{x}}) + \mathbf{F}^T \partial g(\mathbf{F}\hat{\mathbf{x}}) + (\lambda_2 \mathrm{Id} + \tilde{\lambda}_2 \mathbf{F}^{\mathrm{T}}\mathbf{F})\hat{\mathbf{x}} = 0 \tag{4.56}$$

and defines an analytic function of $(\lambda_2, \tilde{\lambda}_2)$ for the coordinates of the solution $\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2)$, and thus for $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$ as well, using the analytic inverse function theorem from [83]. We also know that the proximal of a convex (and differentiable) function with an addition ridge penalty reads:

$$\mathrm{Prox}_{\gamma(f + \frac{\lambda_2}{2}\|.\|_2^2)}(z) = ((1 + \gamma\lambda_2)\mathrm{Id} + \gamma f')^{-1}(z) \tag{4.57}$$

which is an analytic function in λ_2. Therefore, all equations defining the scalar quantities in the state evolution equations (4.27) at finite N are analytic in $(\lambda_2, \tilde{\lambda}_2)$, which implicitly defines an analytic function for any scalar combination of those quantities [83]. Moreover, the convergence of Oracle MLVAMP on the smoothed problem holds for an open subset of $(\lambda_2, \tilde{\lambda}_2)$; we can therefore use an analytic continuation theorem [83] to extend it to all non-negative values of $(\lambda_2, \tilde{\lambda}_2)$. By choosing $(\lambda_2, \tilde{\lambda}_2) = (0, 0)$; we show that the expressions of all scalar quantities defined by state evolution equations at finite N still hold for the original problem (4.2). To take the asymptotic limit, we must again proceed carefully. We consider the asymptotic MSE as a pointwise limit of a sequence of loss functions for N finite, which are themselves analytic. However, if the asymptotic loss is not analytic in $(\lambda_2, \tilde{\lambda}_2)$, then it could have a different expression

outside MLVAMP's convergence domain, as underlined in the last paragraph of 3.6. To properly conclude the proof, we need to assume analyticity of the MSE in $(\lambda_2, \tilde{\lambda}_2)$ in the asymptotic limit. Proving this reasonable assumption is left for later work. Under this condition, we have safely found that replica and SE equations provide the exact asymptotic expression for the reconstruction performance of estimator $\hat{\mathbf{x}}$.

4.6 Numerical experiments

4.6.1 Matrix parameters and singular values

We perform a few experiments to compare simulation with the theoretical expressions predicted by the state evolution and replica result. The experimental points were obtained using the convex optimization tools of [126], with a data matrix of dimension $N = 200, M = \alpha N$, for $\alpha \in [0.1, 3]$. Each point is an average over 100 realizations. We assume that the ground-truth x_0 is pulled from a Gauss-Bernoulli law of the form:

$$p_{x_0}(x_0) = (1-\rho)\delta(x_0) + \rho \frac{1}{\sqrt{2\pi}} \exp\left(-x_0^2/2\right). \tag{4.58}$$

Our plots correspond to $\rho = 1$. The two types of matrices we use are Gaussian i.i.d., the eigenvalues of $\mathbf{F}^T\mathbf{F}$ are then sampled from the Marchenko-Pastur distribution (B.23); and rotationally invariant matrices with singular values independently sampled from the uniform distribution $\mathcal{U}([(1-\alpha)^2, (1+\alpha)^2])$; which yields p_λ as in (3.56). The theoretical prediction is obtained by iterating the state evolution equations (4.27), until they converge to their fixed point.

4.6.2 Regularization: elastic net

We use elastic net regularization $f = \lambda_1\|\cdot\|_1 + \frac{\lambda_2}{2}\|\cdot\|_2^2$, and the proximal expression is given in (C.37). To run state evolution equations, we need to compute a few expectations over the regularization proximal, that we provide here. Writing $X = \frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}$, where $\xi_{1x} \sim \mathcal{N}(0,1)$, a little calculus shows that:

$$\begin{aligned}\mathbb{E}[\mathrm{Prox}^2_{f/\hat{Q}_{1x}}(X)] = \left(\frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\right)^2 \Bigg[(1-\rho)\left(\frac{\lambda_1^2 + (\hat{Q}_{1x})^2\tau}{(\hat{Q}_{1x})^2}\,\mathrm{erfc}\left(\frac{\lambda_1}{\hat{Q}_{1x}\sqrt{2\tau}}\right) - \frac{\lambda_1\sqrt{2\tau}\exp(-\frac{\lambda_1^2}{2(\hat{Q}_{1x})^2\tau})}{\hat{Q}_{1x}\sqrt{\pi}}\right) \\ + \rho\left(\frac{\lambda_1^2 + (\hat{Q}_{1x})^2(\tau+\tau_0)}{(\hat{Q}_{1x})^2}\,\mathrm{erfc}\left(\frac{\lambda_1}{\hat{Q}_{1x}\sqrt{2(\tau+\tau_0)}}\right) - \frac{\lambda_1\sqrt{2(\tau+\tau_0)}\exp(-\frac{\lambda_1^2}{2(\hat{Q}_{1x})^2(\tau+\tau_0)})}{\hat{Q}_{1x}\sqrt{\pi}}\right)\Bigg],\end{aligned} \tag{4.59}$$

$$\mathbb{E}[\mathrm{Prox}'_{f/\hat{Q}_{1x}}(X)] = \frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\left[(1-\rho)\,\mathrm{erfc}\left(\frac{\lambda_1}{\hat{Q}_{1x}\sqrt{2\tau}}\right) + \rho\,\mathrm{erfc}\left(\frac{\lambda_1}{\hat{Q}_{1x}\sqrt{2(\tau+\tau_0)}}\right)\right], \tag{4.60}$$

$$\mathbb{E}[x_0\mathrm{Prox}_{f/\hat{Q}_{1x}}(X)] = \frac{\rho\sqrt{\tau_0}}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\,\mathrm{erfc}\left(\frac{\lambda_1}{\hat{Q}_{1x}\sqrt{2(\tau+\tau_0)}}\right), \tag{4.61}$$

where we write $\tau_0 = (\hat{m}_{1x}/\hat{Q}_{1x})^2$ and $\tau = \hat{\chi}_{1x}/\hat{Q}_{1x}^2$.

4.6.3 Loss functions

We provide the proximal of the loss function for several cases, which are necessary to run state evolution equations. The involved expressions cannot always be simplified, and we had to perform a numerical integration. In the present model, if the teacher y is chosen as a sign, one-dimensional integrals can be reached, leading to stable and reasonably fast implementation

(a few minutes to generate a curve comparable to those of Figure 4.2 for the non-linear models, the ridge regression being very fast).

Square loss The square loss is defined as:

$$g(p, y) = \frac{1}{2}(p - y)^2, \tag{4.62}$$

its proximal and partial derivative then read:

$$\mathrm{Prox}_{g/\gamma}(p) = \frac{\gamma}{1+\gamma}p + \frac{1}{1+\gamma}y \tag{4.63}$$

$$\frac{\partial}{\partial p}\mathrm{Prox}_{g/\gamma}(p) = \frac{\gamma}{1+\gamma}. \tag{4.64}$$

Using this form with a plain ridge penalty (elastic net with $\lambda_1 = 0$) leads to great simplification in the state evolution equations, and recovers the classical expressions obtained for ridge regression in papers such as [71, 61].

Hinge loss and SVMs The hinge loss is used for "maximum-margin" classification, notably for support vector machines (SVMs). It reads:

$$g(p, y) = \max(0, 1 - py). \tag{4.65}$$

Assuming $y \in \{-1, +1\}$, its proximal and partial derivative then read:

$$\mathrm{Prox}_{g/\gamma}(p) = \begin{cases} p + \frac{y}{\gamma} & \text{if} \quad \gamma(1-yp) \geqslant 1 \\ y & \text{if} \quad 0 \leqslant \gamma(1-yp) \leqslant 1 \\ p & \text{if} \quad \gamma(1-yp) \leqslant 0 \end{cases} \tag{4.66}$$

$$\frac{\partial}{\partial p}\mathrm{Prox}_{g/\gamma}(p) = \begin{cases} 1 & \text{if} \quad \gamma(1-yp) \geqslant 1 \\ 0 & \text{if} \quad 0 \leqslant \gamma(1-yp) \leqslant 1 \\ 1 & \text{if} \quad \gamma(1-yp) \leqslant 0. \end{cases} \tag{4.67}$$

Logistic loss The logistic loss reads:

$$g(p, y) = \log(1 + \exp(-py)) \tag{4.68}$$

Its proximal (at point p) is the solution to the fixed point problem:

$$x = p + \frac{y}{\gamma(1 + \exp(py))}, \tag{4.69}$$

and its derivative, given that the logistic loss is twice differentiable, reads

$$\frac{\partial}{\partial p}\mathrm{Prox}_{g/\gamma}(p) = \left(1 + \frac{1}{\gamma}\frac{\partial^2}{\partial p^2} g(\mathrm{Prox}_{g/\gamma}(p))\right)^{-1} \tag{4.70}$$

$$= \left(1 + \frac{1}{\gamma}\frac{1}{(2 + 2\cosh(\mathrm{Prox}_{g/\gamma}(p))}\right)^{-1}. \tag{4.71}$$

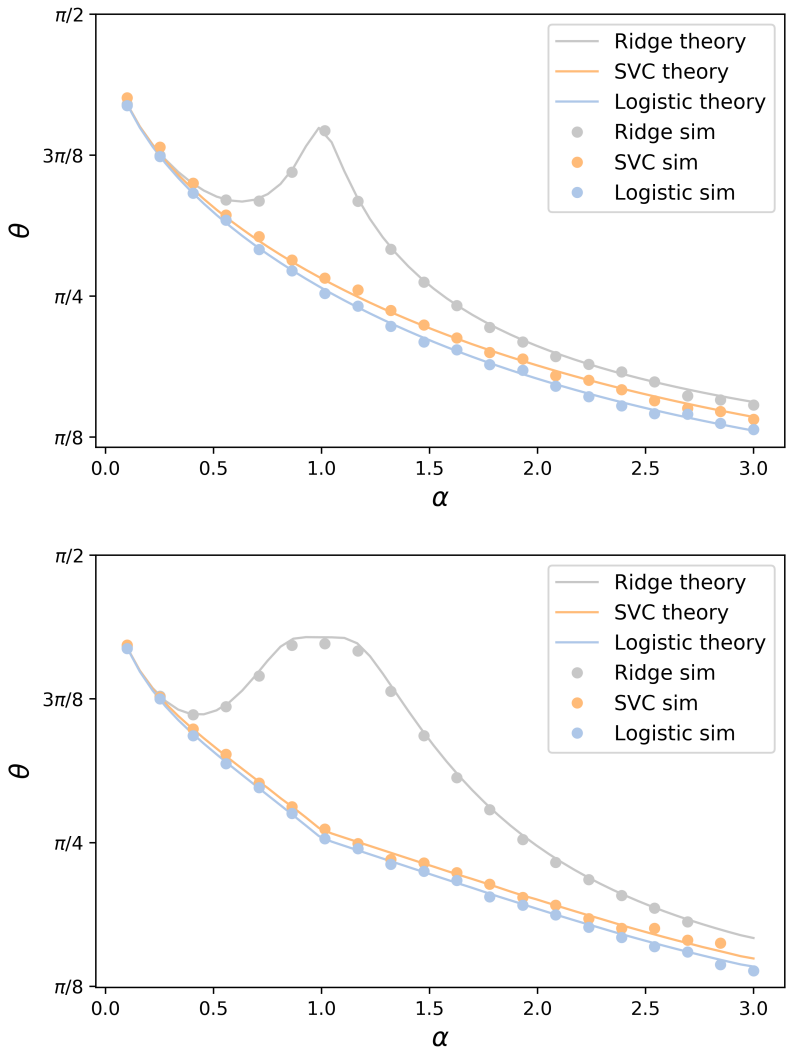


FIGURE 4.2: Comparison between simulation and theory (from the state evolution equations) predictions of scalar quantities characterizing reconstruction performance. We consider a binary classification problem with data generated as $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x}_0 + \omega_0)$ with matrix $\mathbf{F}$ being in the top figure Gaussian i.i.d., and in the bottom a rotationally invariant matrix with a squared uniform density of eigenvalues. We plot the angle between the estimator and the ground-truth vector $\theta = \arccos(m_x^*/(\sqrt{\rho_x q_x^*}))$ as a function of the aspect ratio $\alpha = M/N$ taking ℓ_2 penalty with $\lambda_2 = 10^{-3}$. The losses correspond to a square loss (ridge regression), Support Vector Machine through hinge loss, and a logistic regression. The theoretical prediction (full line) is compared with numerical experiments (points) conducted using standard convex optimization solvers from [126].

Summary of Chapter 4 In this chapter, we proved the replica formula that characterizes reconstruction performance for a teacher-student scenario in a generalized linear model, when the data matrix is rotationally invariant, and under an analyticity assumption for the asymptotic mean squared error. This performance is determined through scalar parameters that describe minimizer (4.2), which is the solution of an unconstrained convex problem. Extremizing the replica free energy yields the same equations as the fixed point of the state evolution equations of multi-layer VAMP, in its two-layer version. We study an Oracle version of MLVAMP, that can be recast as a linear dynamical system, and use a linear matrix inequality that mediates a convergence condition. By adding ridge penalties to the regularization and loss functions, we

can find a regime where Oracle MLVAMP is convergent. Besides, we show that its fixed point corresponds to the minimizer we are interested in, indicating that it is indeed characterized by the replica formula. Performing an analytic continuation on the added regularization parameters extends the validity of the expressions of all scalar quantities coming from the state evolution equations; which shows the replica formula for the original minimization problem. We compare theory to simulations for matrices with size of a few hundreds, with a sign teacher, two different types of matrices, an ℓ_2 penalty and several losses. We find excellent agreement, despite the small size of the matrices and the asymptotic nature of the state evolution/replica analytic result.The proofs from the two last chapters strongly rely on message passing algorithms state evolution equations. A natural question would be to probe the existence of a similar set of equations (at least empirical) for different algorithms, such as expectation propagation schemes, which would also allow a physical or optimization approach to characterize algorithmic performance.

Chapter 5

Rademacher complexity and free energy: a link between the replica and statistical theories of learning

Until now, we have mostly focused on inference problems formulated in statistical physics terms, and pointed out links with convex optimization, message passing algorithms, and random matrix theory tools. In this chapter, we take a Diagon Alley[1] and begin by shortly introducing the well-known statistical learning problem of generalization, and later underline how some angles can be rephrased through the prism of statistical physics. This chapter is adapted from [1].

5.1 Convergence bounds on the generalization gap

5.1.1 A friendly introduction to the generalization problem

We consider binary classifications functions on a space $\mathcal{X}$ into $\{-1, +1\}$. $\mathcal{X}$ is an environment, such as a set of images, on which we define a concept through a teacher function f_T; such as the absence of presence of a bird on the picture. The goal is to find a function that approximates the concept, which will be pulled from a *hypothesis class* $\mathcal{F}$. The latter could be for instance a neural network or a linear function, with respective weights or parameters $\mathbf{w}$. We want to compare f_T with functions from $\mathcal{F}$ on the whole space $\mathcal{X}$, but instead we use a *test set* of M examples $\{y^{(\mu)}, \mathbf{x}^{(\mu)}\}_{\mu=1,..,M}$ where $y^{(\mu)} = f_T(\mathbf{x}^{(\mu)})$; and $\mathbf{x}^{(\mu)}$ is assumed to be drawn from a distribution p_x. We choose a loss function $\mathcal{L}$, such as the square loss $\mathcal{L}(y_1, y_2) = (y_1 - y_2)^2$, or in the following the indicator $\mathcal{L}(y_1, y_2) = \mathbb{1}(y_1, y_2)$. A general definition to measure the performance of $f_{\mathbf{w}} \in \mathcal{F}$ is the *empirical risk*

$$\mathcal{R}^M_{\text{empirical}}(f_{\mathbf{w}}) = \frac{1}{M} \sum_{\mu=1}^{m} \mathcal{L}\left(y^{(\mu)}, f_{\mathbf{w}}(\mathbf{x}^{(\mu)})\right). \tag{5.1}$$

What we actually hope to minimize is the *population risk*, defined as

$$\mathcal{R}_{\text{population}}(f_{\mathbf{w}}) = \mathbb{E}_{y,\mathbf{x}}\left[\mathcal{L}(y, f_{\mathbf{w}}(\mathbf{x}))\right]. \tag{5.2}$$

We typically pick a function $f_{\mathbf{w}}$ after a learning phase, where we try to obtain the value of parameters $\mathbf{w}$ that guarantees a good match with the teacher function on a given set of examples. Taking the same set as above turns it into a *training set*. This learning process aims at *empirical risk minimization*. The *generalization* question amounts to knowing if the chosen function $f_{\mathbf{w}}$ performs well on the complete set $\mathcal{X}$, knowing that it is a good approximation of the concept on the training set. However, the discrepancy between empirical and population risks could be arbitrarily large. Our goal is thus to bound the difference between them, that we

[1] see *Harry Potter and the Philosopher's stone*, J.K. Rowling.

call *generalization gap*.

For simplicity, let us call $\epsilon_{\mathbf{w}}$ the population risk, and $\nu_{\mathbf{w}}$ the probability of f_T and $f_{\mathbf{w}}$ being equal on a random element from the training set. Luckily, the law of large numbers states that for a randomly chosen set of parameters $\mathbf{w}$, $\nu_{\mathbf{w}}^M$ converges to $\epsilon_{\mathbf{w}}$ as $M \to \infty$. The convergence rate for one function $f_{\mathbf{w}}$ is then provided by the *Hoeffding inequality*, where we take δ a tolerance parameter:

$$\text{Prob}\Big(|\nu_{\mathbf{w}}^M - \epsilon_{\mathbf{w}}| > \delta\Big) \leq 2e^{-\delta^2 M}. \tag{5.3}$$

The convergence of $\nu_{\mathbf{w}}^M$ to $\epsilon_{\mathbf{w}}$ is called *convergence of frequencies to probabilities*. (5.3) notably shows that the convergence rate is of order $1/\sqrt{M}$. However, this inequality holds for a given function $f_{\mathbf{w}}$, but we would like to have a uniform convergence result, for all functions in the hypothesis class. Indeed, we may pick through the learning process a function $f_{\mathbf{w}}$ such that $\nu_{\mathbf{w}}^M$ converges very slowly to $\epsilon_{\mathbf{w}}$, and the generalization gap would be large. Therefore we focus on the quantity

$$\text{Prob}\left(\sup_{f_{\mathbf{w}} \in \mathcal{F}} |\nu_{\mathbf{w}}^M - \epsilon_{\mathbf{w}}| > \delta\right).$$

This probability accounts for the worst-case scenario since it takes the supremum on all of the hypothesis class.

5.1.2 The Vapnik-Chervonenkis theorem on uniform convergence

If $\mathcal{F}$ is finite, we can invoke Hoeffding's inequality on each function from the hypothesis class, thus reaching

$$\text{Prob}\Big(\sup_{f_{\mathbf{w}} \in \mathcal{F}} |\nu_{\mathbf{w}}^M - \epsilon_{\mathbf{w}}| > \delta\Big) \leq 2|\mathcal{F}|e^{-\delta^2 M}. \tag{5.4}$$

The key for the case where $\mathcal{F}$ is infinite is the Vapnik-Chervonenkis theorem, that states:

$$\text{Prob}\Big(\sup_{f_{\mathbf{w}} \in \mathcal{F}} |\nu_{\mathbf{w}}^M - \epsilon_{\mathbf{w}}| > \delta\Big) \leq C_1 \Delta_{\mathcal{F}}(2M) e^{-\delta^2 M}. \tag{5.5}$$

where C_1 is a constant, and $\Delta_{\mathcal{F}}(2M)$ represents an effective number of functions in $\mathcal{F}$, mediating the number of concepts that can be expressed on the training set by the hypothesis class. To reach uniform convergence, we need $\Delta_{\mathcal{F}}(2M)e^{-\delta^2 M}$ to go to zero as $M \to \infty$, therefore $\Delta_{\mathcal{F}}$'s growth should not be too fast. In particular, a polynomial growth would be perfect. Taking a random set $\{\mathbf{x}^{(\mu)}\} \in \mathcal{X}^M$, $\Delta_{\mathcal{F}}(M)$ is the *growth function* of the hypothesis class, and is defined by the maximum number of different classifications which can be induced by its functions. If $\mathcal{F}$ is finite, it is clear that $\Delta_{\mathcal{F}}(M) \leq |\mathcal{F}|$, since the hypothesis class cannot induce more classifications than its number of functions. Besides, $\Delta_{\mathcal{F}}(M) \leq 2^M$ which is the total number of classifications. Thankfully, Vapnik and Chervonenkis one side [33, 171], Sauer on the other [140], showed that for every class of functions $\mathcal{F}$, there exists a unique integer $d_{\text{VC}}(\mathcal{F})$, called the *VC dimension*, such that for $M \leq d_{\text{VC}}$, all 2^M classifications can be induced by $\mathcal{F}$. For $M > d_{\text{VC}}$, $\Delta_{\mathcal{F}}(M)$ is bounded by a polynomial function. More precisely:

$$\Delta_{\mathcal{F}}(M) = 2^M \qquad \text{if } M \leq d_{\text{VC}} \tag{5.6a}$$

$$\Delta_{\mathcal{F}}(M) \leq \left(\frac{eM}{d_{\text{VC}}}\right)^{d_{\text{VC}}} \qquad \text{if } M > d_{\text{VC}}. \tag{5.6b}$$

Another way of defining the VC dimension is the size of the set that can be fully shattered by the hypothesis class $\mathcal{F}$. In other words, if we can find a set of k points from $\mathcal{X}$ such that any

classification on these examples can be achieved by a function from the hypothesis class, but cannot find any such set of $k+1$ examples, then the VC dimension is k. For instance, if $\mathcal{X} = \mathbb{R}^2$, and $\mathcal{F}$ is the space of linear classifiers, then the associated VC dimension is 3, as explained in Fig. 5.1. The VC theorem (5.5) thus provides a uniform convergence result, which informally

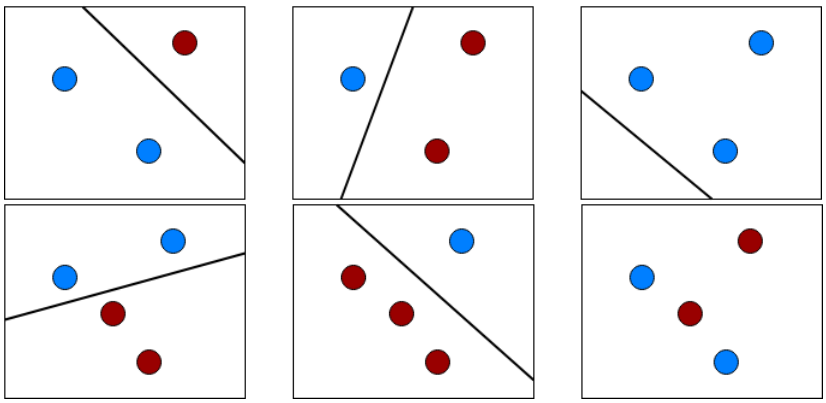


FIGURE 5.1: First line: a set of three points that can be shattered by a linear classifier, no matter what the chosen labels are (red or blue), as shown in the three examples. Second line: a set of four points that cannot always be shattered by a linear classifier, as shown in the last picture. Any set of 4 points can be attributed a blue/red classification that cannot be shattered; hence the VC dimension of linear classifiers on $\mathbb{R}^2$ is 3.

reads

$$\mathcal{R}_{\text{population}}(f_{\mathbf{w}}) - \mathcal{R}^{M}_{\text{empirical}}(f_{\mathbf{w}}) = \mathcal{O}\left(\sqrt{\frac{d_{\text{VC}}(\mathcal{F})}{M}}\right). \tag{5.7}$$

However, this approach clearly does not give tight enough bounds: not only is it a worst-case analysis, but it also does not depend on the data distribution. It holds whatever the distribution is, and fails to exploit information about the problem's specificity. Moreover, it only applies to binary classification.

5.1.3 The Rademacher complexity

Rademacher bound on uniform convergence

Another quantity provides a uniform convergence bound on the generalization gap, and takes into account the data distribution: the Rademacher complexity [17]. Let $f_{\mathbf{w}}$ be any function in the hypothesis class $\mathcal{F}$ on $\mathbb{R}^N$, and let $\epsilon \in \{\pm 1\}^M$ be drawn uniformly at random. The *empirical Rademacher complexity* is defined as

$$\hat{\mathfrak{R}}_M(\mathcal{F}, \mathbf{X}) \equiv \mathbb{E}_\epsilon \left[\sup_{f_{\mathbf{w}} \in \mathcal{F}} \frac{1}{M} \sum_{\mu=1}^{M} \epsilon_\mu f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right) \right], \tag{5.8}$$

and depends on the sample examples $\mathbf{X} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(M)}\} \in \mathbb{R}^{N \times M}$. The *Rademacher complexity* is defined as the population average

$$\mathfrak{R}_M(\mathcal{F}) \equiv \mathbb{E}_{\mathbf{X}} \left[\hat{\mathfrak{R}}_M(\mathcal{F}, \mathbf{X}) \right]. \tag{5.9}$$

We keep our focus on binary classification and consider the corresponding indicator loss function $\mathcal{L}(y_1, y_2) = \mathbb{1}(y_1 - y_2)$ that counts the number of misclassified samples. We will be interested

in a hypothesis class $\mathcal{F} = \left\{f_{\mathbf{w}} : \mathbb{R}^N \longrightarrow \{\pm 1\}\right\}$. We define the training error $\epsilon^M_{\text{train}}(.)$ and generalization error $\epsilon_{\text{gen}}(.)$ for any function $f_{\mathbf{w}} \in \mathcal{F}$ by

$$\epsilon^M_{\text{train}}(f_{\mathbf{w}}) \equiv \frac{1}{M}\sum_{\mu=1}^{M} \mathbb{1}\left(y^{(\mu)} \neq f_{\mathbf{w}}(\mathbf{x}^{(\mu)})\right) \tag{5.10}$$

$$\epsilon_{\text{gen}}(f_{\mathbf{w}}) \equiv \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathbb{1}(y \neq f_{\mathbf{w}}(\mathbf{x}))\right]. \tag{5.11}$$

The Rademacher complexity provides a uniform convergence bound on binary classification (see e.g. [17, 146, 113]), that can be stated as follows: Fix a distribution p_x and let $\delta > 0$. Let $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots \mathbf{x}^{(M)}\} \in \mathbb{R}^{N\times M}$ a set of examples identically and independently drawn from p_x. Then with probability at least $1-\delta$ (over the draw of $\mathbf{X}$),

$$\forall f_{\mathbf{w}} \in \mathcal{F}, \quad \epsilon_{\text{gen}}(f_{\mathbf{w}}) - \epsilon^M_{\text{train}}(f_{\mathbf{w}}) \leq \mathfrak{R}_M(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{M}}. \tag{5.12}$$

The Rademacher complexity becomes a uniform bound of the generalization gap: in our favorite asymptotic limit when $M, N \to \infty$ the $\delta-$dependent term goes to zero, and only the first term on the right-hand side remains finite.

Recovering the VC uniform convergence bound

This inequality can be used to recover the classical VC result (5.7). To do so, we invoke Massart's lemma [102], which states that for a finite subset $\mathcal{A} \subset \mathbb{R}^M$ and $\epsilon \in \{\pm 1\}^M$ uniformly drawn:

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{\mu=1}^{M} \epsilon_\mu a_\mu\right] \leq \sup_{a\in\mathcal{A}}\left[(\textstyle\sum_{\mu=1}^M a_\mu^2)^{1/2}\right]\sqrt{2\log|A|}. \tag{5.13}$$

Consider a particular sample set $\mathbf{X}$, with M elements. Taking any hypothesis class $\mathcal{F}$, it can at most induce $\Delta_{\mathcal{F}}(M)$ different classifications on those M points, i.e. a finite number, where $\Delta_{\mathcal{F}}$ is the growth function. We pick $\Delta_{\mathcal{F}}(M)$ functions from the hypothesis class such that they cover all possible classifications described by the class, and call this ensemble of functions $\mathcal{F}_X$. We then define the finite set $\mathcal{A}_X = \{\left(f(\mathbf{x}^{(1)}), ..., f(\mathbf{x}^{(M)})\right) | f \in \mathcal{F}_X\}$. Clearly $\sup_{a\in\mathcal{A}_X}\left[(\sum_{\mu=1}^M a_\mu^2)^{1/2}\right] = \sqrt{M}$. Writing the Rademacher definition yields:

$$\mathfrak{R}_M(\mathcal{F}) \equiv \mathbb{E}_{\mathbf{X},\epsilon}\left[\mathfrak{R}_M(\mathcal{F},\mathbf{X})\right] \leq \mathbb{E}_{\mathbf{X}}\left[\frac{1}{M}\sup_{a\in\mathcal{A}_X}\frac{1}{M}\sum_{\mu=1}^{M}\epsilon_\mu a_\mu\right] \tag{5.14}$$

$$\leq \mathbb{E}_{\mathbf{X}}\left[\sqrt{\frac{2\log|\mathcal{A}_{\mathbf{X}}|}{M}}\right] \leq \sqrt{\frac{2\log\Delta_{\mathcal{F}}(M)}{M}} \tag{5.15}$$

$$\mathfrak{R}_M(\mathcal{F}) \leq C\sqrt{\frac{d_{\text{VC}}(\mathcal{F})}{M}} \tag{5.16}$$

where C is a constant, according to (5.6). (5.12) thus agrees with the bound given in (5.7). In the rest of the chapter, we will look at a few different hypothesis classes, and see how the corresponding Rademacher complexities can be computed or understood through statistical physics.

5.2 Rademacher complexities on some hypothesis classes for i.i.d. data

We start by considering that each vector from data points $\mathbf{X} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(M)}\} \in \mathbb{R}^{N\times M}$ has been generated identically and independently from a factorized distribution, such that $\forall \mu \in [\![1; M]\!], p_x\left(\mathbf{x}^{(\mu)}\right) = \prod_{i=1}^{N} p_x(x_i^{(\mu)})$. A simple example would be to take matrix $\mathbf{X}$ Gaussian i.i.d. For now, we focus on i.i.d. data distribution, but sec. 5.4.5 presents a generalization to rotationally invariant matrices $\mathbf{X}$. Our approach is to work on *typical case* problems, instead of the worst-case analysis achieved through statistical bounds [145, 178, 118, 49, 184]. Real-world applications imply complicated structures of data, but we still hope to gain insight on them by understanding simple settings and computing closely or exactly the associated Rademacher complexities. We focus on the usual asymptotic limit where $M, N \to \infty$, with $\alpha = \frac{M}{N}$ of order 1.

5.2.1 Linear model

To start with a simple example, we tackle the computation of the Rademacher complexity for a simple function class containing all linear models with weights $\mathbf{w} \in \mathbb{R}^N$ lying on the sphere $S_N(\Gamma)$ of radius $\Gamma > 0$:

$$\mathcal{F}_{\text{linear}} = \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^N \longrightarrow \mathbb{R} \\ \mathbf{x} \longrightarrow \frac{1}{\sqrt{N}} \mathbf{w}^T \mathbf{x} \end{cases} , \mathbf{w} \in S_N(\Gamma) \right\} . \tag{5.17}$$

From eq. (5.9),

$$\mathfrak{R}_M\left(\mathcal{F}_{\text{linear}}\right) = \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{M\sqrt{N}} \sup_{\mathbf{w}\in S_N(\Gamma)} \left[\mathbf{y}^T \mathbf{X}^T \mathbf{w} \right] \right] \tag{5.18}$$

where $\mathbf{y}$ is replacing ϵ. Computing the empirical Rademacher complexity amounts to finding the vector $\mathbf{w}^\star$ that maximizes the scalar product between $\mathbf{y}$ and $\mathbf{X}^T\mathbf{w}$, which is $\mathbf{w}^\star = \frac{\mathbf{X}^T\mathbf{y}}{\|\mathbf{X}^T\mathbf{y}\|_2}\sqrt{N}\Gamma$. The empirical Rademacher complexity thus reads

$$\mathfrak{R}_m\left(\mathcal{F}_{\text{linear}}\right) = \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{\Gamma}{M} \|\mathbf{X}\mathbf{y}\|_2 \right] . \tag{5.19}$$

Since $\mathbf{X}$ has i.i.d. entries, we can apply the central limit theorem, which states:

$$\forall i \in [\![1, N]\!],\ (\mathbf{X}\mathbf{y})_i = \sum_{\mu=1}^{M} x_i^{(\mu)} y_\mu \sim \mathcal{N}(0, M). \tag{5.20}$$

hence

$$\mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\|\mathbf{X}\mathbf{y}\|_2^2 \right] = \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\sum_{i=1}^{N} (x_i^{(\mu)} y_\mu)^2 \right] = NM \tag{5.21}$$

and $\mathbb{E}_{\mathbf{y},\mathbf{X}}[\|\mathbf{X}\mathbf{y}\|_2] = \sqrt{NM}$, which finally gives

$$\mathfrak{R}_M\left(\mathcal{F}_{\text{linear}}\right) = \frac{\Gamma}{\sqrt{\alpha}} , \tag{5.22}$$

where $\alpha = \frac{M}{N}$. This first result for the simple linear function hypothesis class allows to grasp the behavior of the Rademacher complexity in the asymptotic limit. At fixed input dimension N, it decreases with the number of samples as $1/\sqrt{\alpha}$. The generalization gap thus goes to zero as α goes to infinity. Note that increasing the radius of the weights expands the function complexity

and might help for fitting the data-set, but unfortunately leads to a looser generalization bound, which illustrates the bias variance trade-off [60]. The Rademacher complexity scales as $\alpha^{-1/2}$, therefore it remains finite in the asymptotic limit, while the term $\sqrt{\log(1/\delta)/M}$ in (5.12) goes to zero as $M \to \infty$.

5.2.2 Perceptron model

We now turn to a different hypothesis class: the perceptron denoted $\mathcal{F}_{\text{sign}}$. This class contains linear classifiers which output binary variables, and will fit much better labels in the binary classification task. The class writes

$$\mathcal{F}_{\text{sign}} = \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^N \longrightarrow \{\pm 1\} \\ \mathbf{x} \longrightarrow \text{sign}\left(\frac{1}{\sqrt{N}} \mathbf{w}^T \mathbf{x}\right) \end{cases} , \mathbf{w} \in \mathbb{R}^N \right\} . \tag{5.23}$$

We assume that $\mathbf{X} \in \mathbb{R}^{N \times M}$ is i.i.d. with all elements of $\mathbf{x}^{(\mu)}$ sampled from $\mathcal{N}(0,1)$. We want to show that the Rademacher complexity for this class asymptotically scales as $1/\sqrt{\alpha}$ when α grows large, i.e. the same behavior as the Rademacher complexity for $\mathcal{F}_{\text{linear}}$. To do this, we will point out an upper and lower bound of the Rademacher complexity that scale as such.

Upper bound

For a linear classifier with binary ouputs such as the perceptron, the VC dimension is easy to compute and $d_{\text{VC}} = N$. Hence we know from Massart's theorem [102] that $\mathfrak{R}_M(\mathcal{F}_{\text{sign}})$ is bounded by a term of order $\sqrt{d_{\text{VC}}(\mathcal{F}_{\text{sign}})/M} = 1/\sqrt{\alpha}$.

Lower bound

Let us consider the following estimator, known as the Hebb rule [72]: $\mathbf{w}^\star = \frac{1}{\sqrt{N}} \sum_{\nu=1}^{M} y^{(\nu)} \mathbf{x}^{(\nu)}$. It gives the following function

$$f_{\mathbf{w}^\star}\left(\mathbf{x}^{(\mu)}\right) = \text{sign}\left(\frac{1}{\sqrt{N}} \mathbf{w}^{\star T} \mathbf{x}^{(\mu)}\right) = \text{sign}\left(\left(\frac{1}{N} \sum_{\nu=1}^{M} y^{(\nu)} \mathbf{x}^{(\nu)}\right)^T \mathbf{x}^{(\mu)}\right) .$$

By definition, the Rademacher complexity reads

$$\begin{aligned} \mathfrak{R}_M(\mathcal{F}_{\text{sign}}) &\equiv \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\sup_{\mathbf{w}} \frac{1}{M} \sum_{\mu=1}^{M} y^{(\mu)} f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right) \right] \\ &\geq \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{M} \sum_{\mu=1}^{M} y^{(\mu)} f_{\mathbf{w}^\star}\left(\mathbf{x}^{(\mu)}\right) \right] \\ &= \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{M} \sum_{\mu=1}^{M} \text{sign}\left(y^{(\mu)} \frac{1}{N} \left(\sum_{\nu=1}^{M} y^{(\nu)} \mathbf{x}^{(\nu)} \right)^T \mathbf{x}^{(\mu)} \right) \right] \\ &= \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{M} \sum_{\mu=1}^{M} \text{sign}\left(1 + \frac{1}{N} \sum_{\nu \neq \mu}^{M} y^{(\mu)} y^{(\nu)} \mathbf{x}^{(\nu)T} \mathbf{x}^{(\mu)} \right) \right] . \end{aligned}$$

As $\mathbf{x}^{(\mu)} \sim \mathcal{N}(0,1)^N$ and $y^{(\mu)} \sim \frac{1}{2}\delta(y^{(\mu)}+1) + \frac{1}{2}\delta(y^{(\mu)}-1)$, $\mathbf{z}^{(\mu)} \equiv y^{(\mu)}\mathbf{x}^{(\mu)} \sim \mathcal{N}(0,1)^N$. Let us define the random variable

$$\theta_\mu \equiv \frac{1}{N}\sum_{\nu\neq\mu}^{M} y^{(\mu)}y^{(\nu)}\mathbf{x}^{(\nu)T}\mathbf{x}^{(\mu)} = \frac{1}{N}\sum_{\nu\neq\mu}^{M}\mathbf{z}^{(\nu)T}\mathbf{z}^{(\mu)}.$$

The central limit theorem states that θ_μ is a Gaussian variable. We compute its two first moments:

$$\mathbb{E}\left[\theta_\mu\right] = \mathbb{E}_{\mathbf{z}}\left[\frac{1}{N}\sum_{\nu\neq\mu}^{M}\mathbf{z}^{(\nu)T}\mathbf{z}^{(\mu)}\right] = \mathbb{E}_{\mathbf{z}}\left[\frac{1}{N}\sum_{\nu\neq\mu}^{M}\sum_{i=1}^{N}\mathbf{z}_i^{(\nu)T}\mathbf{z}_i^{(\mu)}\right] = 0$$

$$\mathbb{E}\left[\theta_\mu^2\right] = \mathbb{E}\left[\frac{1}{N^2}\left(\sum_{\nu\neq\mu}^{M}\mathbf{z}^{(\nu)T}\mathbf{z}^{(\mu)}\right)^2\right] = \frac{(M-1)}{N} \underset{M\to\infty}{\longrightarrow} \alpha\,.$$

Therefore $\theta_\mu \sim \mathcal{N}(0,\alpha)$. Finally

$$\begin{aligned}\mathfrak{R}_M(\mathcal{F}_{\text{sign}}) &\geq \mathbb{E}_\theta\left[\frac{1}{M}\sum_{\mu=1}^{M}\text{sign}\left(1+\theta_\mu\right)\right] = \mathbb{E}_\theta\left[\text{sign}\left(1+\theta\right)\right] \\ &= \mathbb{P}\left[\theta \geq -1\right] - \mathbb{P}\left[\theta \leq -1\right] = 2\mathbb{P}\left[\theta \geq -1\right] - 1.\end{aligned}$$

Noting that

$$\mathbb{P}\left[\theta \geq -1\right] = \int_{-\frac{1}{\sqrt{\alpha}}}^{\infty}\mathrm{D}\theta = \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2\alpha}}\right) \sim_{\alpha\to\infty} \frac{1}{2} + \frac{1}{\sqrt{2\pi\alpha}}\,,$$

we obtain a lower bound for the Rademacher complexity

$$\mathfrak{R}_M\left(\mathcal{F}_{\text{sign}}\right) \geq \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\alpha}} = \mathcal{O}\left(\frac{1}{\sqrt{\alpha}}\right).$$

Combing the lower and upper bounds shows that $\mathfrak{R}_M(\mathcal{F}_{\text{sign}}) = \mathcal{O}(1/\sqrt{\alpha})$ in the asymptotic and large α limit.

5.3 The statistical physics approach

5.3.1 The Gardner capacity for classification problems

Let us now see how statistical physics approaches the question of classification. We stay focused on Gaussian i.i.d. data where $\mathbf{x} \sim \mathcal{N}(0,1)^N$. Consider a function class, for instance the perceptron one $\mathcal{F}_{\text{sign}}$: $\{f_{\mathbf{w}} : \mathbf{x} \to \text{sign}\left(\frac{1}{\sqrt{N}}\mathbf{w}^T\mathbf{x}\right)\}$. A classic question in the literature was to compute how many misclassified examples can be obtained for a given rule, used to generate the labels [49], for instance using a function $f_{\mathbf{w}}$ from the perceptron class. Given M samples $\{y^{(\mu)}, \mathbf{x}^{(\mu)}\}_{\mu=1,\dots,M}$, the number of wrongly classified training samples can be defined through the Hamiltonian:

$$\mathcal{H}\left(\{\mathbf{y},\mathbf{X}\},\mathbf{w}\right) \equiv \sum_{\mu=1}^{M}\mathbb{1}\left[y^{(\mu)} \neq f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right)\right] = \frac{1}{2}\left(M - \sum_{\mu=1}^{M}y^{(\mu)}f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right)\right). \tag{5.24}$$

Note that this Hamiltonian is similar to the one we are familiar with through the compressed sensing calculation, but taken in the limit where the inverse temperature β goes to infinity. The

Boltzmann distribution then turns into a delta function, hence the indicator in (5.24). A classical problem in statistical physics is to compute the critical "storage" capacity: we take M examples $\{\mathbf{x}^{(\mu)}\}_{\mu=1,\dots,M}$ and labels $\{y^{(\mu)}\}_{\mu=1,\dots,M}$ randomly chosen between ± 1. We then need to find the maximal number of samples M_c that can be correctly classified by the hypothesis class, and from there we define the critical capacity, also called the *Gardner capacity* [59] as $\alpha_c = M_c/N$. This computation was first achieved thanks to the replica method, let us briefly explain its main idea. The replicated free energy is presented as an integral on the space of weights. Inside the integral, we impose a constraint to select only functions that properly classify the examples i.e. verify $f_{\mathbf{w}}(\mathbf{x}^{(\mu)}) = y^{(\mu)}$ for $\mu = 1, ..., M$. The free energy yields a function of the order parameter q, which quantifies the overlap between perceptron weights $\mathbf{w}$ of functions that are allowed by the proper classification constraint. We extremize the free energy which yields a saddle-point equation on q. We then take the limit $q \to 1$, which physically means that only one set of weights exists such that the corresponding function $f_{\mathbf{w}}$ can classify the examples, hence that we have reached the limit number of patterns that can be correctly classified. The limit $q \to 1$ becomes an equation on α, its solution is α_c.

It turns out there exists a deep connection between the Gardner capacity and the VC dimension. Indeed, both of them are linked to the maximum number of points M_c such that there exists a function in the hypothesis class that is able fit the data set. In particular, using Sauer's lemma [140] in the large size limit $M, N \to \infty$, where $\alpha_c = M_c/N$ and $\alpha_{\mathrm{VC}} = d_{\mathrm{VC}}/N$ are kept finite, it is possible to show that the Gardner capacity α_c provides a lower-bound of the VC dimension [49]:

$$\alpha_c \leq 2\alpha_{\mathrm{VC}}\,. \tag{5.25}$$

For instance, in the case of perceptron hypothesis class $\mathcal{F}_{\mathrm{sign}}$, the VC dimension is equal to the input space dimension: $d_{\mathrm{VC}} = N$, yielding $\alpha_{\mathrm{VC}} = 1$; while the Gardner capacity is $\alpha_c = 2$ [37, 59]. Computing the Gardner capacity gathered a lot of work from the statistical physics community, starting with a series of papers in the 90s [59, 85], and kept fueling research through more recent rigorous works [154, 155, 41, 8].

5.3.2 The Rademacher complexity and the ground state energy

The Rademacher complexity might seem mysterious due to its somewhat complicated definition. However, it is in fact linked to a very usual statistical physics quantity, which is none other than the *ground state energy*. We define the Gibbs measure at inverse temperature β as:

$$\langle \dots \rangle_\beta \equiv \frac{\int \mathrm{d}\mathbf{w} \dots e^{-\beta \mathcal{H}(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}}{\int \mathrm{d}\mathbf{w} e^{-\beta \mathcal{H}(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}}. \tag{5.26}$$

We now take the Gibbs measure of the Hamiltonian in (5.24) for any function $f_{\mathbf{w}} \in \mathcal{F}$, then average over $\{\mathbf{y}, \mathbf{X}\}$:

$$\mathbb{E}_{\mathbf{y},\mathbf{X}} \left\langle \frac{\mathcal{H}\left(\{\mathbf{y},\mathbf{X}\},\mathbf{w}\right)}{N} \right\rangle_\beta = \frac{\alpha}{2} \left[1 - \mathbb{E}_{\mathbf{y},\mathbf{X}} \left\langle \frac{1}{M} \sum_{\mu=1}^{M} y^{(\mu)} f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right) \right\rangle_\beta \right], \tag{5.27}$$

where $\alpha = M/N$. Taking the zero temperature limit, i.e. $\beta \to \infty$, we obtain the ground state energy e_{gs}, a quantity commonly used in physics. Interestingly, we recognize the definition of

the Rademacher complexity $\mathfrak{R}_M(\mathcal{F})$ within:

$$\begin{aligned} e_{\text{gs}} &\equiv \lim_{\beta\to\infty}\lim_{N\to\infty}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left\langle\frac{\mathcal{H}\left(\{\mathbf{y},\mathbf{X}\},\mathbf{w}\right)}{N}\right\rangle_{\beta} \\ &= \frac{\alpha}{2}\left(1-\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\sup_{f_{\mathbf{w}}\in\mathcal{F}}\frac{1}{M}\sum_{\mu=1}^{M}y^{(\mu)}f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right)\right]\right) \\ &= \frac{\alpha}{2}\left(1-\mathfrak{R}_M\left(\mathcal{F}\right)\right)\,, \end{aligned} \tag{5.28}$$

where random labels $\mathbf{y}$ play the role of the Rademacher variable ϵ in (5.9). We thus underline a (surprisingly!) simple correspondence between the ground state energy on the perceptron model and the Rademacher complexity. This connection means that the Rademacher complexity can be computed (rather than bounded) for many models using the replica method.

5.3.3 A flavor of understanding of Rademacher bounds on generalization

Before going to actual ground state energy computations, we would like to understand more intuitively why the Rademacher complexity is linked with the generalization gap bound, through a hand-waving explanation. Consider the fraction of mistakes performed by a classifier $f_{\mathbf{w}}$ on the training set, i.e. the training error $\epsilon^M_{\text{train}}(f_{\mathbf{w}})$; and on unknown samples, i.e. the generalization error $\epsilon_{\text{gen}}(f_{\mathbf{w}})$. The worst case scenario while trying to learn how to fit patterns is the absence of underlying rule, meaning that labels are purely random and uncorrelated from input. The generalization error will then remain equal to $1/2$, which is what we get from a random guess. We can then sketch the following heuristic generalization bound:

$$\begin{aligned} \epsilon_{\text{gen}}(f_{\mathbf{w}})-\epsilon^M_{\text{train}}(f_{\mathbf{w}}) &\leq \epsilon^{\text{random labels}}_{\text{gen}}(f_{\mathbf{w}})-\epsilon^{\text{random labels},M}_{\text{train}}(f_{\mathbf{w}}) \\ &= \frac{1}{2}-\epsilon^{\text{random labels},M}_{\text{train}}(f_{\mathbf{w}}) \\ &= \frac{1}{2}\left(1-2\epsilon^{\text{random labels},M}_{\text{train}}(f_{\mathbf{w}})\right) \\ &= \frac{1}{2}\mathbb{E}_{\mathbf{y}}\left[\frac{1}{M}\sum_{\mu=1}^{M}y_\mu f_{\mathbf{w}}\left(\mathbf{x}^{(\mu)}\right)\right] \\ &\leq \frac{1}{2}\hat{\mathfrak{R}}_M\left(\mathcal{F}\right)\,. \end{aligned} \tag{5.29}$$

Note that this heuristic reasoning does not give the *exact* Rademacher generalization bound, but it does bring to light a similar quantity. In fact, the actual stronger and uniform (over all possible $\mathbf{w}\in\mathbb{R}^N$) bound does not have a factor $1/2$, and cannot be fully captured by the simple above argument. Nevertheless, this simple point reflects the gist idea within Rademacher bound. It provides a very pessimistic bound by assuming the worst possible scenario, i.e. fitting data and trying to make predictions while the labels are random. Of course, in real data problems the rule is not random and uncorrelated with inputs; it is then no surprise that the Rademacher bound is not tight [185].

5.4 Consequences and bounds for simple models

5.4.1 Ground state energies of the perceptron

We focus on the perceptron hypothesis class, and take α_c the Gardner capacity for Gaussian i.i.d. data. If the number of samples M is smaller than M_c the maximum number of patterns that can be classified, i.e. $\alpha<\alpha_c$, it is by definition possible to fit all random labels $\mathbf{y}$. Accordingly,

the number of misclassified examples is zero and the ground state energy $e_{\text{gs}} = 0$. This means that the Rademacher complexity is asymptotically equal to 1 for $\alpha < \alpha_c$, by virtue of (5.28). Above the Gardner capacity $\alpha > \alpha_c$, the estimator $f_{\mathbf{w}}$ cannot perfectly fit the random labels and will misclassify some of them, equivalently $e_{\text{gs}} > 0$. We have seen in 5.2 that the Rademacher complexity scales as $1/\sqrt{\alpha}$ when α becomes large; therefore we can make the following reasonable guess:

$$\mathfrak{R}_M(\mathcal{F}) = 1 \quad \text{for} \quad \alpha < \alpha_c\,, \tag{5.30}$$

$$\mathfrak{R}_M(\mathcal{F}) \approx \Theta\left(\sqrt{\frac{\alpha_c}{\alpha}}\right) \quad \text{for} \quad \alpha \gg \alpha_c\,. \tag{5.31}$$

Using the replica method and the mapping with ground state energies (5.28), we shall now see how one can go beyond these simple arguments, and compute the actual precise asymptotic value of the Rademacher complexity.

5.4.2 Computing the ground state energy with the replica method

Reminder on the replica ansatz and calculation

We will now draw out our favorite tool to compute free energies, i.e. the replica method. In Chapter 2, we computed the free energy for compressed sensing, for rotationally invariant matrices as well as Gaussian i.i.d. ones. In Chapter 4, we focused on the replica-symmetric Kabashima free energy formula on rotationally invariant matrices for generalized models, which includes our current interest. However, the formula (4.5) is somewhat difficult to handle, and we will use here its simplification for Gaussian i.i.d. data. We consider a simple generalization of the linear functions hypothesis class. Fix any activation function $\varphi : \mathbb{R} \mapsto \{\pm 1\}$, and define the following hypothesis class

$$\mathcal{F}_\varphi \equiv \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^N \mapsto \{-1, 1\} \\ \mathbf{x} \mapsto \varphi\left(\frac{1}{\sqrt{N}} \mathbf{w}^T \mathbf{x}\right) \end{cases} , \mathbf{w} \in \mathbb{R}^N \right\}. \tag{5.32}$$

As usual, we use the Bayesian framework and start by writing the posterior distribution

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{w}, \mathbf{X}) P(\mathbf{w})}{P(\mathbf{y}, \mathbf{X})} = \frac{e^{-\beta \mathcal{H}(\{\mathbf{y}, \mathbf{X}\}, \mathbf{w})} P_w(\mathbf{w})}{\mathcal{Z}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta)}\,, \tag{5.33}$$

where P_w is the prior distribution of the weights, $\mathcal{Z}$ is the partition function associated to the Hamiltonian eq. (5.24) at inverse temperature β

$$\mathcal{Z}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta) = \int_{\mathbb{R}^N} \mathrm{d}\mathbf{w} e^{-\beta \mathcal{H}(\{\mathbf{y}, \mathbf{X}\}, \mathbf{w})} P_w(\mathbf{w})\,. \tag{5.34}$$

From there, we define the free energy at inverse temperature β:

$$\Phi_{\mathbf{y}, \mathbf{X}}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta) = -\lim_{N \to \infty} \frac{1}{N\beta} \log \mathcal{Z}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta)\,. \tag{5.35}$$

Being interested in the *typical case*, we want to compute the averaged free energy

$$\Phi(\alpha, \beta) \equiv \mathbb{E}_{\mathbf{y}, \mathbf{X}} \left[\Phi_{\mathbf{y}, \mathbf{X}}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta)\right]. \tag{5.36}$$

We follow up by writing the replica trick

$$-\frac{1}{N\beta}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\log \mathcal{Z}(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)\right] = -\frac{1}{N\beta}\lim_{n\to 0}\frac{\partial \log \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^n\right]}{\partial n}\,, \tag{5.37}$$

and after computing the replicated free energy, we assume there exists an analytical continuation for $n \in \mathbb{R}$ and that we can revert both limits, so that

$$\Phi(\alpha,\beta) = \lim_{n\to 0}\left[\lim_{N\to\infty} -\frac{1}{N\beta}\frac{\partial \log \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^n\right]}{\partial n}\right]. \tag{5.38}$$

This computation can be found in [184], and we provide some more detail in appendix E. Now remember that in previous chapters, we stayed in clear waters and only performed replica-symmetric calculations. Indeed, we were either in Bayes-optimal teacher-student scenarios which guarantees replica symmetry thanks to Nishimori's identity, or studying convex penalized problems which made sure that the free energy extremum corresponds to a single configuration. We will now need to row on and break replica symmetry. Recall that computing Φ (5.38) reduces to an optimization problem over two symmetric matrices $\mathbf{Q}, \hat{\mathbf{Q}} \in \mathbb{R}^{n\times n}$, where off-diagonal terms measure the *overlaps* between the weights of different replicas, while the diagonal term is fixed to $\mathbb{E}\left[\frac{1}{N}\|\mathbf{w}\|_2^2\right]$. As explained in 1.1.4, we can define a hierarchy of ansatz on these matrices $\mathbf{Q}$ and $\hat{\mathbf{Q}}$, starting with replica-symmetry (RS), the going to one-step or more replica symmetry breaking (RSB). The different ansatz describe different solution space structures. While in some problems the RS or 1RSB ansatz are sufficient, in others only the infinite step solution (full-RSB) gives the exact ansatz [104, 154, 155]. However, the 1RSB approach is usually an accurate approximation, so we might not need to use the full-RSB ansatz, as each additional step of symmetry breaking makes the computation more tedious.

General expressions of RS and 1RSB free energy for Gaussian i.i.d. data

The RS and 1RSB computation of average free energy for Gaussian i.i.d. matrices [98, 50, 180] yield:

$$\Phi^{(\mathrm{RS})}_{\mathrm{iid}}(\alpha,\beta) = -\frac{1}{\beta}\operatorname*{Extr}_{q_0,\hat{q}_0}\left\{\frac{1}{2}\left(q_0\hat{q}_0 - 1\right) + \Psi^{(\mathrm{RS})}_{\mathrm{w}}(\hat{q}_0) + \alpha\Psi^{(\mathrm{RS})}_{\mathrm{out}}(q_0,\beta)\right\},$$

$$\Phi^{(\mathrm{1RSB})}_{\mathrm{iid}}(\alpha,\beta) = -\frac{1}{\beta}\operatorname*{Extr}_{q_0,q_1,\hat{q}_0,\hat{q}_1,m}\left\{\frac{1}{2}\left(q_1\hat{q}_1 - 1\right) + \frac{m}{2}\left(q_0\hat{q}_0 - q_1\hat{q}_1\right) + \Psi^{(\mathrm{1RSB})}_{\mathrm{w}}(\hat{q}_0,\hat{q}_1) + \alpha\Psi^{(\mathrm{1RSB})}_{\mathrm{out}}(q_0,q_1,\beta)\right\}$$

where q_0, q_1 denote the overlap order parameters. The auxiliary functions read

$$\Psi^{\mathrm{m}}_{w}(\hat{q}_0) \equiv \mathbb{E}_{\xi_0}\log\mathbb{E}_w\left[\exp\left(\frac{(1-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right] \tag{5.39}$$

$$\Psi^{(\mathrm{RS})}_{\mathrm{out}}(q_0,\beta) \equiv \mathbb{E}_y\mathbb{E}_{\xi_0}\log\mathbb{E}_z\left[\mathcal{I}\left(y\middle|\sqrt{Q-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right] \tag{5.40}$$

$$\Psi^{(\mathrm{1RSB})}_{\mathrm{w}}(\hat{q}_0,\hat{q}_1) \equiv \frac{1}{m}\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_w\left[\exp\left(\frac{(1-\hat{q}_1)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1-\hat{q}_0}\xi_1\right)w\right)\right]^m\right) \tag{5.41}$$

$$\Psi^{(\mathrm{1RSB})}_{\mathrm{out}}(q_0,q_1,\beta) \equiv \frac{1}{m}\mathbb{E}_y\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_z\left[\mathcal{I}(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{1-q_1}z,\beta)\right]^m\right) \tag{5.42}$$

where ξ_0, ξ_1 are i.i.d. normal random variables, and $y \sim P_y(.)$ the distribution of the random labels. $\mathcal{I}(y|z) = e^{-\beta V(y|z)}$ is a temperature dependent cost function, where the generic cost function V reads in our case $V(y|z) = \mathbb{1}\left[y \neq \varphi(z)\right]$. In another setting, we could have replaced it with a different loss, such as the squared one. Those expressions are valid for any generic weight distribution $P_w(.)$ and non-linearity φ. The detailed computation can be found in Appendix E, in particular (E.19) and (E.31). The general method to find the ground state energy is to take

the zero temperature limit

$$e_{\text{gs,iid}}(\alpha) \equiv \lim_{\beta\to\infty} \Phi_{\text{iid}}(\alpha,\beta)\,, \tag{5.43}$$

while handling carefully the scaling of the optimized order parameters in this limit.

Spherical perceptron

The most commonly studied model, ever since Gardner's initial works [58, 57, 59, 58] is the spherical model of the perceptron with continuous weights, i.e. with weights $\mathbf{w} \in \mathbb{R}^N$ such that $\|\mathbf{w}\|_2^2 = N$. The spherical constraint allows to have a well-defined model which excludes diverging or vanishing weights. In this case, the Gardner capacity is rigorously known to be equal to $\alpha_c = 2$ [37].

The RS, 1RSB and 2RSB [98, 50, 180] are computed in appendix E.6. In the RS case, we take the zero temperature limit $\beta \to \infty$ with $q_0 \to 1$ and $\chi = \beta(Q - q_0)$ finite. In the 1RSB ansatz, we take $q_1 \to 1, m \to 0$ keeping $\chi \equiv \beta(Q - q_1)$ and $\Omega_0 \equiv \frac{m\beta}{\chi}$ finite. Those limits and scalings lead to the following expressions of the ground states energies:

$$e^{(\text{RS})}_{\text{gs,iid}} = \operatorname*{Extr}_{\chi} \left\{ -\frac{1}{2\chi} + \alpha \mathbb{E}_{y,\xi_0} \min_z \left[V(y|z) + \frac{(z-\xi_0)^2}{2\chi} \right] \right\} \tag{5.44}$$

$$e^{(\text{1RSB})}_{\text{gs,iid}} = \operatorname*{Extr}_{\chi,\Omega_0,q_0} \left\{ \frac{1}{2\Omega_0\chi} \log\left(1 + \Omega_0(1-q_0)\right) + \frac{q_0}{2\chi\left(1+\Omega_0(1-q_0)\right)} \right. \tag{5.45}$$
$$\left. + \frac{\alpha}{\chi\Omega_0} \mathbb{E}_{\xi_0} \left[\log \mathbb{E}_{\xi_1} \left[e^{-\Omega_0\chi \min_z \left[V(y|z) + \frac{1}{2\chi}\left(z - \sqrt{q_0}\xi_0 - \sqrt{1-q_0}\xi_1\right)^2 \right]} \right] \right] \right\}$$

where the cost function is $V(y|z) = \mathbb{1}\left[y \neq \varphi(z)\right]$. The details of the derivation via the replica method and the expression for the 2RSB ansatz are given in appendix E.6. The results for Rademacher variable y and with $\varphi(z) = \text{sign}(z)$ are depicted in Fig. 5.2.

Till now, we have thought of using statistical physics to gain knowledge about the Rademacher complexity. Nevertheless, this exchange of goods works is double-sided, and the bounds on the Rademacher complexity can also bear consequences for statistical physics. Indeed, we do not know a priori what large α scaling the ground state energy has. However, guessing that the Rademacher complexity scales as $1/\sqrt{\alpha}$ for large values of α – namely that there exists a constant $\mathcal{C}$ such that $\mathfrak{R}_m(\mathcal{F}) \underset{\alpha\to\infty}{\approx} \frac{\mathcal{C}}{\sqrt{\alpha}}$ – implies that the ground state energy behaves for large α as

$$e_{\text{gs}}(\alpha) = \frac{\alpha}{2}\left(1 - \mathfrak{R}_M(\mathcal{F})\right) \underset{\alpha\to\infty}{\longrightarrow} \frac{\alpha}{2}\left(1 - \frac{\mathcal{C}}{\sqrt{\alpha}}\right). \tag{5.46}$$

Let us see how this prediction compares with our actual ground state expressions. We first notice that the replica-symmetric solution complexity fails to deliver the correct scaling as sketched in Fig. 5.2: the scaling drawn from (5.46) must not be entirely trivial. On the other hand, the 1RSB and 2RSB solutions we used (which are expected to be numerically very close to the full-RSB one, that is harder to evaluate), seem in Fig. 5.2 to display the correct scaling. The connection between the Rademacher complexity and the ground state free energy allows to forecast through (5.46) the scaling of the energy in the large α regime that is only satisfied through a replica-symmetry breaking ansatz. An intriguing question would be to determine the constant $\mathcal{C}$ through statistical physics, which is still beyond our grasp.

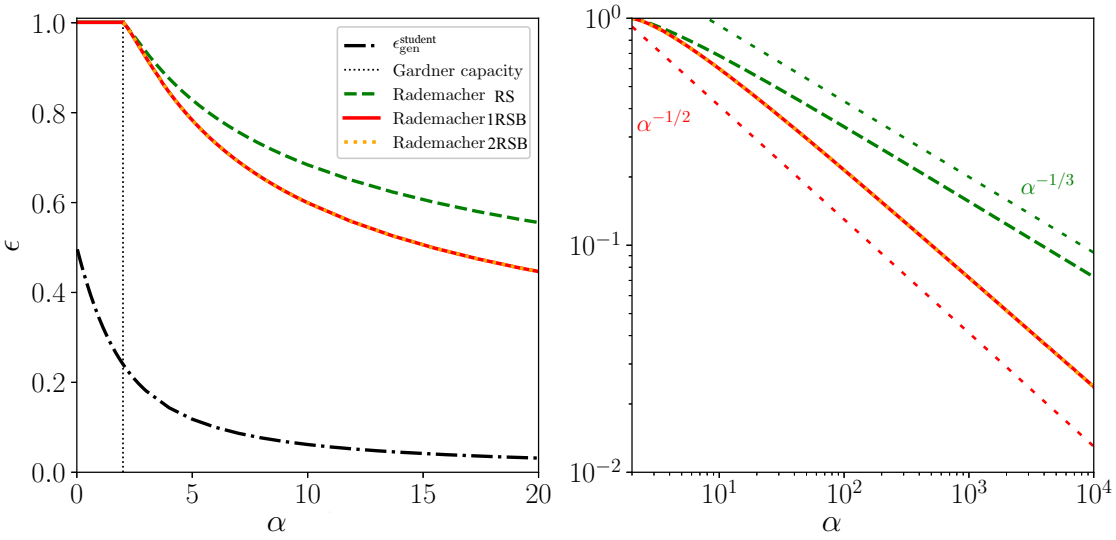


FIGURE 5.2: Explicit Rademacher complexity for the spherical perceptron (with Gardner capacity $\alpha_c = 2$). For $\alpha < \alpha_c$ the problem of properly classifying patterns is satisfiable, so the number of errors is zero and the Rademacher complexity is constant to 1. For $\alpha > \alpha_c$, the problem is not satisfiable and $e_{\text{gs}} > 0$. In the case of the spherical perceptron, RS (dashed green) and 1RSB (red) ansatz provide neatly different results that scale respectively with $\alpha^{-1/3}$ and $\alpha^{-1/2}$ in the large α limit, as shown ont the right picture where scalings are represented by colored slightly dashed lines. Performing the 2RSB calculation (see appendix E.6) does not change the scaling and the difference with respect to 1RSB is visually imperceptible. The black dotted-dashed curve is the generalization error in the teacher-student scenario [15]. Note the large gap between the worst case Rademacher bound and the actual teacher-student generalization error.

Binary perceptron

Another common choice for the weights distribution is the binary prior $P_w(w) = \frac{1}{2}[\delta(w-1) + \delta(w+1)]$ [85]. In this case, the replica-symmetric prediction of the Gardner capacity gives a wrong result, which can be seen in different ways: the resulting capacity is larger than 1, which is impossible since the N weights cannot store more than N patterns. Moreover, the obtained value does not match simulation. Finally, computing the stability of the saddle-point of the RS free energy shows that it is unstable [58, 98] past the so-called Almeida-Thouless capacity [39], and local stability is a necessary (but not sufficient) condition for the result to be correct. Upgrading from the RS calculation leads to the 1RSB scheme, in the binary perceptron, the landscape of the model is said to be *frozen* 1RSB, i.e. clustered in point-like dominant solutions, and it turns out that the equation yielding the 1RSB storage capacity amounts to finding an effective temperature that sets the RS entropy to zero [49]. Proceeding with it yields $\alpha_c \approx 0.83$, which is believed to be correct and matches numerical simulation [86]. This prediction is remarkably still not entirely proven, although a lower bound is derived in [41], and the capacity for different types of step perceptrons are rigorously found in [8].

Again, we note that even though we are unsure of the 1RSB global stability, and should possibly replace by a more complex (and ultimately full-RSB) solution, 1RSB already gives the good scaling $\mathfrak{R}_M(\mathcal{F}) = \mathcal{O}(1/\sqrt{\alpha})$, and satisfies the scaling (5.46) for large α, as shown in Fig. 5.3.

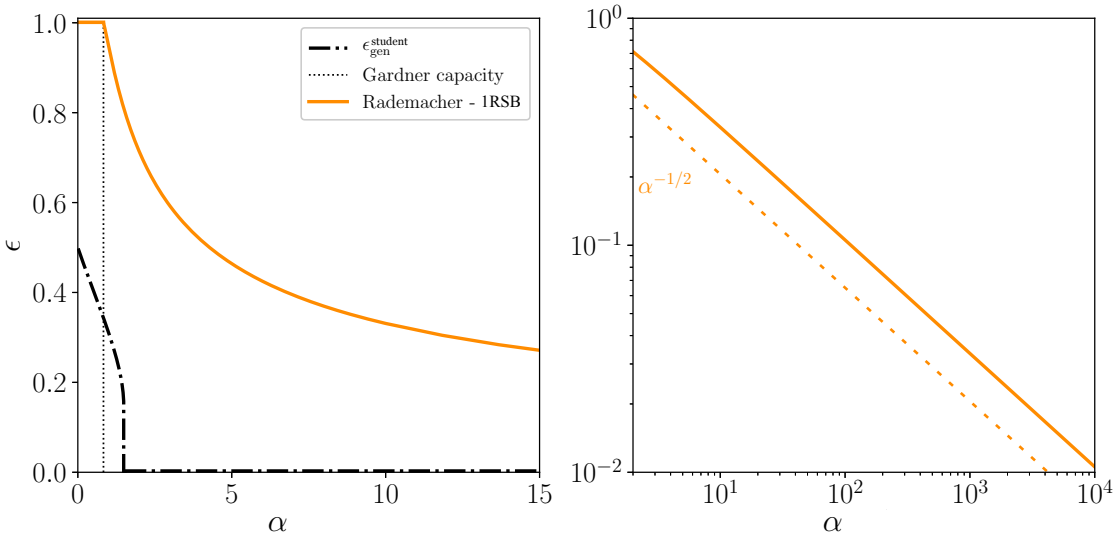


FIGURE 5.3: Explicit Rademacher complexity for the binary perceptron ($\alpha_c \approx 0.83...$). The replica solution (orange) leads again to a $\alpha^{-1/2}$ scaling (dashed orange) of the Rademacher complexity at large α, shown on the right picture. The dotted-dashed black curve is the generalization error in the teacher-student scenario. Again, we obersve a large discrepancy between the worst case bound (Rademacher) and the teacher-student generalization error.

5.4.3 Teacher-student scenario versus worst case Rademacher

The Rademacher bounds are really interesting as they depend only on data distribution, and are valid for *any rule* used to generate the labels, no matter how complicated. In this sense, it is a worst-case scenario on the rule that prescribes labels to data. An opposite take is to consider the teacher-student approach, where we know the teacher rule used to generate labels. This would actually provide the *best-case* generalization gap. by fitting the labels according to the same teacher rule.

We shall assume that the actual labels on the training set are given by the rule

$$y = \text{sign}\left(\frac{1}{\sqrt{N}}\mathbf{w}^{\star\intercal}\mathbf{x}\right), \tag{5.47}$$

with $\mathbf{w}^\star$ the *teacher weights* that can be taken as Rademacher ± 1 variables, or Gaussian ones. Labels are thus generated by feeding i.i.d. random samples to a neural network architecture (the teacher) and are then presented to another neural network (the student) that is trained using this data. Let us compare the worst case Rademacher bound with the actual generalization error achieved by the student.

We first consider the error of a *typical* solution $\mathbf{w}$ sampled from the posterior distribution $\epsilon_{\text{gen}}^{\text{student}}$ for the student. From the generalization error definition (5.11), and since labels are ± 1 variables, we write:

$$\epsilon_{\text{gen}}^{\text{student}} = 1 - \mathbb{E}_{\mathbf{x},\mathbf{w}^\star}\left[\langle f_{\mathbf{w}^\star}(\mathbf{x}) \times f_{\mathbf{w}}(\mathbf{x})\rangle\right] = 1 - q^\star \tag{5.48}$$

where $q^\star = \mathbb{E}_{\mathbf{x},\mathbf{w}^\star}\left[\langle f_{\mathbf{w}^\star}(\mathbf{x}) \times f_{\mathbf{w}}(\mathbf{x})\rangle\right]$ is none other than the replica-symmetric overlap. Recall that it can be computed through the replica method [145, 178, 118], and was rigorously obtained as well in [15], which showed on the way that the resulting generalization error is equal to the

Bayes-optimal error for the quadratic loss. The two *optimistic* (teacher-student) and *pessimistic* (Rademacher) errors can be seen in Fig. 5.2 for spherical weights, and in Fig. 5.3 for binary weights. In this case, since a perfect fit is always possible, the training error is zero and the Rademacher complexity is itself the bound on the generalization error. These two figures show how different the worst and teacher-student case can be in practice, and underline that one should not be surprised when the empirical Rademacher complexity does not give a sharp bound on the generalization gap [185].

5.4.4 Committee machine with Gaussian weights

Knowing that a large gap lies between the Rademacher bound and the teacher-student setting generalization error, we wonder whether we can find a case where the Rademacher bound is void: it means having the Rademacher complexity equal to 1, but achieving good generalization in the teacher-student setting. This actually happens for two-layer networks. Consider a simple hypothesis class, the *committee machine* [49]. It consists in a two-layer network where the second layer has been fixed, such that only weights of the first layer $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_K\} \in \mathbb{R}^{N\times K}$ are learnt. The function class for a committee machine with K hidden units is defined by

$$\mathcal{F}_{\text{com}} \equiv \left\{ f_{\mathbf{W}} : \begin{cases} \mathbb{R}^N \mapsto \{-1, 1\} \\ \mathbf{x} \mapsto \text{sign}\Big(\sum_{k=1}^K \text{sign}\Big(\frac{1}{\sqrt{N}}\mathbf{w}_k^\intercal \mathbf{x}\Big)\Big) \end{cases} \quad \mathbf{W} \in \mathbb{R}^{N\times K} \right\} . \tag{5.49}$$

Instead of computing the Rademacher complexity with the replica method, it will be enough for our current purpose to understand its rough behavior. As discussed in 5.4.1, it is linked to the Gardner capacity. A generic bound by [111] states that α_c it is upper bounded by a term of order $K\log(K)$. Additionally, the Gardner capacity has been computed by the replica method in [115, 170, 181], yielding $\alpha_c = \mathcal{O}(K\sqrt{\log(K)})$. We thus expect that the Rademacher complexity is equal to 1 when α is smaller than a term of order $K\log(K)$, but grows proportionally to $1/\sqrt{\alpha}$ in the large α limit. A reasonable scaling sketch is:

$$\begin{aligned} &\mathfrak{R}_M\left(\mathcal{F}_{\text{com}}\right) = 1 && \text{for} \quad \alpha < \mathcal{O}\left(K\sqrt{\log(K)}\right) \\ &\mathfrak{R}_M\left(\mathcal{F}_{\text{com}}\right) \approx \mathcal{O}\left(\sqrt{\frac{K\sqrt{\log K}}{\alpha}}\right) && \text{for} \quad \alpha \gg \mathcal{O}\left(K\sqrt{\log K}\right) . \end{aligned} \tag{5.50}$$

Let us compare this with the generalization error achieved by the teacher-student case, when the labels are produced by a teacher committee machine as

$$y = \text{sign}\left(\sum_{k=1}^K \text{sign}\left(\frac{1}{\sqrt{N}}\mathbf{w}_k^{\star\intercal}\mathbf{x}\right)\right), \tag{5.51}$$

the error of the student reads

$$\epsilon_{\text{gen}}^{\text{student}} = 1 - \mathbb{E}_{\mathbf{x},\mathbf{W}^\star}\left[\langle f_{\mathbf{W}^\star}(\mathbf{x}) \times f_{\mathbf{W}}(\mathbf{x})\rangle\right] = 1 - q^\star \tag{5.52}$$

where, again $q^\star = \mathbb{E}_{\mathbf{x},\mathbf{W}^\star}\left[\langle f_{\mathbf{W}^\star}(\mathbf{x}) \times f_{\mathbf{W}^\star}(\mathbf{x})\rangle\right]$, has been computed in a series of papers in statistical physics [73, 143], and rigorously derived using the Guerra interpolation method in [7]. Interestingly, in this case, one can get an error that decays as $1/\alpha$ as soon as $\alpha \gg K$. In the large α limit, there is a huge gap between the Rademacher bound $\mathfrak{R}_M\left(\mathcal{F}_{\text{com}}\right)$ that scales as $(K\sqrt{\log(K)}/\alpha)^{1/2}$ and the actual generalization error $\epsilon_{\text{gen}}^{\text{student}} = \mathcal{O}(K/\alpha)$. This disparity further illustrates the substantial difference in behavior one can get between the worst case and teacher-student case analysis, as seen in Fig. 5.4.

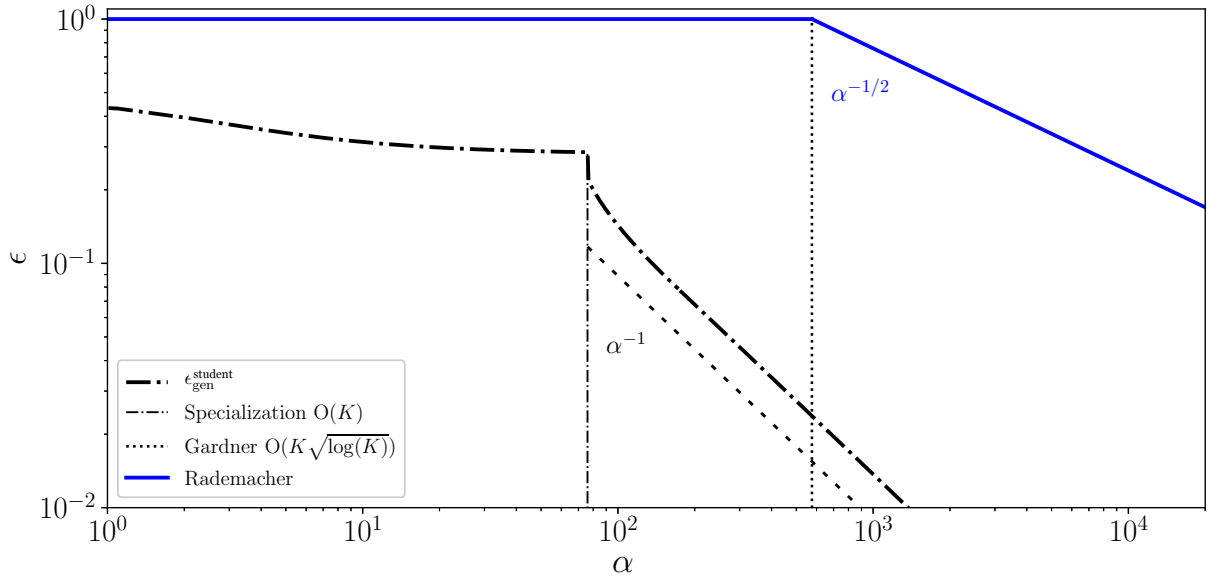


FIGURE 5.4: Illustration of the scaling of the Rademacher complexity (blue) for the fully connected committee machine, compared to the exact generalization error in the teacher-student scenario (dotted-dashed black), scaling as $1/\alpha$ at large α. We observe a large gap between the worst case bound (Rademacher) and the teacher-student result.

5.4.5 Extension to rotationally invariant matrices

Until now, we have focused on Gaussian i.i.d. data and relied on the corresponding replica calculation to compute the exact Rademacher complexity. Fortunately, we also know how to carry out the replica calculation for rotationally invariant data matrices, and the result depends on p_λ, the asymptotic eigenvalue distribution of $\mathbf{X}^T\mathbf{X}$.

For the teacher-student setting, this calculation is performed in [152] and was at the heart of Chapter 4, where we only needed the replica-symmetric ansatz since we added a convex penalty. The replica-symmetric free energy in the limit $\beta \to \infty$ is given in (4.5). The corresponding saddle-point equations can be iterated through (4.27). [152] further establishes the 1RSB free energy. The ground state 1RSB free energy can be simply obtained by taking the $\beta \to \infty$ limit in it. A local stability analysis for both the RS and 1RSB case is also detailed, providing a stability condition that depends on p_λ, and can be checked to gain intuition about the range of validity of the RS and 1RSB ansatz.

Recall that another derivation of the replica free energy (both RS and 1RSB) was provided in [149], and in [78] for the teacher-student case, but as explained in footnote 1, it might not be true for some eigenvalue distributions p_λ. Indeed, it summons a function F which depends on p_λ, but implies the equality between a quenched and annealed average, which is violated if the larger eigenvalue of $\mathbf{X}^T\mathbf{X}$ has non-negligible deviation when $N \to \infty$. Nevertheless, in the case of the Gaussian i.i.d. data matrix, the calculation is correct if we directly replace the function F by its i.i.d. counterpart $F_{iid}(x, y) = -\frac{\alpha}{2}xy$, and we easily recover our i.i.d. free energy expressions.

Summary of Chapter 5 In this chapter, we seeked to establish a connection between statistical learning theory and statistical physics, by looking at the problem of generalization in binary classification. In particular, we focus on the generalization gap which quantifies the difference between the error achieved by an estimator on a training set, and the error on the complete data ensemble. This gap can be bounded in the asymptotic large size limit thanks to the VC dimension, which only depends on the function hypothesis class, or the Rademacher complexity, which additionally depends on data distribution but not on the rule used to generate the labels. We show that the Rademacher complexity is equal to the ground state free energy of the corresponding inference problem, that can be computed thanks to the physics replica method, at least for i.i.d. or rotationally invariant matrices. The simplest replica-symmetric ansatz might fail to describe the dominant configurations of the energy landscape, therefore we also conduct 1RSB and 2RSB calculations. The link between Rademacher and ground state allows to obtain in some settings the precise value of the Rademacher complexity, but also to predict empirically a scaling of $1/\sqrt{\alpha}$ for large α for the ground state energy. We compute the generalization error for the ideal teacher-student case as well, which is considerably better than the worst-case approach (since the latter assumes random labels) embodied by the Rademacher complexity. Open questions arise: it would be interesting to prove the general scaling of the Rademacher complexity in the large α limit (including hypothesis classes with multi-layer networks), and from there deduce the behavior of ground state energies. Extending our analysis to more complicated and structured classes of data matrices (that are not rotationally invariant) also remains a challenge.

Afterword

In this thesis, we went over several inference problems: the inverse Ising model with sparse teacher weights, noiseless compressed sensing, penalized linear regression, penalized generalized linear model, and generalization for binary classification. Our goal was to provide a physical description, enabling us to provide theoretical predictions on reconstruction performance, or statistical properties of estimates and order parameters. In short, we try to think of the problem as a physical system, to pour insight gained from the study of disordered systems, through meaningful quantities such as the free energy, entropy, magnetization...

Our preferred tool was the replica method, which allows to compute the free energy of a spin system in some settings. One major limitation of the replica is the need for randomness assumptions, in particular about the structure of the data matrix. Initial results concerned i.i.d. matrices, and now encompass rotationally invariant matrices through the use of the Harish-Chandra-Itzykson-Zuber integral. However, performing the replica calculation for structured data is a tough challenge, and an ongoing direction of research in statistical physics. Another instrument is the cavity approach as well as TAP equations and the Bethe free energy, which lead to a highway of variational methods on tree like graphs.

A natural concern is to make heuristic results from statistical physics rigorous. This would strengthen physical results, and make them more reliable for different communities. From a mathematical perspective, it is also interesting to develop a formalism that accounts for physical intuition. One common technique is based on the Guerra interpolation method, but we resorted in this thesis to constructive proofs based on convergence analysis of proper algorithms. Indeed, we aimed at drawing concepts from different fields closer and building bridges between them, rather than staying enclosed in a purely physical framework. We centered our work on message passing algorithms, to exploit their rich connections with statistical physics. In particular, we were fond of their state evolution equations, which coincide at their fixed point with replica equations, and provide information to track the algorithm step-by-step. Message passing algorithms can converge very fast, but they suffer from lack of stability which makes them less popular in the optimization community. However, they remain powerful theoretical tools.

Several questions emerge from our interest in message passing algorithms. From a numerical point of view, it would be interesting to design a way of implementing state evolution equations with more stability. In fact, we had to put up quite an effort to successfully implement the state evolution of multi-layer vector approximate message passing. Another promising path would be to extend message passing algorithms to different types of matrices, as we have been again limited to rotationally invariant matrices. In fact, an even broader target would be to find an equivalent of state evolution equations (even heuristically) for a larger class of algorithms, such as expectation propagation schemes. Basically, we would like to enhance ways of tracking algorithmic behavior through analytic equations.

On the whole, we wish to account for more correlated and complicated structures of data. Although rotationally invariant matrices are well-defined and can be generated, understanding what type of data they describe is not so clear. They might also provide a good approximation for structured matrices, or share universal properties as seen for noiseless compressed sensing. A better understanding of the effect of data correlation on algorithms such as belief propagation or expectation propagation, and on replica or TAP equations, is a crucial and flourishing direction of research. While we focused on "simple" scenarios, we strived to refine proof techniques to push back the language frontier between different fields, that necessarily entangle over machine learning and inference problems, and lay foundations towards understanding of multi-layer networks.

Appendix A

Calculation details for inverse Ising problem with sparse teacher

A.1 Computations for L and $\mathcal{S}$

For computing L, the following decomposition of the cavity fields becomes useful:

$$h^* = \sqrt{Q^* - \frac{m^2}{q}}v_* + \sqrt{\frac{m^2}{q}}z, \tag{A.1}$$

$$h^a = \sqrt{Q-q}v_a + \sqrt{q}z, \tag{A.2}$$

where v_a, v_*, z are i.i.d Gaussian variables with zero mean and unit variance. It is easy to confirm that this decomposition reproduces the covariances among $\{h^*, h^1, \cdots, h^n\}$. Using this and performing the integration with respect to v_*, we get

$$L(Q^*, Q, q, m) = \int \mathrm{D}z\ e^{\sqrt{\frac{m^2}{q}}z - \frac{1}{2}\frac{m^2}{q}} \left(\int \mathrm{D}v\ e^{-\beta\ell\left(\sqrt{Q-q}v + \sqrt{q}z\right)}\right)^n, \tag{A.3}$$

where we use the relation $Z_0 = 2e^{\frac{1}{2}Q^*}$, which was canceled with a factor appearing by the integration of v_*. (1.95) is easily derived from this.

To compute the entropic term $\mathcal{S}(\mathbf{C}^{\backslash \mathbf{0}}, \mathbf{J}^*, Q, q, m)$, we use the rescaled variable $\mathbf{W} = \sqrt{N}\mathbf{J}$ and set the integration measure as $\mathrm{Tr}_{\mathbf{J}} = \int d\mathbf{W}$. Here, we use the uniform measure because in the present setting the student has no prior information about the teacher couplings. If certain prior knowledge is available such as the teacher coupling sparseness, it can be suitable to introduce another measure. Further, we represent the delta functions by the Fourier expressions as follows:

$$\delta\left(Q - \frac{1}{N}\sum_{i,j} C_{ij}^{\backslash 0} W_i^a W_j^a\right) = C_1 \int \mathrm{d}\tilde{Q}\ e^{\frac{1}{2}N\tilde{Q}Q - \frac{1}{2}\tilde{Q}\sum_{i,i} C_{ij}^{\backslash 0} W_i^a W_j^a}, \tag{A.4a}$$

$$\delta\left(q - \frac{1}{N}\sum_{i,j} C_{ij}^{\backslash 0} W_i^a W_j^b\right) = C_2 \int \mathrm{d}\tilde{q}\ e^{-N\tilde{q}q + \tilde{q}\sum_{i,j} C_{ij}^{\backslash 0} W_i^a W_j^b}, \tag{A.4b}$$

$$\delta\left(m - \frac{1}{N}\sum_{i,j} C_{ij}^{\backslash 0} W_i^* W_j^a\right) = C_2 \int \mathrm{d}\tilde{m}\ e^{-N\tilde{m}m + \tilde{m}\sum_{i,j} C_{ij}^{\backslash 0} W_i^* W_j^a}, \tag{A.4c}$$

where the integration contour is the imaginary axis and C_1, C_2 are appropriate normalization constants; however, these points are irrelevant and ignored hereafter. Inserting (A.4) into (1.92),

we get

$$e^{N\mathcal{S}} = \int \mathrm{d}\tilde{Q}\,\mathrm{d}\tilde{q}\,\mathrm{d}\tilde{m}\; e^{\mathcal{S}_X} \int \prod_a \mathrm{d}\mathbf{W}^a\; e^U, \tag{A.5}$$

where

$$\mathcal{S}_X = N\left(\frac{1}{2}n\hat{Q}Q - \frac{1}{2}n(n-1)\tilde{q}q - n\tilde{m}m\right), \tag{A.6}$$

$$\begin{aligned} U &= -\frac{1}{2}\tilde{Q}\sum_a (\mathbf{W}^a)^T\,\mathbf{C}^{\backslash\mathbf{0}}\mathbf{W}^a + \tilde{q}\sum_{a<b}(\mathbf{W}^a)^T\,\mathbf{C}^{\backslash\mathbf{0}}\mathbf{W}^b + \tilde{m}\sum_a (\mathbf{J}^*)^T\,\mathbf{C}^{\backslash\mathbf{0}}\mathbf{W}^a \\ &= -\frac{1}{2}(\tilde{Q}+\tilde{q})\sum_a (\mathbf{W}^a)^T\,\mathbf{C}^{\backslash\mathbf{0}}\mathbf{W}^a + \frac{1}{2}\tilde{q}\sum_{a,b}(\mathbf{W}^a)^T\,\mathbf{C}^{\backslash\mathbf{0}}\mathbf{W}^b + \tilde{m}\sum_a (\mathbf{J}^*)^T\,\mathbf{C}^{\backslash\mathbf{0}}\mathbf{W}^a. \end{aligned} \tag{A.7}$$

To decouple different replicas and components of $\{\mathbf{W}^a\}_a$, we use the expression $\mathbf{C}^{\backslash\mathbf{0}} = \mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{\Lambda}$ is the diagonal matrix consisting of the eigenvalues $\{\lambda_i\}_i$ and $\mathbf{U}$ is the appropriate orthogonal matrix. Performing the variable transformation $\tilde{\mathbf{W}} = \mathbf{U}\mathbf{W}$ and applying the Hubbard–Stratonovich transformation, we get

$$\begin{aligned} \int \prod_a \mathrm{d}\mathbf{W}^a\; e^U &= \int \prod_i \mathrm{D}z_i \int \prod_a \mathrm{d}\tilde{\mathbf{W}}^a\; e^{-\frac{1}{2}(\tilde{Q}+\tilde{q})\sum_a\sum_i \lambda_i\left(\tilde{W}_i^a\right)^2 + \sum_a\sum_i\left(\lambda_i\tilde{W}_i^*\tilde{m} + \sqrt{\lambda_i\tilde{q}}z_i\right)\tilde{W}_i^a} \\ &= \int \prod_i \mathrm{D}z_i\; e^{n\sum_i\left\{\frac{1}{2}\frac{\left(\sqrt{\lambda_i}\tilde{W}_i^*\tilde{m}+\sqrt{\tilde{q}}z_i\right)^2}{(\tilde{Q}+\tilde{q})} + \frac{1}{2}\left(\log 2\pi - \log\lambda_i - \log(\tilde{Q}+\tilde{q})\right)\right\}} \\ &= e^{-\frac{N}{2}\log\left(1-\frac{n\tilde{q}}{\tilde{Q}+\tilde{q}}\right) + \frac{1}{2}n\sum_i \frac{\lambda_i\left(\tilde{W}_i^*\right)^2\tilde{m}^2}{\tilde{Q}+\tilde{q}(1-n)} + \frac{n}{2}\sum_i\left(\log 2\pi - \log\lambda_i - \log(\tilde{Q}+\tilde{q})\right)} \equiv e^{\mathcal{S}_J}. \end{aligned} \tag{A.8}$$

Note that the definition of $\tilde{\mathbf{W}}$ implies that $\tilde{\mathbf{W}}$ essentially obeys a Gaussian distribution, and thus the estimator $\hat{\mathbf{J}}$ also does. This knowledge of the distribution can be used for hypothesis testing.

In the asymptotic limit $N \to \infty$, we can use the saddle-point (or Laplace) method to avoid the explicit integrations with respect to $\tilde{Q}, \tilde{q}, \tilde{m}$, yielding

$$\begin{aligned} S = \underset{\tilde{Q},\tilde{q},\tilde{m}}{\mathrm{Extr}}\left\{\frac{\mathcal{S}_X + \mathcal{S}_J}{N}\right\} = \underset{\tilde{Q},\tilde{q},\tilde{m}}{\mathrm{Extr}}\Bigg\{&\frac{1}{2}n\tilde{Q}Q - \frac{1}{2}n(n-1)\tilde{q}q - n\tilde{m}m - \frac{1}{2}\log\left(1 - \frac{n\tilde{q}}{\tilde{Q}+\tilde{q}}\right) \\ &+\frac{1}{2}n\frac{Q^*\tilde{m}^2}{\tilde{Q}+\tilde{q}(1-n)} + \frac{n}{2}\left(\log 2\pi - \log(\tilde{Q}+\tilde{q})\right) - \frac{n}{2N}\,\mathrm{Tr}\log\mathbf{C}^{\backslash\mathbf{0}}\Bigg\}. \end{aligned} \tag{A.9}$$

where we used the relations $\sum_i \lambda_i\left(\tilde{W}_i^*\right)^2 = NQ^*$, $\sum_i \log\lambda_i = \mathrm{Tr}\log\mathbf{C}^{\backslash\mathbf{0}}$. The limit $n \to 0$ leads to

$$\begin{aligned} \lim_{n\to 0}\frac{\mathcal{S}}{n} = \underset{\tilde{Q},\tilde{q},\tilde{m}}{\mathrm{Extr}}\Bigg\{&\frac{1}{2}\tilde{Q}Q + \frac{1}{2}\tilde{q}q - \tilde{m}m + \frac{1}{2}\frac{\tilde{q}+Q^*\tilde{m}^2}{\tilde{Q}+\tilde{q}} \\ &+\frac{1}{2}\left(\log 2\pi - \log(\tilde{Q}+\tilde{q})\right) - \frac{1}{2N}\,\mathrm{Tr}\log\mathbf{C}^{\backslash\mathbf{0}}\Bigg\}, \end{aligned} \tag{A.10}$$

and the extremization condition gives

$$\tilde{Q} = \frac{Q - 2q + m^2/Q^*}{(Q-q)^2},\ \tilde{q} = \frac{q - m^2/Q^*}{(Q-q)^2},\ \tilde{m} = \frac{m/Q^*}{Q-q}. \tag{A.11}$$

Substituting these relations into (A.10), we obtain (1.94). If we ignore the terms related to m and $\tilde{m}$, we have (1.114).

A.2 Derivation of macroscopic parameters R and ρ

To derive the expressions of R and ρ, we can employ the technique of auxiliary variables. We introduce two terms $h_R \sum_a (\mathbf{W}^a)^T \mathbf{W}^a$ and $h_\rho \sum_a (\mathbf{W}^*)^T \mathbf{W}^a$ in (A.7), and perform the same line of computations as in A.1. As a result, the entropic term is modified to the following expression:

$$\begin{aligned}\lim_{n\to 0} \frac{\mathcal{S}}{n} = \underset{\tilde{Q},\tilde{q},\tilde{m}}{\mathrm{Extr}} \Bigg\{ & \frac{1}{2}\tilde{Q}Q + \frac{1}{2}\tilde{q}q - \tilde{m}m \\ & + \frac{1}{2N}\sum_i \left(\frac{(\tilde{m}\lambda_i + h_\rho)^2 \left(\tilde{W}_i^*\right)^2 + \lambda_i \tilde{q}}{(\tilde{Q}+\tilde{q})\lambda_i - 2h_R} + \log 2\pi - \log\left((\tilde{Q}+\tilde{q})(\lambda_i - 2h_R)\right) \right) \Bigg\}. \end{aligned} \tag{A.12}$$

Taking the differentiation with respect to h_ρ and taking the limit $h_\rho, h_R \to 0$, we get

$$\rho = \lim_{h_\rho, h_R \to 0} \frac{\partial}{\partial h_\rho} \lim_{n\to 0} \frac{\mathcal{S}}{n} = \frac{1}{N}\sum_i \frac{\tilde{m}\left(\tilde{W}_i^*\right)^2}{\tilde{Q}+\tilde{q}} = \frac{m}{Q^*} R^*. \tag{A.13}$$

The last expression is obtained by using (A.11). Similarly,

$$\begin{aligned} R &= \lim_{h_\rho, h_R \to 0} \frac{\partial}{\partial h_R} \lim_{n\to 0} \frac{\mathcal{S}}{n} = \frac{1}{N}\sum_i \left(\frac{\tilde{m}^2 \left(\tilde{W}_i^*\right)^2}{\left(\tilde{Q}+\tilde{q}\right)^2} + \frac{\tilde{Q}+2\tilde{q}}{\left(\tilde{Q}+\tilde{q}\right)^2}\frac{1}{\lambda_i} \right) \\ &= \left(\frac{m}{Q^*}\right)^2 R^* + \left(Q - \frac{m^2}{Q^*}\right) \frac{1}{N} \mathrm{Tr}\left(\mathbf{C}^{\backslash \mathbf{0}}\right)^{-1}. \end{aligned} \tag{A.14}$$

which gives (1.99). In the sparse case, we need to compute $\sum_{i\in\bar{\Omega}} \Delta_i^2$ for computing the RSS. By construction, this is equivalent to R when m is absent. Hence $\sum_{i\in\bar{\Omega}} \Delta_i^2$ is given by putting $m = 0$ in (A.14), leading to (1.121).

A.3 Details of numerical experiments

The actual experimental procedures are summarized as follows. We first generated a random graph and the teacher couplings on it, and obtained spin snapshots using MC sampling. Then, we randomly chose a center spin S_0 and learnt the couplings connected to S_0 by minimizing the PL cost function defined with a dataset obtained from the sampled spin configurations. This single sequence of operations provided single values of the quantities of interest, such as $\mathcal{E}$ and Q. To obtain the error bars of those quantities, we repeated this sequence many times. Here, the experiment had three different sources of fluctuations: the generated teacher model (graph shape and couplings), the choice of the center spin, and the MC sampling. We did not discriminate between these three fluctuations unless explicitly mentioned, and we defined the error bar as the standard error among the obtained values according to their recurrence; the number of datasets

obtained this way is denoted as N_{set}. In the MC sampling, we started from a random initial configuration and updated the state by the standard Metropolis method; one MC step (MCS) is defined by N trial flips of spins, where N is the total number of spins. We discarded the first 10^5 MCSs as burn-in to avoid systematic errors from the initialization. Furthermore, to avoid possible correlations in samples, each dataset for learning was generated by subsampling from a much larger dataset, which consists of all the configurations recorded after every few numbers of MCS. The size of the subsampled dataset was chosen to be at least five times smaller than that of the larger dataset. The optimization algorithm is a standard trust-region method using the second-order expansion of the objective function.

Appendix B

Replica calculation for right rotationally invariant matrices

B.1 General setting

We measure $\mathbf{y} \in \mathbb{R}^M$, which has been generated through

$$\mathbf{y} = \mathbf{F}\mathbf{x_0} + \mathbf{w} \tag{B.1}$$

with $\mathbf{w} \sim \mathcal{N}(0, \Delta_0)^N$ a Gaussian noise, and $\mathbf{F} \in \mathbb{R}^{M\times N}$ the sensing matrix. We focus on the asymptotic limit, i.e. $M, N \to \infty$ but $\alpha = M/N$ is fixed of order 1. We want to sample a vector $\mathbf{x}$ from the posterior probability measure

$$P(\mathbf{x}|\mathbf{y}, \mathbf{F}) \cong \frac{1}{\mathcal{Z}} \prod_{i=1}^{n} P(x_i) \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i)^2} \tag{B.2}$$

where $\mathcal{Z}$ is the partition function. In the Bayes-optimal setting, we assume the true prior distribution p_{x_0} is separable and we take $P(x_i) = p_{x_0}(x_i)$. In the case of penalized reconstruction, we want to find

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x}) \right\} \tag{B.3}$$

where f is a convex and separable penalty function, for instance the ℓ_1 norm. To incorporate it in our Bayesian approach, we plug in $P(x_i) \cong e^{-\frac{f(x_i)}{\Delta}}$.

To proceed with the replica calculation, we want to compute Φ the free energy averaged on the randomness of the model, i.e. $\mathbf{F}, \mathbf{x_0}, \mathbf{w}$, which can be done through the replica trick

$$\Phi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}} (\log \mathcal{Z}) = \lim_{N\to\infty} \frac{1}{N} \lim_{n\to 0} \frac{\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(\mathcal{Z}^n) - 1}{n}. \tag{B.4}$$

Introducing n replicas of the system, we want the replicated partition function

$$\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(\mathcal{Z}^n) = \int \prod_{i,a} dx_i^a \prod_{i,a} P(x_i^a) \prod_{\mu} \mathbb{E}_{\mathbf{F},\mathbf{x_0},\mathbf{w}} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\sum_{a=1}^{n}\left(\sum_{i=1}^{N} F_{\mu i} x_{0,i} + w_\mu - \sum_{i=1}^{N} F_{\mu i} x_i^a\right)^2}. \tag{B.5}$$

B.2 Reminder for the case of a Gaussian i.i.d. sensing matrix

If $\mathbf{F}$ is sampled from a Gaussian i.i.d. distribution with zero mean and variance $1/N$, the calculation is straightforward as we can directly compute the average on matrix elements. The

detail can be found in [87]. In particular, looking at the term

$$X_\mu = \mathbb{E}_{\mathbf{F},\mathbf{w}}\left[e^{-\frac{1}{2\Delta}\sum_{a=1}^{n}(\sum_{i=1}^{N}F_{\mu i}(x_{0,i}-x_i^a)+w_\mu)^2}\right], \tag{B.6}$$

we observe that the variables $v_\mu^a \equiv \sum_{i=1}^{N} F_{\mu i}(x_{0,i}-x_i^a)+w_\mu$ obey a joint Gaussian distribution, and we average on it. This average will result in the first two terms inside the final free energy. We introduce the following order parameters for all $a = 1, .., n$:

$$m_a = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_{0,i} \tag{B.7}$$

$$Q_a = \frac{1}{N}\sum_{i=1}^{N} (x_i^a)^2 \tag{B.8}$$

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_i^b \tag{B.9}$$

and we rely on the replica-symmetric ansatz, which is correct in the teacher-student Bayes-optimal setting.

$$\begin{aligned}\Phi_{iid}(Q,q,m,\hat{Q},\hat{q},\hat{m}) = &-\frac{\alpha}{2}\frac{q-2m+\mathbb{E}[x_0^2]+\Delta_0}{\Delta+Q-q} - \frac{\alpha}{2}\log(\Delta+Q-q) + \frac{Q\hat{Q}}{2} - m\hat{m} + \frac{q\hat{q}}{2} \\ &+ \int \mathrm{d}x_0 p_{x_0}(x_0)\int \mathrm{D}z \log\left\{\int \mathrm{d}x P(x) e^{-\frac{\hat{Q}+\hat{q}}{2}x^2+\hat{m}xx_0+z\sqrt{\hat{q}}x}\right\}.\end{aligned} \tag{B.10}$$

B.3 Free energy for right rotationally invariant matrices

Let us now see how this computation adapts to a right rotationally invariant matrix $\mathbf{F}$, as done in [79]. We assume that the distribution p_λ of the eigenvalues of $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ converges and has compact support. An essential point of the calculation is to compute $\prod_\mu X_\mu$ when $\mathbf{F}$ is not i.i.d. Let us write this quantity without splitting it into a product of M terms, in its matricial shape

$$\prod_\mu X_\mu = \mathbb{E}_{\mathbf{F},\mathbf{w}}\left[\exp\left\{-\frac{1}{2\Delta}\sum_{a=1}^{n}\left[\mathbf{F}(\mathbf{x_0}-\mathbf{x}^a)+\xi\right]^T\left[\mathbf{F}(\mathbf{x^0}-\mathbf{x}^a)+\xi\right]\right\}\right]. \tag{B.11}$$

First, we perform the average on the Gaussian noise $\mathbf{w}$:

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}}\left[\exp\left\{-\frac{1}{2\Delta}\sum_{a=1}^{n}\left[\mathbf{F}(\mathbf{x_0}-\mathbf{x}^a)+\mathbf{w}\right]^T\left[\mathbf{F}(\mathbf{x_0}-\mathbf{x}^a)+\mathbf{w}\right]\right\}\right] \\ &= \exp\left\{-\frac{1}{2\Delta}\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\mathbf{F}^T\mathbf{F}(\mathbf{x_0}-\mathbf{x^a})\right\}\mathbb{E}_{\mathbf{w}}\left[\exp\left\{-\frac{\|\mathbf{w}\|_2^2}{2\Delta_0}-\frac{n\|\mathbf{w}\|_2^2}{2\Delta}-\sum_{a=1}^{n}\frac{\mathbf{w}^T\mathbf{F}(\mathbf{x_0}-\mathbf{x^a})}{\Delta}\right\}\right] \\ &\cong \exp\left\{-\frac{1}{2\Delta}\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\mathbf{F}^T\mathbf{F}(\mathbf{x_0}-\mathbf{x^a}) + \frac{\Delta_0}{2\Delta(\Delta+n\Delta_0)}\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\right]\mathbf{F}^T\mathbf{F}\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})\right]\right\}.\end{aligned} \tag{B.12}$$

We are left to average on $\mathbf{F}$:

$$\begin{aligned}&\mathbb{E}_{\mathbf{F}}\left[\exp\left\{-\frac{1}{2\Delta}\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\mathbf{F}^T\mathbf{F}(\mathbf{x_0}-\mathbf{x^a}) + \frac{\Delta_0}{2\Delta(\Delta+n\Delta_0)}\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\right]\mathbf{F}^T\mathbf{F}\left[\sum_{a=1}^{n}(\mathbf{x^0}-\mathbf{x^a})\right]\right\}\right] \\ &= \mathbb{E}_{\mathbf{F}}\left[\exp\left\{\frac{1}{2}\mathrm{Tr}\left(\mathbf{F}^T\mathbf{F}\mathbf{L}(n)\right)\right\}\right]\end{aligned} \tag{B.13}$$

where $\mathbf{L}(n) \equiv -\frac{1}{\Delta}\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})(\mathbf{x_0}-\mathbf{x^a})^T + \frac{\Delta_0}{\Delta(\Delta+n\Delta_0)}\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})\right]\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\right]$.
p_λ is the asymptotic eigenvalue distribution of $\mathbf{F}^T\mathbf{F}$. We assume that it has at least one non-zero eigenvalue and compact support. We define $\lambda_{min}, \lambda_{max}$ the minimum and maximum of its support[1].

Useful transforms We take a detour to introduce a few useful functions that depend on the distribution p_λ, starting with

$$\mathcal{G}_{\mathbf{C}}(x) = \frac{1}{2}\mathrm{Sup}_\Lambda\left\{-\int \mathrm{d}\lambda p_\lambda(\lambda)\log|\Lambda-\lambda| + \Lambda x\right\} - \frac{1}{2}\log|x| - \frac{1}{2}. \tag{B.14}$$

The Stieltjes transform associated with $\mathbf{C}$ is

$$\mathcal{S}_{\mathbf{C}}(x) = \int_{\lambda_{min}}^{\lambda_{max}} \mathrm{d}\lambda \frac{\rho(\lambda)}{x-\lambda}. \tag{B.15}$$

It is properly defined outside of p_λ's support. If x is in the appropriate range i.e. $x < -\mathcal{S}_{\mathbf{C}}(\lambda_{max})$, then we can differentiate with respect to Λ inside $\mathcal{G}_{\mathbf{C}}$ and obtain $\Lambda^* = S^{-1}(-x)$,to reach

$$\mathcal{G}'_{\mathbf{C}}(x) = \frac{1}{2}\mathcal{S}_{\mathbf{C}}^{-1}(-x) - \frac{1}{2x} = \frac{1}{2}\mathcal{R}_{\mathbf{C}}(x) \tag{B.16}$$

where

$$\mathcal{R}_{\mathbf{C}}(x) = \mathcal{S}_{\mathbf{C}}^{-1}(-x) - \frac{1}{x} \tag{B.17}$$

is the R-transform related to the asymptotic eigenvalue distribution p_λ. However, if x lies in a different range and we cannot express $\mathcal{G}_{\mathbf{C}}$ simply as the R-transform's integral. If $x > -\mathcal{S}_{\mathbf{C}}(\lambda_{max})$, then Λ "saturates" at λ_{max} so that

$$\mathcal{G}_{\mathbf{C}}(x > \mathcal{S}_{\mathbf{C}}(\lambda_{max})) = \frac{1}{2}\left[-\int_{\lambda_{min}}^{\lambda_{max}} \mathrm{d}\lambda p_\lambda(\lambda)\log|\lambda_{max}-\lambda| + \lambda_{max}x - \log x - 1\right]. \tag{B.18}$$

In the following, we use the R-transform for the sake of elegance but we should remember that this expression only holds in the right range, otherwise the valid expression would be to replace the R-transform by $2\mathcal{G}'_{\mathbf{C}}$.

Harish-Chandra-Itzykson-Zuber integral We need to average on $\mathbf{F}$ in (B.13), which can be done using the asymptotic form of the Harish-Chandra-Itzykson-Zuber integral [158]. The general idea is the following: considering a rotationally invariant matrix $\mathbf{M} = \mathbf{U}'\mathbf{\Sigma}'\mathbf{V}'$, for any function ϕ of $\mathbf{M}$:

$$\mathbb{E}_{\mathbf{M}}\left[\phi(\mathbf{M})\right] = \mathbb{E}_{\mathbf{M}}\left[\mathcal{D}\mathbf{U}'\,\mathcal{D}\mathbf{V}'\,\phi(\mathbf{U}'\mathbf{\Sigma}'\mathbf{V}'^T)\right] \tag{B.19}$$

where integrating on $\mathcal{D}\mathbf{U}'$, $\mathcal{D}\mathbf{V}'$ represents averages over the ensemble of orthogonal matrices using the Haar measure. The Harish-Chandra-Itzykson-Zuber integral then allows to write the result as a function that depends only on the asymptotic singular value distribution of $\mathbf{M}$. Applied to $\mathbf{M} = \mathbf{F}^T\mathbf{F}$, it states that:

$$\mathbb{E}_{\mathbf{F}}\left[\exp\left\{\frac{1}{2}\mathrm{Tr}\left(\mathbf{F}^T\mathbf{F}\mathbf{L}(n)\right)\right\}\right] = \exp\left\{N \sum_{\lambda \text{ e.v. of } \mathbf{L}(n)/N} G_{\mathbf{F}^T\mathbf{F}}(\lambda)\right\}. \tag{B.20}$$

[1]Note that $\mathbf{F}^T\mathbf{F} = \mathbf{V}^T\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}$: the matrix $\mathbf{U}$ does not show up in the calculation, which explains why we only need rotational invariance for $\mathbf{V}$. This will be different for a generalized linear model, as the dependency upon $\mathbf{F}$ will not be as simple.

To write this explicitly, we need to compute the eigenvalues of matrix $\mathbf{L}(n)$. Some linear algebra shows that $\frac{1}{N}\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})(\mathbf{x_0}-\mathbf{x^a})^T$ and $\frac{1}{N}\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})\right]\left[\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a})^T\right]$ have the same set of eigenvectors

$$\left\{\sum_{a=1}^{n}(\mathbf{x_0}-\mathbf{x^a}),\left((n-1)(\mathbf{x_0}-\mathbf{x^a})-\sum_{b\neq a}(\mathbf{x_0}-\mathbf{x^b})\right)_{a=1,\dots,n}\right\} \tag{B.21}$$

with eigenvalues respectively $\{n(\mathbb{E}[x_0^2]-2m+q)+Q-q, Q-q\}$ and $\{n(Q-q),0\}$. Hence the eigenvalues of $\frac{1}{N}\mathbf{L}(n)$ are

$$-\frac{Q-q}{\Delta}-\frac{n}{\Delta}(\mathbb{E}[x_0^2]-2m+q)+n\frac{\Delta_0}{\Delta_0(\Delta+n\Delta_0)(Q-q)}$$

with multiplicty 1, and $-\dfrac{Q-q}{\Delta}$ with multiplicity $(n-1)$. We clarify the Harish-Chandra-Itzykson-Zuber integral:

$$\mathbb{E}_{\mathbf{F}}\left[\frac{1}{2}\mathrm{Tr}\left(\mathbf{F}^T\mathbf{F}\mathbf{L}(n)\right)\right]=(n-1)\mathcal{G}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right)+\mathcal{G}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}-\frac{n}{\Delta}(\mathbb{E}[x_0^2]-2m+q)+n\frac{\Delta_0}{\Delta_0(\Delta+n\Delta_0)}(Q-q)\right).$$

Further in the replica calculation, we need to take the log of this quantity and keep the term of linear order in n:

$$\lim_{n\to 0}\frac{\partial}{\partial n}\frac{1}{N}\log\left\{\mathbb{E}_{\mathbf{F}}\left[\frac{1}{2}\mathrm{Tr}\left(\mathbf{F}^T\mathbf{F}\mathbf{L}(n)\right)\right]\right\}=\mathcal{G}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right)+\left(-\frac{\mathbb{E}[x_0^2]+q-2m}{\Delta}+\frac{\Delta_0}{\Delta^2}(Q-q)\right)\mathcal{G}'_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right).$$

This contribution to the free energy boils down to the first two terms inside the free energy for a rationally invariant matrix:

$$\begin{aligned}\Phi(Q,q,m,\hat{Q},\hat{q},\hat{m})=&\,\mathcal{G}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right)+\left(-\frac{\mathbb{E}[x_0^2]-2m+q}{\Delta}+\frac{\Delta_0(Q-q)}{\Delta^2}\right)\mathcal{G}'_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right)\\&+\frac{Q\hat{Q}}{2}-m\hat{m}+\frac{q\hat{q}}{2}+\int \mathrm{d}x_0 p_{x_0}(x_0)\int \mathrm{D}z\log\left\{\int \mathrm{d}x P(x)e^{-\frac{\hat{Q}+\hat{q}}{2}x^2+\hat{m}xx_0+z\sqrt{\hat{q}}x}\right\}.\end{aligned} \tag{B.22}$$

We can check the particular case of a Gaussian i.i.d. matrix, which can be seen as a rotationally invariant matrix whose singular values obey the Marchenko-Pastur distribution, defined as

$$p_{\mathrm{MP}}(\lambda)=\left(1-\frac{1}{\alpha}\right)^+\delta(\lambda)+\frac{\sqrt{(\lambda-a)^+(b-\lambda)^+}}{2\alpha\lambda} \tag{B.23}$$

where $(z)^+=\max(0,z)$ and

$$a=(1-\sqrt{\alpha})^2 \qquad\qquad b=(1+\sqrt{\alpha})^2. \tag{B.24}$$

The associated transforms read

$$\mathcal{G}_{\mathrm{MP}}(z)=\frac{\alpha}{2}\log|1-z| \tag{B.25}$$

$$\mathcal{S}_{\mathrm{MP}}(z)=\frac{1-\alpha\pm\sqrt{z^2-2(\alpha+1)z+(\alpha-1)^2}}{2\alpha z} \tag{B.26}$$

$$\mathcal{R}_{\mathrm{MP}}(z)=\frac{\alpha}{1-z} \tag{B.27}$$

and (B.22) happily becomes the free energy for Gaussian i.i.d. matrices (B.10).

B.4 Density evolution equations

To obtain the density evolution equations on (E, V) we start by taking the fixed point equations by differentiating (B.22) with respect to parameters m, q, Q, $\hat{m}$, $\hat{q}$ and $\hat{Q}+\hat{q}$:

$$\hat{m} = \frac{1}{\Delta}\mathcal{R}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) \tag{B.28a}$$

$$\hat{q} = -\frac{1}{\Delta^2}\left(-\mathbb{E}[x_0^2] + 2m - q + \frac{\Delta_0}{\Delta}(Q-q)\right)\mathcal{R}'_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) + \frac{\Delta_0}{\Delta^2}\mathcal{R}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) \tag{B.28b}$$

$$\hat{Q}+\hat{q} = \frac{1}{\Delta}\mathcal{R}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) \tag{B.28c}$$

$$m = \int \mathrm{d}x_0\, x_0 p_{x_0}(x_0) \int \mathrm{D}z f_a\left(\hat{m}, x_0 + z\frac{\sqrt{\hat{q}}}{\hat{m}}\right) \tag{B.28d}$$

$$Q-q = \int \mathrm{d}x_0 p_{x_0}(x_0) \int \mathrm{D}z f_v\left(\hat{m}, x_0 + z\frac{\sqrt{\hat{q}}}{\hat{m}}\right) \tag{B.28e}$$

$$q = \int \mathrm{d}x_0 p_{x_0}(x_0) \int \mathrm{D}z f_a^2\left(\hat{m}, x_0 + z\frac{\sqrt{\hat{q}}}{\hat{m}}\right) \tag{B.28f}$$

with f_a, f_v defined in (2.56). Recall that the squared error and variance of the estimator can be written as functions of order parameters as $E = q - 2m + \mathbb{E}[x_0^2]$, $V = Q - q$. We can easily recognize E and V in the right-hand side of equations (B.28a),(B.28b) and (B.28c). Writing $\hat{m}$, $\hat{q}$ in terms of E, V and combining it with equations (B.28e), (B.28f) we reach

$$E = \mathbb{E}\left[\left\{f_a\left(\frac{\Delta}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}, x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\sqrt{(E - \frac{\Delta_0}{\Delta}V)\mathcal{R}'_{\mathbf{C}}(-\frac{V}{\Delta}) + \Delta_0\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\right) - x_0\right\}^2\right] \tag{B.29a}$$

$$V = \mathbb{E}\left[f_v\left(\frac{\Delta}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}, x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\sqrt{(E - \frac{\Delta_0}{\Delta}V)\mathcal{R}'_{\mathbf{C}}(-\frac{V}{\Delta}) + \Delta_0\mathcal{R}_{\mathbf{C}}(-\frac{V}{\Delta})}\right)\right] \tag{B.29b}$$

with the expectation taken on $x_0 \sim p_{x_0}(x_0)$ and $z \sim \mathcal{N}(0,1)$.

Proximal formulation for linear reconstruction with convex penalization In the case of reconstruction with convex penalty f, we can slightly modify those equations to write them in terms of proximal operators defined as

$$\forall \gamma \in \mathbb{R}^+, x, y \in \mathbb{R} \quad \mathrm{Prox}_{\gamma f}(y) \equiv \arg\min_x \left\{f(x) + \frac{1}{2\gamma}(x-y)^2\right\}. \tag{B.30}$$

They will appear naturally in place of functions f_a and f_v when we derive the saddle-point equations, this time taking $P(x) = e^{-\frac{f(x)}{\Delta}}$, then the limit $\Delta \to 0$. For instance, (B.28d) yields

$$m = \int dx_0 \phi(x_0) x_0 \int Dz \int \frac{dx}{\tilde{Z}} x \exp\left\{-\frac{f(x)}{\Delta} + \frac{xx_0}{\Delta}\mathcal{R}_{\mathbf{C}}\left(-\frac{V}{\Delta}\right) - \frac{x^2}{2\Delta}\mathcal{R}_{\mathbf{C}}\left(-\frac{V}{\Delta}\right) + z\sqrt{\hat{q}}x\right\} \tag{B.31}$$

with $\tilde{Z} = \int dx\, e^{-\frac{1}{\Delta}f(x) - \frac{\hat{Q}+\hat{q}}{2}x^2 + \hat{m}xx_0 + z\sqrt{\hat{q}}x}$. Equation (B.28b) also reads

$$\hat{q} = \frac{\Delta_0}{2\Delta^2}\mathcal{R}_{\mathbf{C}}\left(-\frac{V}{\Delta}\right) + \frac{1}{2\Delta}\left(E - \frac{\Delta_0}{\Delta}V\right)\mathcal{R}'_{\mathbf{C}}\left(-\frac{V}{\Delta}\right). \tag{B.32}$$

which can be inserted into (B.31). We then want to take the limit $\Delta \to 0$. We rescale the variance parameter V into V/Δ, but keep the same name for simplicity. We are now left to do

a Laplace approximation in the integral term of (B.31), to reach

$$
\begin{aligned}
m = \lim_{\Delta\to 0} \int \mathrm{d}x_0 p_{x_0}(x_0) x_0 \int \mathrm{D}z \int \frac{\mathrm{d}x}{\tilde{Z}} x \\
\exp\left\{-\frac{1}{\Delta}\left[f(x) + \frac{x^2}{2}\mathcal{R}_{\mathbf{C}}(-V) - \frac{x x_0}{2}\mathcal{R}_{\mathbf{C}}(-V) - zx\sqrt{\Delta_0 \mathcal{R}_{\mathbf{C}}(-V) + (E-\Delta_0 V)\mathcal{R}'_{\mathbf{C}}(-V)}\right]\right\} \\
= \mathbb{E}\left[x_0 \arg\min_x \left\{ f(x) + \frac{\mathcal{R}_{\mathbf{C}}(-V)}{2}\left(x - \left[x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V)}\right]\right)^2 \right\}\right],
\end{aligned}
$$

clearly yielding a proximal operator such that

$$
m = \mathbb{E}\left[x_0 \mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V)}\right)\right]. \tag{B.33}
$$

The same reasoning on (B.28e), (B.28f) also invoke a proximal operator in place of f_a and f_v, and combining them leads to

$$
V = \mathbb{E}\left[\frac{1}{\mathcal{R}_{\mathbf{C}}(-V)}\mathrm{Prox}'_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V)}\right)\right] \tag{B.34a}
$$

$$
E = \mathbb{E}\left[\left\{\mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-V)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V)}\sqrt{(E-\Delta_0 V)\,\mathcal{R}'_{\mathbf{C}}(-V) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V)}\right) - x_0\right\}^2\right]. \tag{B.34b}
$$

where V is the rescaled variance of the estimator.

Appendix C

Details for Oracle VAMP convergence proof

C.1 Definitions and assumptions

Proper and closed convex functions A convex function is *proper* if its domain is non-empty, and if it never attains $-\infty$. A convex function $f : \mathbb{R}^N \to \mathbb{R}$ with domain $\mathrm{dom}(f)$ is *closed* if for each $\alpha \in \mathbb{R}$, the sublevel set $\{\mathbf{x} \in \mathrm{dom}(f) | f(\mathbf{x}) \leq \alpha\}$ is a closed set.

Empirical convergence with p-th order moment This paragraph reproduces appendix B of [132], which reviews the analysis framework from [19]. It anchors the definitions necessary for a rigorous analysis of VAMP and state evolution statement. The building blocks are the notions of *vector sequence* and *pseudo-Lipschitz function*, which allow to define the *empirical convergence with p-th order moment*. Consider a vector of the form

$$\mathbf{x}(N) = (\mathbf{x}_1(N), ..., \mathbf{x}_N(N)) \tag{C.1}$$

where each sub-vector $\mathbf{x}_n(N) \in \mathbb{R}^r$ for any given $r \in \mathbb{N}^*$. For $r = 1$, which is the case we are in, $\mathbf{x}(N)$ is denoted a *vector sequence*.

Given $p \geqslant 1$, a function $\mathbf{f} : \mathbb{R}^r \to \mathbb{R}^s$ is said to be *pseudo-Lipschitz continuous of order p* if there exists a constant $C > 0$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^s$:

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\|_2 \leqslant C\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \left[1 + \|\mathbf{x}_1\|_2^{p-1} + \|\mathbf{x}_1\|_2^{p-1}\right] \tag{C.2}$$

Then, a given vector sequence $\mathbf{x}(N)$ *converges empirically with p-th order moment* if there exists a random variable $X \in \mathbb{R}^r$ such that:

i) $\mathbb{E}|X|^p < \infty$; and

ii) for any scalar-valued pseudo-Lipschitz continuous $\mathbf{f}(.)$ of order p,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{f}(x_n(N)) = \mathbb{E}[f(X)] \text{ a.s.} \tag{C.3}$$

Note that defining an empirically converging singular value distribution implicitly defines a sequence of matrices $\mathbf{F}(N)$ which are rotationally invariant.

We also recall the definition of *uniform Lipschitz continuity*. For a given mapping $\phi(\mathbf{x}, A)$ defined on $\mathbf{x} \in \mathcal{X}$ and $A \in \mathbb{R}$, we say it is *uniform Lipschitz continuous* in $\mathbf{x}$ at $A = \bar{A}$ if there

exists constants L_1 and $L_2 \geqslant 0$ and an open neighborhood U of $\bar{A}$ such that:

$$\|\phi(\mathbf{x}_1, A) - \phi(\mathbf{x}_2, A)\|_2 \leqslant L_1 \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \tag{C.4}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $A \in U$; and

$$\|\phi(\mathbf{x}, A_1) - \phi(\mathbf{x}, A_2)\|_2 \leqslant L_2(1 + \|\mathbf{x}\|_2)|A_1 - A_2| \tag{C.5}$$

for all $\mathbf{x} \in \mathcal{X}$ and $A_1, A_2 \in U$.

State evolution assumptions The state evolution theorem (see Theorem 1 (i-ii-iii) from [132]) holds if we fulfill the main assumptions stated in 3.1, as well as the following three conditions:

- $\alpha_1^{(t)}$ must be in $[0, 1]$. This is always verified because a proximal operator has Lipschitz constant 1, hence its derivative is smaller than one, and α_1 is an expectation value of this derivative.
- The functions defining A_i and $\mathcal{E}_i$ must be continuous at the points prescribed by the SE equations. This holds true as well since proximals of convex functions are continuous.
- Finally the denoisers (here the proximals) and their derivatives need to be uniformly Lipschitz in their arguments at their parameters. This is again verified from properties of proximal operators.

C.2 Equivalence of replica equations and state evolution fixed point

In this section, we play with the fixed point equations of state evolution (3.20) to write them under the same form of replica fixed point equations. The fixed point of SE, where we have inserted the fixed point conditions $V_1 = V_2$, $\mathcal{E}_1 = \mathcal{E}_2$, and removed the time subscripts, reads:

$$\alpha_1 = \mathbb{E}\left[\mathrm{Prox}'_{\frac{1}{A_1}f}(x_0 + P_1)\right] \tag{C.6a}$$
$$\alpha_2 = A_2 \mathcal{S}_{\mathbf{C}}(-A_2) \tag{C.6b}$$
$$V_1 = \frac{\alpha_1}{A_1} = \frac{\alpha_2}{A_2} \tag{C.6c}$$
$$\frac{1}{V_1} = A_1 + A_2 \tag{C.6d}$$
$$\tau_2 = \frac{1}{(1-\alpha_1)^2}\left[\mathcal{E}_1 - \alpha_1^2 \tau_1\right] \tag{C.6e}$$
$$\tau_1 = \frac{1}{(1-\alpha_2)^2}\left[\mathcal{E}_1 - \alpha_2^2 \tau_2\right] \tag{C.6f}$$
$$\mathcal{E}_1 = \mathbb{E}\left[\left(\mathrm{Prox}_{f/A_1}(x_0 + P_1) - x_0\right)^2\right] = \mathbb{E}\left[\frac{\Delta_0 \lambda_{\mathbf{C}} + \tau_2 A_2^2}{(\lambda_{\mathbf{C}} + A_2)^2}\right]. \tag{C.6g}$$

Our goal is to rewrite those relations involving only two variables, $\mathcal{E}_1$ and V_1. First off, (C.6b) and (C.6c) give

$$V_1 = \mathcal{S}_{\mathbf{C}}(-A_2). \tag{C.7}$$

Then (C.6d) turns into

$$A_1 = \frac{1}{V} - A_2 = \frac{1}{V_1} + \mathcal{S}_{\mathbf{C}}^{-1}(-V_1) = \mathcal{R}_{\mathbf{C}}(-V_1). \tag{C.8}$$

Combined with (C.6a) and (C.6c):

$$V_1 = \frac{1}{\mathcal{R}_{\mathbf{C}}(-V_1)} \mathbb{E}_{x_0,P_1}\left[\mathrm{Prox}'_{f/A_1}(x_0 + P_1)\right] \tag{C.9}$$

where P_1 is a Gaussian variable of variance τ_1. We now want to explicit the expression of τ_1. Starting from (C.6e), using (C.6c) and (C.6d) provides

$$\tau_2 = \frac{1}{(A_2V_1)^2}\left[\mathcal{E}_1 - \tau_1(1 - A_2V_1)^2\right]. \tag{C.10}$$

Turning to (C.6f), and plugging in the error definition from (C.6g) yields

$$\tau_1 = \frac{1}{(1 - A_2V_1)^2}\left(\mathbb{E}\left[\Delta_0 \frac{\lambda_{\mathbf{C}}}{(\lambda_{\mathbf{C}} + A_2)^2} + \tau_2 \frac{A_2^2}{(\lambda_{\mathbf{C}} + A_2)^2}\right] - \tau_2(A_2V_1)^2\right) \tag{C.11}$$

$$= \frac{1}{(1 - A_2V_1)^2}\left(\mathbb{E}\left[\frac{\Delta_0}{(\lambda_{\mathbf{C}} + A_2)} - \frac{A_2\Delta_0}{(\lambda_{\mathbf{C}} + A_2)^2} + \tau_2 \frac{A_2^2}{(\lambda_{\mathbf{C}} + A_2)^2}\right] - \tau_2(A_2V_1)^2\right) \tag{C.12}$$

$$\tau_1 = \frac{1}{(1 - A_2V_1)^2}\left(\Delta_0\mathcal{S}_{\mathbf{C}}(-A_2) - A_2\Delta_0\mathcal{S}'_{\mathbf{C}}(-A_2) + \tau_2A_2^2\mathcal{S}'_{\mathbf{C}}(-A_2) - \tau_2(A_2V_1)^2\right), \tag{C.13}$$

then incorporating τ_2's expression (C.10) reads

$$\tau_1 = \frac{\Delta_0V_1^2}{(1 - A_2V_1)^2\mathcal{S}'_{\mathbf{C}}(-A_2)}(\mathcal{S}_{\mathbf{C}}(-A_2) - A_2\mathcal{S}'_{\mathbf{C}}(-A_2)) + \frac{\mathcal{E}_1}{(1 - A_2V_1)^2\mathcal{S}'_{\mathbf{C}}(-A_2)}(\mathcal{S}'_{\mathbf{C}}(-A_2) - V_1^2). \tag{C.14}$$

We have now expressed the variance of the Gaussian variable P_1 as a function of $\mathcal{E}_1$ and V_1. A somewhat heavy (but easy) step is now necessary. We repeatedly use equalities $A_1V_1 = 1 - A_2V_1$; $A_2 = -\mathcal{S}_{\mathbf{C}}^{-1}(V_1)$, and the identity

$$\mathcal{R}'_{\mathbf{C}}(x) = (\mathcal{S}_{\mathbf{C}}^{-1}(-x))' + \frac{1}{x^2} = \frac{-1}{\mathcal{S}'_{\mathbf{C}}(\mathcal{S}_{\mathbf{C}}^{-1}(-x))} + \frac{1}{x^2} \tag{C.15}$$

to reach

$$\tau_1 = \frac{\mathcal{E}_1}{\mathcal{R}_{\mathbf{C}}^2(-V_1)}\left(\frac{-1}{\mathcal{S}'_{\mathbf{C}}(\mathcal{S}_{\mathbf{C}}^{-1}(V_1))} + \frac{1}{V_1^2}\right) + \frac{\Delta_0}{\mathcal{R}_{\mathbf{C}^2}(-V_1)}\left(\mathcal{R}_{\mathbf{C}}(-V_1) - V_1\left(\frac{-1}{\mathcal{S}'_{\mathbf{C}}(\mathcal{S}_{\mathbf{C}}^{-1}(V_1))} + \frac{1}{V_1^2}\right)\right)$$

$$= \frac{1}{\mathcal{R}_{\mathbf{C}}^2(-V_1)}\left(\left(\mathcal{E}_1 - \Delta_0V_1\right)\mathcal{R}'_{\mathbf{C}}(-V_1) + \Delta_0\mathcal{R}_{\mathbf{C}}(-V_1)\right). \tag{C.16}$$

Adding this to equation (C.17) yields

$$V_1 = \frac{1}{\mathcal{R}_{\mathbf{C}}(-V_1)}\mathbb{E}\left[\mathrm{Prox}'_{f/\mathcal{R}_{\mathbf{C}}(-V_1)}(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V_1)}\sqrt{\left(\mathcal{E}_1 - \Delta_0V_1\right)\mathcal{R}'_{\mathbf{C}}(-V_1) + \Delta_0\mathcal{R}_{\mathbf{C}}(-V_1)}\right]. \tag{C.17}$$

Using the established expression for τ_1 inside $\mathcal{E}_1$'s first definition in (C.6g) also reads

$$\mathcal{E}_1 = \mathbb{E}\left[\left\{\mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-V_1)}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}(-V_1)}\sqrt{(\mathcal{E}_1 - \Delta_0 V_1)\,\mathcal{R}'_{\mathbf{C}}(-V_1) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-V_1)}\right) - x_0\right\}^2\right]. \quad \text{(C.18)}$$

Equations (C.17) and (C.18) immediately follow from the fixed point of state evolution, and they are exactly the same as (B.34a) and (B.34b), i.e. the replica equations on (E, V).

C.3 Lipschitz constants of Oracle VAMP operators

In this section, we establish the Lipschitz constants of operators $\mathcal{O}_1$ and $\mathcal{O}_2$, which are successively applied to define one interation of Oracle VAMP (3.23). We will resort to properties of proximal operators, but first cite a few definitions from [18]:

- An operator $T : \mathbb{R}^k \to \mathbb{R}^l$ is *nonexpansive* if it is Lipschitz-continuous with Lipschitz constant equal to 1.
- T is *firmly nonexpansive* if

$$\forall \mathbf{x}, \forall \mathbf{y} \in \mathbb{R}^k \quad \|T(\mathbf{x}) - T(\mathbf{y})\|_2^2 + \|(\mathrm{Id} - T)\mathbf{x} - (\mathrm{Id} - T)\mathbf{y}\|_2^2 \leq \|\mathbf{x} - \mathbf{y}\|_2^2. \quad \text{(C.19)}$$

- Let $\gamma > 0$. T is γ-*cocoercive* if and only if γT is firmly nonexpansive, which is equivalent to writing $T = \frac{1}{2\gamma}(\mathrm{Id} + S)$ with S a nonexpansive operator.

We now cite Proposition 2 from [64], which will help us characterizing the proximal operators: assume that f is σ-strongly convex and β-smooth and that $\gamma > 0$.

i) If $\beta > \sigma$, then $\mathrm{Prox}_{\gamma f} - \frac{1}{1+\gamma\beta}\mathrm{Id}$ is $(\frac{1}{1+\gamma\sigma} - \frac{1}{1+\gamma\beta})^{-1}$-cocoercive. This result also includes the case where f has no smoothness assumption, by taking the limit $\beta = +\infty$.

ii) If $\beta = \sigma$, $\mathrm{Prox}_{\gamma f}$ is 0-Lipschitz.

iii) If we do not have a strong convexity assumption (i.e. f is simply convex and $\sigma = 0$), the property still implies that $\mathrm{Prox}_{\gamma f}$ is firmly nonexpansive.

C.3.1 Lipschitz constant of $\mathcal{O}_1$

Case 1: $0 < \sigma_1 < \beta_1$ The previous proposition yields

$$\mathrm{Prox}_{\frac{1}{A_1}f} = \frac{1}{2}\left(\frac{1}{1+\sigma_1/A_1} + \frac{1}{1+\beta_1/A_1}\right)\mathrm{Id} + \frac{1}{2}\left(\frac{1}{1+\sigma_1/A_1} - \frac{1}{1+\beta_1/A_1}\right)S_1 \quad \text{(C.20)}$$

where S_1 is a non-expansive operator. Replacing in the expression of $\mathcal{O}_1$ leads to:

$$\mathcal{O}_1 = \left(\frac{1}{2V_1}\left(\frac{1}{A_1+\sigma_1} + \frac{1}{A_1+\beta_1}\right) - 1\right)\mathrm{Id} + \frac{1}{2V_1}\left(\frac{1}{1+\sigma_1/A_1} - \frac{1}{1+\beta_1/A_1}\right)S_1\left(\frac{\cdot}{A_1}\right) \quad \text{(C.21)}$$

which, knowing that $A_1 + A_2 = \frac{1}{V_1}$, $\mathcal{O}_1$ has Lipschitz constant:

$$\mathcal{L}_1 = \max\left(\frac{|A_2 - \sigma_1|}{A_1 + \sigma_1}, \frac{|\beta_1 - A_2|}{A_1 + \beta_1}\right). \quad \text{(C.22)}$$

Case 2: $0 < \sigma_1 = \beta_1$ In this case, the proposition above yields

$$\|\mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{x}) - \mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{y})\|_2^2 = \left(\frac{1}{1+\sigma_1/A_1}\right)^2 \|\mathbf{x}-\mathbf{y}\|_2^2 \tag{C.23}$$

which, with the firm non-expansiveness of the proximal operator gives:

$$\begin{aligned}
\|\mathcal{O}_1(\mathbf{x}) - \mathcal{O}_1(\mathbf{y})\|_2^2 &= \frac{1}{V^2}\|\mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{x}/A_1) - \mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{y}/A_1)\|_2^2 && \text{(C.24)}\\
&\quad - 2\frac{A_1}{V_1}\left\langle \frac{\mathbf{x}}{A_1} - \frac{\mathbf{y}}{A_1}, \mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{x}/A_1) - \mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{y}/A_1)\right\rangle + \|\mathbf{x}-\mathbf{y}\|_2^2 && \text{(C.25)}\\
&\leqslant \left(\frac{1}{V_1^2} - 2\frac{A_1}{V_1}\right)\|\mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{x}/A_1) - \mathrm{Prox}_{\frac{1}{A_1}f}(\mathbf{y}/A_1)\|_2^2 + \|\mathbf{x}-\mathbf{y}\|_2^2 && \text{(C.26)}\\
&= \left(\left(\frac{1}{V_1^2} - 2\frac{A_1}{V_1}\right)\left(\frac{1}{A_1+\sigma_1}\right)^2 + 1\right)\|\mathbf{x}-\mathbf{y}\|_2^2 && \text{(C.27)}\\
&= \left(\frac{A_2^2 - A_1^2}{(A_1+\sigma_1)^2} + 1\right)\|\mathbf{x}-\mathbf{y}\|_2^2. && \text{(C.28)}
\end{aligned}$$

The upper bound on the Lipschitz constant is therefore:

$$\mathcal{L}_1 = \sqrt{\frac{(A_2^2 - A_1^2)}{(A_1+\sigma_1)^2} + 1}. \tag{C.29}$$

Case 3: no strong convexity or smoothness assumption This setting will not be needed for our proof (because we will only handle penalty functions which have a strictly positive strong convexity constant, by adding a ridge term), but we still go through it for completeness. In this case the only information we have is the firm nonexpansiveness of the proximal operator, which gives the same proof as in the previous case but stops at (D.45), immediately giving the upper bound:

$$\mathcal{L}_1 = \max\left(1, \frac{A_1}{A_2}\right). \tag{C.30}$$

C.3.2 Lipschitz constant of $\mathcal{O}_2$

We have assumed that the data matrix is non-trivial, implying $\lambda_{max}(\mathbf{F}^T\mathbf{F}) \neq 0$. In this case we use the explicit form of $\mathcal{O}_2$, which is linear:

$$\begin{aligned}
\|\mathcal{O}_2(\mathbf{x}) - \mathcal{O}_2(\mathbf{y})\|_2 &= \|\left(\frac{1}{V_1}(\mathbf{F}^T\mathbf{F} + A_2\mathrm{Id})^{-1} - \mathrm{Id}\right)(\mathbf{x}-\mathbf{y})\|_2 && \text{(C.31)}\\
&\leqslant \|\left(\frac{1}{V_1}(\mathbf{F}^T\mathbf{F} + A_2\mathrm{Id})^{-1} - \mathrm{Id}\right)\|\,\|\mathbf{x}-\mathbf{y}\|_2. && \text{(C.32)}
\end{aligned}$$

The spectral norm of matrix $\frac{1}{V_1}(\mathbf{F}^T\mathbf{F} + A_2\mathrm{Id})^{-1} - \mathrm{Id}$ gives the upper bound on the Lipschitz constant:

$$\mathcal{L}_2 = \max\left(\frac{|A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})|}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1|}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})}\right). \tag{C.33}$$

C.4 State evolution equations for the elastic net problem

This section provides some technical details for our minimization problem in the elastic net setting, i.e.

$$\hat{\mathbf{x}}^*_{\boldsymbol{\lambda_2}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\left\{\frac{1}{2}\|\mathbf{y}-\mathbf{F}\mathbf{x}\|_2^2 + \lambda_1\|\mathbf{x}\|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2\right\}. \tag{C.34}$$

We first give the definitions of the proximal operators associated with the separable ℓ_1 norm, the ℓ_2 norm, and the elastic net regularization through their element-wise expression:

$$\mathrm{Prox}_{\lambda_1\|\cdot\|_1}(x) = \begin{cases} x+\lambda_1 & \text{if } x < -\lambda_1 \\ 0 & \text{if } -\lambda_1 < x < \lambda_1 \\ x-\lambda_1 & \text{if } x > \lambda_1 \end{cases} \tag{C.35}$$

which is called the soft thresholding function.

$$\mathrm{Prox}_{\frac{\lambda_2}{2}\|\cdot\|_2^2} = \frac{1}{1+\lambda_2} \tag{C.36}$$

$$\mathrm{Prox}_{\lambda_1\|\cdot\|_1+\frac{\lambda_2}{2}\|\cdot\|_2^2} = \frac{1}{1+\lambda_2}\mathrm{Prox}_{\lambda_1\|\cdot\|_1}. \tag{C.37}$$

For our numerical simulations, we consider an i.i.d. teacher vector $\mathbf{x}_0$ pulled from the Gauss-Bernoulli distribution :

$$p_{x_0}(x_0) = (1-\rho)\delta(x_0) + \rho\frac{1}{\sqrt{2\pi}}\exp\left(-x_0^2/2\right). \tag{C.38}$$

To run Oracle VAMP, we first had to determine the constants coming from the fixed point of state evolution, hence we had to solve SE equations first. Let us give some detail about the set of equations (3.20), specifically for the elastic net minimization problem. We need to specify all quantities that depend on the proximal operators.

$$\alpha_1^{(t)} = \mathbb{E}\left[\mathrm{Prox}'_{f/A_1^{(t)}}(x_0+P_1^{(t)})\right] \tag{C.39}$$

where the expectation is over $x_0 \sim p_{x_0}$ and $P_1 \sim \mathcal{N}(0,\tau_1^{(t)})$, can be explicitly computed and yields

$$\alpha_1^{(t)} = \frac{1}{1+\frac{\lambda_2}{A_1^{(t)}}}\left[(1-\rho)\,\mathrm{erfc}\left(\frac{\lambda_1}{A_1^{(t)}\sqrt{2\tau_1^{(t)}}}\right) + \rho\,\mathrm{erfc}\left(\frac{\lambda_1}{A_1^{(t)}\sqrt{2(\tau_1^{(t)}+1)}}\right)\right] \tag{C.40}$$

with $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{+\infty} e^{-t^2}\,\mathrm{d}t$. We now turn to the error

$$\mathcal{E}_1(A_1^{(t)},\tau_1^{(t)}) = \mathbb{E}\left[\left(\mathrm{Prox}_{f/A_1^{(t)}}(x_0+P_1^{(t)}) - x_0\right)^2\right]. \tag{C.41}$$

This term can be tricky to evaluate numerically, as it involves two-dimensional integrals. Some algebra allows to write it in terms of one-dimensional integrals and error functions, which are supported by most scientific coding library:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t.$$

We denote $s = \left(1 + \frac{\lambda_2}{A_1^{(t)}}\right)^{-1}$, and directly skip to the result:

$$\begin{aligned}
\mathcal{E}_1 = (1-\rho)s^2 &\left[\operatorname{erfc}\left(\frac{\lambda_1/A_1^{(t)}}{\sqrt{2\tau_1^{(t)}}}\right)\left(\left(\frac{\lambda_1}{A_1^{(t)}}\right)^2 + \tau_1^{(t)}\right) - \exp\left\{-\frac{(\lambda_1/A_1^{(t)})^2}{2\tau_1^{(t)}}\right\}\sqrt{2\tau_1^{(t)}/\pi}\,\frac{\lambda_1}{A_1^{(t)}}\right] \\
&+ \rho\mathbb{E}_{x_0}\left[\frac{1}{2}x_0^2\left\{\operatorname{erf}\left(\frac{\lambda_1/A_1^{(t)} - x_0}{\sqrt{2\tau_1^{(t)}}}\right) + \operatorname{erf}\left(\frac{\lambda_1/A_1^{(t)} + x_0}{\sqrt{2\tau_1^{(t)}}}\right)\right\} + x_0^2 - 2sx_0^2 + s^2\left[\tau_1^{(t)} + (\lambda_1/A_1^{(t)})^2 + x_0^2\right]\right. \\
&\quad + s\sqrt{\tau_1^{(t)}/(2\pi)}\exp\left\{\left(-\frac{\lambda_1/A_1^{(t)} - x_0^2}{2\tau_1^{(t)}}\left((s-2)x_0 - s\frac{\lambda_1}{A_1^{(t)}}\right)\right)\right\} \\
&\quad + s\sqrt{\tau_1^{(t)}/(2\pi)}\exp\left\{\left(-\frac{\lambda_1/A_1^{(t)} + x_0^2}{2\tau_1^{(t)}}\left((2-s)x_0 - s\frac{\lambda_1}{A_1^{(t)}}\right)\right)\right\} \\
&\quad + \frac{1}{2}\left[s^2(\tau_1^{(t)} + (\lambda_1/A_1^{(t)} - x_0)^2) + 2s(\lambda_1/A_1^{(t)} - x_0)x_0 + x_0^2\right]\operatorname{erf}\left(\frac{\lambda_1/A_1^{(t)} - x_0}{\sqrt{2\tau_1^{(t)}}}\right) \\
&\quad \left. - \frac{1}{2}\left[x_0^2 - 2sx_0(\lambda_1/A_1^{(t)} + x_0) + s^2(\tau_1^{(t)} + (\lambda_1/A_1^{(t)} + x_0)^2)\right]\operatorname{erf}\left(\frac{\lambda_1/A_1^{(t)} + x_0}{\sqrt{2\tau_1^{(t)}}}\right)\right].
\end{aligned} \tag{C.42}$$

Using expression (C.40) and (C.42) allows to numerically solve the state evolution equations for the elastic net problem.

Appendix D

Details for Oracle MLVAMP convergence proof

D.1 From replica potentials to Moreau envelopes

Here we show how the potentials defined for the replica free energy (4.5) can be mapped to Moreau envelopes in the $\beta \to \infty$ limit. We consider the scalar case since the replica expressions are scalar. All functions are separable here, so any needed generalization to the multidimensional case is immediate. We start by reminding the definition of the Moreau envelope [18, 122] $\mathcal{M}_{\gamma f}$ of a proper, closed and convex function f for a given $\gamma \in \mathbb{R}_+^*$ and any $z \in \mathbb{R}$:

$$\mathcal{M}_{\gamma f}(z) = \inf_{x \in \mathbb{R}} \left\{ f(x) + (1/2\gamma)\|x - z\|_2^2 \right\} \tag{D.1}$$

The Moreau envelope can be interpreted as a smoothed version of a given objective function with the same minimizer. For ℓ_1 minimization for example, it allows to work with a differentiable objective. By definition of the proximal operator we have the following identity:

$$\text{Prox}_{\gamma f}(z) = \arg\min_{x \in \mathbb{R}} \left\{ f(x) + (1/2\gamma)\|x - z\|_2^2 \right\}, \tag{D.2}$$

$$\mathcal{M}_{\gamma f}(z) = f(\text{Prox}_{\gamma f}(z)) + \frac{1}{2}\|\text{Prox}_{\gamma f}(z) - z\|_2^2. \tag{D.3}$$

Let us see how to match the replica potentials with the Moreau envelope. We start from the definition of ϕ_x, and apply Laplace's approximation:

$$\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) = \lim_{\beta \to \infty} \frac{1}{\beta} \log \int e^{-\frac{\beta \hat{Q}_{1x}}{2} x^2 + \beta(\hat{m}_{1x} x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - \beta f(x)} dx \tag{D.4}$$

$$= -\frac{\hat{Q}_{1x}}{2}(x^*)^2 + (\hat{m}_{1x} x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x^* - f(x^*) \tag{D.5}$$

where

$$x^* = \arg\min_{x \in \mathbb{R}} \left\{ -\frac{\hat{Q}_{1x}}{2} x^2 + (\hat{m}_{1x} x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - f(x) \right\}. \tag{D.6}$$

This is an unconstraint convex optimization problem, thus its optimality condition is enough to characterize its unique minimizer:

$$-\hat{Q}_{1x} x^* + (\hat{m}_{1x} x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}) - \partial f(x^*) = 0 \tag{D.7}$$

$$\iff x^* = (\text{Id} + \frac{1}{\hat{Q}_{1x}}\partial f)^{-1} \left(\frac{\hat{m}_{1x} x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}} \right) \tag{D.8}$$

$$\iff x^* = \text{Prox}_{f/\hat{Q}_{1x}} \left(\frac{\hat{m}_{1x} x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}} \right) \tag{D.9}$$

Replacing this in the replica potential and completing the square, we get:

$$\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) = -f(\mathrm{Prox}_{f/\hat{Q}_{1x}}(X)) - \frac{\hat{Q}_{1x}}{2}\|X - \mathrm{Prox}_{f/\hat{Q}_{1x}}(X)\|_2^2 + \frac{X^2}{2}\hat{Q}_{1x} \quad \text{(D.10)}$$

$$= \hat{Q}_{1x}\frac{X^2}{2} - \mathcal{M}_{f/\hat{Q}_{1x}}(X) \quad \text{(D.11)}$$

where we used the shorthand $X = \frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}$. The same calculation provides a similar expression for ϕ_y.

D.2 Rigorous state evolution statement

D.2.1 Making assumption (4.26) rigorous

We look into the state evolution equations derived for MLVAMP in [132], that we call (SE1). Those equations are proven to be exact in the asymptotic limit, and follow the same algorithm as (4.11). In particular, they provide statistical properties of vectors $\mathbf{h}_{1x}, \mathbf{h}_{2x}, \mathbf{h}_{1z}, \mathbf{h}_{2z}$. We would like to check whether the starting assumption (4.26), that serves as building block of the state evolution derived in [152] (SE2), are rigorous too. We can read relations from [54] using the following dictionary between our notations and theirs, valid at each iteration of the algorithm:

$$\hat{Q}_{1x}, \hat{Q}_{2x}, \hat{Q}_{1z}, \hat{Q}_{2z} \longleftrightarrow \gamma_0^-, \gamma_0^+, \gamma_1^+, \gamma_1^- \quad \text{(D.12a)}$$

$$\chi_{1x}\hat{Q}_{1x}, \chi_{2x}\hat{Q}_{2x}, \chi_{1z}\hat{Q}_{1z}, \chi_{2z}\hat{Q}_{2z} \longleftrightarrow \alpha_0^-, \alpha_0^+, \alpha_1^-, \alpha_1^+ \quad \text{(D.12b)}$$

$$\mathbf{x_0}, \mathbf{z_0}, \rho_x, \rho_z, \mathbf{h}_{1x}, \mathbf{h}_{2x}, \mathbf{h}_{1z}, \mathbf{h}_{2z} \longleftrightarrow \mathbf{Q}_0^0, \mathbf{Q}_1^0, \tau_0^0, \tau_1^0, \mathbf{r}_0^-, \mathbf{r}_0^+, \mathbf{r}_1^+, \mathbf{r}_1^-. \quad \text{(D.12c)}$$

Let us see what (SE1) says about the distribution of vectors $\mathbf{r}_0^-, \mathbf{r}_0^+, \mathbf{r}_1^+, \mathbf{r}_1^-$ (which are precisely the subjects of assumption (4.26)). Placing ourselves in the asymptotic limit, [54] shows the following equalities:

$$\mathbf{r}_0^- = \mathbf{Q}_0^0 + \mathbf{Q}_0^- \quad \text{(D.13a)}$$

$$\mathbf{r}_0^+ = \mathbf{Q}_0^0 + \mathbf{Q}_0^+ \quad \text{(D.13b)}$$

$$\mathbf{r}_1^- = \mathbf{Q}_1^0 + \mathbf{Q}_1^- \quad \text{(D.13c)}$$

$$\mathbf{r}_1^+ = \mathbf{Q}_1^0 + \mathbf{Q}_1^+ \quad \text{(D.13d)}$$

where $\mathbf{Q}_0^- \sim \mathcal{N}(0, \tau_0^-)^N$ and $\mathbf{Q}_1^- \sim \mathcal{N}(0, \tau_1^-)^N$ are i.i.d. Gaussian vectors. $\mathbf{Q}_0^+$, $\mathbf{Q}_1^+$ have the following norms and non-zero correlations with ground-truth vectors $\mathbf{Q}_0^0, \mathbf{Q}_1^0$:

$$\tau_0^+ \equiv \frac{\|\mathbf{Q}_0^+\|_2^2}{N} \qquad c_0^+ \equiv \frac{\mathbf{Q}_0^{0T}\mathbf{Q}_0^+}{N} \quad \text{(D.14)}$$

$$\tau_1^+ \equiv \frac{\|\mathbf{Q}_1^+\|_2^2}{M} \qquad c_1^+ \equiv \frac{\mathbf{Q}_1^{0T}\mathbf{Q}_1^+}{M}. \quad \text{(D.15)}$$

With simple manipulations, we can rewrite (D.13) as:

$$\mathbf{r}_0^- \overset{\mathrm{d}}{=} \mathbf{Q}_0 + \mathbf{Q}_0^- \tag{D.16a}$$

$$\mathbf{V}^T\mathbf{r}_0^+ \overset{\mathrm{d}}{=} \left(1 + \frac{c_0^+}{\tau_0^0}\right)\mathbf{V}^T\mathbf{Q}_0^0 + \mathbf{V}^T\tilde{\mathbf{Q}}_0^+ \tag{D.16b}$$

$$\mathbf{r}_1^- \overset{\mathrm{d}}{=} \mathbf{Q}_1^0 + \mathbf{Q}_1^- \tag{D.16c}$$

$$\mathbf{U}^T\mathbf{r}_1^+ \overset{\mathrm{d}}{=} \left(1 + \frac{c_1^+}{\tau_1^0}\right)\mathbf{U}^T\mathbf{Q}_1^0 + \mathbf{U}^T\tilde{\mathbf{Q}}_1^+ \tag{D.16d}$$

where for $k \in \{1, 2\}$ vectors

$$\tilde{\mathbf{Q}}_k^+ = -\frac{c_k^+}{\tau_k^0}\mathbf{Q}_k^0 + \mathbf{Q}_k^+ \tag{D.17}$$

and $\mathbf{Q}_0^-, \mathbf{Q}_1^-$ have no correlation with ground-truth vectors $\mathbf{Q}_0^0$, $\mathbf{Q}_1^0$, $\mathbf{U}^T\mathbf{Q}_0^0$, $\mathbf{V}^T\mathbf{Q}_1^0$. Besides, Lemma 5 from [132] states that $\mathbf{V}^T\tilde{\mathbf{Q}}_0^+$ and $\mathbf{U}^T\tilde{\mathbf{Q}}_1^+$ have components that converge empirically to Gaussian variables, respectively $\mathcal{N}(0, \tau_0^+)$ and $\mathcal{N}(0, \tau_1^+)$. Let us now translate this in our own terms, using the following relations that complete our dictionary with state evolution parameters:

$$\frac{\hat{m}_{1x}}{\hat{Q}_{1x}} \longleftrightarrow 1 \qquad \frac{\hat{m}_{2z}}{\hat{Q}_{2z}} \longleftrightarrow 1 \tag{D.18a}$$

$$\frac{\hat{m}_{2x}}{\hat{Q}_{2x}} \longleftrightarrow 1 + \frac{c_0^+}{\tau_0^0} \qquad \frac{\hat{m}_{1z}}{\hat{Q}_{1z}} \longleftrightarrow 1 + \frac{c_1^+}{\tau_1^0} \tag{D.18b}$$

$$\frac{\hat{\chi}_{1x}}{\hat{Q}_{1x}^2} \longleftrightarrow \tau_0^- \qquad \frac{\hat{\chi}_{2z}}{\hat{Q}_{2z}^2} \longleftrightarrow \tau_1^- \tag{D.18c}$$

$$\frac{\hat{\chi}_{2x}}{\hat{Q}_{2x}^2} \longleftrightarrow \tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0} \qquad \frac{\hat{\chi}_{1z}}{\hat{Q}_{1z}^2} \longleftrightarrow \tau_1^+ - \frac{(c_1^+)^2}{\tau_1^0}. \tag{D.18d}$$

Simple bookkeeping transforms equations (D.16) into a rigorous statement of starting assumptions (4.26) from [152]. Since those assumptions are now rigorously established in the asymptotic limit, the remaining derivation of state evolution equations (4.27) holds and provides a mathematically exact statement.

D.2.2 Scalar equivalent model of state evolution

For the sake of completeness, we will provide an overview of the explicit matching between the state evolution formalism from [54] which was developed in a series of papers, and the replica formulation from [152] which relies on statistical physics methods. Although not necessary to our proof, it is interesting to develop an intuition about the correspondence between those two faces of the same coin. We have seen in the previous subsection that [54] introduces ground-truth vectors $\mathbf{Q}_0^0, \mathbf{Q}_1^0$, estimates $\mathbf{r}_0^\pm, \mathbf{r}_1^\pm$ which are related to vectors $\mathbf{Q}_0^\pm, \mathbf{Q}_1^\pm$. Let us introduce a few more vectors using matrices from the singular value decomposition $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Let $\mathbf{s}_\nu \in \mathbb{R}^N$ be the vector containing all square roots of eigenvalues of $\mathbf{F}^T\mathbf{F}$ with p_ν its element-wise distribution; and $\mathbf{s}_\mu \in \mathbb{R}^M$ the vector containing all square roots of eigenvalues of $\mathbf{F}\mathbf{F}^T$ with p_μ its element-wise distribution. Note that those two vectors contain the singular values of $\mathbf{F}$, but one of them also contains $\max(M, N) - \min(M, N)$ zero values. p_μ and p_ν are both well-defined

since p_λ is properly defined in Assumptions 4.1. We also define

$$\begin{aligned} \mathbf{P}_0^0 &= \mathbf{V}^T\mathbf{Q}_0^0 & \mathbf{P}_0^+ &= \mathbf{V}^T\mathbf{Q}_0^+ & \mathbf{P}_0^- &= \mathbf{V}^T\mathbf{Q}_0^- \\ \mathbf{P}_1^0 &= \mathbf{U}\mathbf{Q}_1^0 & \mathbf{P}_1^+ &= \mathbf{U}\mathbf{Q}_1^+ & \mathbf{P}_1^- &= \mathbf{U}\mathbf{Q}_1^-. \end{aligned}$$

By virtue of Lemma 5 from [132], the six previous vectors have elements that converge empirically to a Gaussian variable. Hence, all defined vectors have an element-wise separable distribution, and we can write the state evolution as a scalar model on random variables sampled from those distributions. To do so, we will simply write the variables without the bold font: for instance $Z_0^0 \sim p_{x_0}$, $s_\nu \sim p_\nu$, and Q_0^- refers to the random variable distributed according to the element-wise distribution of vector $\mathbf{Q}_0^-$. The scalar random variable state evolution from [54] now reads:

Initialize $\gamma_1^{-(0)}, \gamma_0^{-(0)}, \tau_0^{-(0)}, \tau_1^{-(0)}, Q_0^{-(0)} \sim \mathcal{N}(0, \tau_0^{-(0)}), Q_1^{-(0)} \sim \mathcal{N}(0, \tau_1^{-(0)}), \alpha_0^{-(0)}, \alpha_1^{-(0)}$

Initial pass (ground truth only)

$$s_\nu \sim p_\nu, \qquad s_\mu \sim p_\mu, \qquad Q_0^0 \sim p_{x_0} \tag{D.19a}$$

$$\tau_0^0 = \mathbb{E}[(Q_0^0)^2] \qquad P_0^0 \sim \mathcal{N}(0, \tau_0^0) \tag{D.19b}$$

$$Q_1^0 = s_\mu P_0^0 \qquad \tau_1^0 = \mathbb{E}[(s_\mu P_0^0)^2] = \mathbb{E}[(s_\mu)^2]\tau_0^0 \qquad P_1^0 \sim \mathcal{N}(0, \tau_1^0) \tag{D.19c}$$

Forward Pass (estimation):

$$\alpha_0^{+(t)} = \mathbb{E}\left[\eta'_{f/\gamma_0^{-(t)}}(Q_0^0 + Q_0^{-(t)})\right] \tag{D.19d}$$

$$\gamma_0^{+(t)} = \frac{\gamma_0^{(t)}}{\alpha_0^{+(t)}} - \gamma_0^{-(t)} \tag{D.19e}$$

$$Q_0^{+(t)} = \frac{1}{1-\alpha_0^{+(t)}}\left\{\eta_{f/\gamma_0^{-(t)}}(Q_0^0 + Q_0^{-(t)}) - Q_0^0 - \alpha_0^+ Q_0^{-(t)}\right\} \tag{D.19f}$$

$$\mathbf{K}_0^{+(t)} = Cov\left(Q_0^0, Q_0^{+(t)}\right) \qquad \left(P_0^0, P_0^{+(t)}\right) \sim \mathcal{N}\left(0, \mathbf{K}_0^{+(t)}\right) \tag{D.19g}$$

$$\alpha_1^{+(t)} = \mathbb{E}\left[\frac{s_\mu^2 \gamma_1^{-(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}\right] \tag{D.19h}$$

$$\gamma_1^{+(t)} = \frac{\gamma_1^{-(t)}}{\alpha_1^{+(t)}} - \gamma_1^{-(t)} \tag{D.19i}$$

$$\begin{aligned} Q_1^{+(t)} = \frac{1}{1-\alpha_1^{+(t)}}\Bigg\{&\frac{s_\mu^2 \gamma_1^{-(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}(Q_1^{-(t)} + Q_1^0) \\ &+ \frac{s_\mu \gamma_0^{+(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}(P_0^{+(t)} + P_0^0) - Q_1^0 - \alpha_1^{+(t)} Q_1^{-(t)}\Bigg\} \end{aligned} \tag{D.19j}$$

$$\mathbf{K}_1^{+(t)} = Cov\left(Q_1^0, Q_1^{+(t)}\right) \qquad \left(P_1^0, P_1^{+(t)}\right) \sim \mathcal{N}\left(0, \mathbf{K}_1^{+(t)}\right) \tag{D.19k}$$

Backward Pass (estimation):

$$\alpha_1^{-(t+1)} = \mathbb{E}\left[\eta_{g(y,.)/\gamma_1^{+(t)}}(P_1^0 + P_1^{+(t)})\right] \tag{D.19l}$$

$$\gamma_1^{-(t+1)} = \frac{\gamma_1^{+(t)}}{\alpha_1^{-(t+1)}} - \gamma_1^{+(t)} \tag{D.19m}$$

$$P_1^{-(t+1)} = \frac{1}{1-\alpha_1^{-(t+1)}}\left\{\eta_{g(y,.)/\gamma_1^{+(t)}}(P_1^0 + P_1^{+(t)}) - P_1^0 - \alpha_1^{-(t+1)} P_1^{+(t)}\right\} \tag{D.19n}$$

$$\tau_1^{-(t+1)} = \mathbb{E}\left[(P_1^{-(t+1)})^2\right] \qquad Q_1^{-(t+1)} \sim \mathcal{N}(0, \tau_1^{-(t+1)}) \tag{D.19o}$$

$$\alpha_0^{-(t+1)} = \mathbb{E}\left[\frac{\gamma_0^{+(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}\right] \tag{D.19p}$$

$$\gamma_0^{-(t+1)} = \frac{\gamma_0^{+(t)}}{\alpha_0^{-(t+1)}} - \gamma_0^{+(t)} \tag{D.19q}$$

$$\begin{aligned} P_0^{-(t+1)} = \frac{1}{1-\alpha_0^{-(t+1)}}\Bigg\{ & \frac{s_\nu \gamma_1^{-(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}(Q_1^{-(t+1)} + Q_1^0) \\ & + \frac{\gamma_0^{+(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}(P_0^{+(t)} + P_0^0) - P_0^0 - \alpha_0^{-(t+1)} P_0^{+(t)}\Bigg\} \end{aligned} \tag{D.19r}$$

$$\tau_0^{-(t+1)} = \mathbb{E}\left[(P_0^{-(t+1)})^2\right] \qquad Q_0^{-(t+1)} \sim \mathcal{N}(0, \tau_0^{-(t+1)}). \tag{D.19s}$$

D.2.3 Direct matching of the state evolution fixed point equations

To be consistent, we should be able to show that equations (D.19) allow us to recover equations (4.27) at their fixed point. Although somewhat tedious, this task is facilitated using dictionaries (D.12) and (D.18). We shall give here an overview of this matching through a few examples.

- Recovering equation (4.27e)

Let us start from the rigorous scalar state evolution, in particular equation (D.19f) that defines variable Q_0^+. We get rid of time indices here since we focus on the fixed point. We first compute the correlation

$$c_0^+ = \mathbb{E}\left[Q_0^0 Q_0^+\right] = \frac{1}{1-\alpha_0^+}\left\{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right] - \tau_0^0\right\} \tag{D.20}$$

where we have used $\mathbb{E}[(Q_0^0)^2] = \tau_0^0$. At the fixed point, we know from MLVAMP or simply translating equations (4.17), (4.18) that

$$1 - \alpha_0^+ = \alpha_0^-, \qquad \frac{1}{\alpha_0^-} = \frac{\gamma_0^- + \gamma_0^+}{\gamma_0^+}, \qquad \gamma_0^+ \alpha_0^+ = \gamma_0^- \alpha_0^-.$$

Simple manipulations take us to

$$c_0^+ = \frac{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right]}{\alpha_0^-} - \tau_0^0\left(1 + \frac{\gamma_0^-}{\gamma_0^+}\right) \tag{D.21}$$

$$\left(1 + \frac{c_0^+}{\tau_0^0}\right)\gamma_0^+ = \frac{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right]\gamma_0^+}{\tau_0^0 \alpha_0^-} - \gamma_0^-. \tag{D.22}$$

Now let us translate this back into our notations. The term $\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]$ simply translates into m_{1x}, and the rest of the terms can all be changed according to our dictionary. (D.22) exactly becomes

$$\hat{m}_{2x} = \frac{m_{1x}}{\rho_x \chi_x} - \hat{m}_{1x}, \tag{D.23}$$

hence we perfectly recover equations (4.27e) at the fixed point.

- Recovering equation (4.27f)

We start again from (D.19f) and square it:

$$\begin{aligned}\mathbb{E}\left[(Q_0^+)^2\right] = \frac{1}{(1-\alpha_0^+)^2}\Big\{&\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] + \mathbb{E}\left[(Q_0^0)^2\right] + (\alpha_0^+)^2\mathbb{E}\left[(Q_0^-)^2\right] \\ &-2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] - 2\alpha_0^+\mathbb{E}\left[Q_0^-\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\Big\} \qquad \text{(D.24)}\\ \tau_0^+ = \frac{1}{(1-\alpha_0^+)^2}\Big\{&\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] + \tau_0^0 + (\alpha_0^+)^2\tau_0^- \\ &-2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] - 2\alpha_0^+\mathbb{E}\left[Q_0^-\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\Big\}. \qquad \text{(D.25)}\end{aligned}$$

Since Q_0^- is a Gaussian variable, independent from Q_0^0, we can use Stein's lemma and use equation (D.19d) to get

$$\mathbb{E}\left[Q_0^-\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] = \alpha_0^+\tau_0^-. \tag{D.26}$$

Moreover, from (D.20) we have

$$(c_0^+)^2(\alpha_0^-)^2 = \left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] - \tau_0^0\right)^2 \tag{D.27}$$

$$\frac{(c_0^+)^2(\alpha_0^-)^2}{\tau_0^0} - \frac{(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right])^2}{\tau_0^0} = -2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] + \tau_0^0. \tag{D.28}$$

Replacing (D.26) and (D.28) into (D.25), we reach

$$\left(\tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0}\right)(\alpha_0^-)^2 = \mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] - \frac{\left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\right)^2}{\tau_0^0} - (\alpha_0^+)^2\tau_0^- \tag{D.29}$$

$$\begin{aligned}\left(\tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0}\right)(\gamma_0^+)^2 = &\frac{\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right](\gamma_0^+)^2}{(\alpha_0^-)^2} - \frac{\left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\right)^2(\gamma_0^+)^2}{\tau_0^0(\alpha_0^-)^2} \\ &- (\gamma_0^-)^2\tau_0^-. \qquad \text{(D.30)}\end{aligned}$$

Notice that $\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]$ simply translates into our variable q_{1x} from its definition (4.27c), and our dictionary directly transforms (D.30) into equation (4.27f):

$$\hat{\chi}_{2x} = \frac{q_{1x}}{\chi_{1x}^2} - \frac{m_{1x}^2}{\rho_x\chi_{1x}^2} - \hat{\chi}_{1x}. \tag{D.31}$$

- Recovering equation (4.27s)

We first note that for any function h,

$$\mathbb{E}[h(s_\nu)] = \min(1,\alpha)\mathbb{E}[h(s_\mu)] + \max(0,1-\alpha)h(0). \tag{D.32}$$

and $s_\nu^2 \sim p_\lambda$. Applying this to $h(s) = \dfrac{\gamma_1^- s^2}{\gamma_1^- s^2 + \gamma_0^+}$ and starting from (D.19j), we rewrite

$$\alpha_1^+ = \mathbb{E}\left[\frac{\gamma_1^- s_\mu^2}{\gamma_1^- s_\mu^2 + \gamma_0^+}\right] \tag{D.33}$$

$$= \frac{1}{\alpha}\mathbb{E}\left[\frac{\gamma_1^- \lambda}{\gamma_1^- \lambda + \gamma_0^+}\right] \tag{D.34}$$

with $\lambda \sim p_\lambda$, which translates into equation (4.27s):

$$\chi_{2z} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x} + \lambda \hat{Q}_{2z}}\right]. \tag{D.35}$$

In a similar fashion, we can recover all equations (4.27) by writing variances and correlations between scalar random variables defined in (D.19), and using the independence properties established in [54]; thus directly showing the matching between the two state evolution formalisms at their fixed point.

D.3 Operator norms and Lipschitz constants

D.3.1 Operator norms of the matrices $\mathbf{W_1}, \mathbf{W_2}, \mathbf{W_3}, \mathbf{W_4}$

The norms of the linear operators $\mathbf{W_1}, \mathbf{W_2}, \mathbf{W_3}, \mathbf{W_4}$ defined in (4.31) can be computed or bounded with respect to the singular values of the matrix $\mathbf{F}$. The derivations are straightforward and do not require any specific mathematical result. Denoting $\|\mathbf{W}\|$ the operator norm of a given matrix $\mathbf{W}$, we have the following:

$$\|W_1\| = \frac{\hat{Q}_{2x}}{\hat{Q}_{1x}}\left(\frac{|\hat{Q}_{1x} - \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\hat{Q}_{1x} - \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})}\right) \tag{D.36}$$

$$\|W_2\| = \frac{\hat{Q}_{2z}}{\chi_x \hat{Q}_{1x}} \frac{\sqrt{\lambda_{max}(\mathbf{F}^T\mathbf{F})}}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})} \tag{D.37}$$

$$\|W_3\| = \frac{\hat{Q}_{2z}}{\hat{Q}_{1z}}\left(\frac{|\hat{Q}_{2x} - \hat{Q}_{1z}\lambda_{min}(\mathbf{F}\mathbf{F}^T)|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}\mathbf{F}^T)}, \frac{|\hat{Q}_{2x} - \hat{Q}_{1z}\lambda_{max}(\mathbf{F}\mathbf{F}^T)|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{max}(\mathbf{F}\mathbf{F}^T)}\right) \tag{D.38}$$

$$\|W_4\| = \frac{\hat{Q}_{2x}}{\chi_z \hat{Q}_{1z}} \frac{\sqrt{\lambda_{max}(\mathbf{F}^T\mathbf{F})}}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})}. \tag{D.39}$$

D.3.2 Lispchitz constants of $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$

We now derive upper bounds of the Lipschitz constants of $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$ using properties of proximal operators stated in appendix C.3. We lay out some details for $\tilde{\mathcal{O}}_1$, the derivation is identical for $\tilde{\mathcal{O}}_2$. Let $(\sigma_1, \beta_1) \in \mathbb{R}_+^{*2}$ be the strong-convexity and smoothness constants of f. Note that (4.48) states that $\sigma_1 \leqslant \hat{Q}_{2x} \leqslant \beta_1$.

Case 1: $0 < \sigma_1 < \beta_1$ Proposition 2 from [64] gives the following expression:

$$\mathrm{Prox}_{f/\hat{Q}_{1x}} = \frac{1}{2}\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}} + \frac{1}{1+\beta_1/\hat{Q}_{1x}}\right)\mathrm{Id} + \frac{1}{2}\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}} - \frac{1}{1+\beta_1/\hat{Q}_{1x}}\right)S_1 \tag{D.40}$$

where S_1 is a nonexpansive operator. Replacing in the expression of $\tilde{\mathcal{O}}_1$ leads to:

$$\tilde{\mathcal{O}}_1 = \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\left(\left(\frac{1}{2\chi_x}\left(\frac{1}{\hat{Q}_{1x}+\sigma_1}+\frac{1}{\hat{Q}_{1x}+\beta_1}\right)-1\right)\mathrm{Id}+\frac{1}{2\chi_x}\left(\frac{1}{\hat{Q}_{1x}+\sigma_1}-\frac{1}{\hat{Q}_{1x}+\beta_1}\right)S_1\right) \tag{D.41}$$

which, knowing that $\hat{Q}_{1x}+\hat{Q}_{2x}=\frac{1}{\chi_x}$ at the fixed point, and splitting cases where the first term of the sum in D.41 is negative or positive, provides $\tilde{\mathcal{O}}_1$'s Lipschitz constant:

$$\omega_1 = \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\max\left(\frac{\hat{Q}_{2x}-\sigma_1}{\hat{Q}_{1x}+\sigma_1}, \frac{\beta_1-\hat{Q}_{2x}}{\hat{Q}_{1x}+\beta_1}\right). \tag{D.42}$$

Case 2: $0<\sigma_1=\beta_1$ In this case, Proposition 2 from says that [64]

$$\|\mathrm{Prox}_{f/\hat{Q}_{1x}}(x)-\mathrm{Prox}_{f/\hat{Q}_{1x}}(y)\|_2^2 = \left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}}\right)^2\|x-y\|_2^2 \tag{D.43}$$

which, with the firm non-expansiveness of the proximal operator gives, for any $x,y\in\mathbb{R}$:

$$\begin{aligned}
\|\tilde{\mathcal{O}}_1(x)-\tilde{\mathcal{O}}_1(y)\|_2^2 &= \left(\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\right)^2\left(\frac{1}{\hat{Q}_{1x}^2\chi_x^2}\|\mathrm{Prox}_{f/\hat{Q}_{1x}}(x)-\mathrm{Prox}_{f/\hat{Q}_{1x}}(y)\|_2^2\right.\\
&\qquad\left.-2\frac{1}{\chi_x}\left\langle x-y, \mathrm{Prox}_{f/\hat{Q}_{1x}}(x)-\mathrm{Prox}_{f/\hat{Q}_{1x}}(y)\right\rangle+\|x-y\|_2^2\right) &\text{(D.44)}\\
&\leqslant\left(\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\right)^2\left(\left(\frac{1}{\hat{Q}_{1x}^2\chi_x^2}-2\frac{1}{\chi_x}\right)\|\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}}f}(x)-\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}}f}(y)\|_2^2+\|x-y\|_2^2\right)\\
&=\left(\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\right)^2\left(\left(\frac{1}{\hat{Q}_{1x}^2\chi_x^2}-2\frac{1}{\chi_x}\right)\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}}\right)^2+1\right)\|x-y\|_2^2 &\text{(D.45)}\\
&=\left(\frac{\hat{Q}_{2x}^2-\hat{Q}_{1x}^2}{(\hat{Q}_{1x}+\sigma_1)^2}+1\right)\|x-y\|_2^2. &\text{(D.46)}
\end{aligned}$$

The upper bound on the Lipschitz constant is therefore:

$$\omega_1 = \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\sqrt{1+\frac{(\hat{Q}_{2x}^2-\hat{Q}_{1x}^2)}{(\hat{Q}_{1x}+\sigma_1)^2}}. \tag{D.47}$$

Recovering (4.42) In our proof, we make no assumption on the strong-convexity or smoothness of the function f, but adding the ridge penalties through parameters $\lambda_2,\tilde{\lambda}_2$ brings both $\tilde{\mathcal{O}}_1$ and $\tilde{\mathcal{O}}_2$ to either the first of the second case above. It is straightforward to see that the Lipschitz constant (D.47) is a upper bound of (D.42). We thus use (D.47) for generality, and recover expressions (4.42):

$$\omega_1 = \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\sqrt{1+\frac{\hat{Q}_{2x}^2-\hat{Q}_{1x}^2}{(\hat{Q}_{1x}+\lambda_2)^2}} \tag{D.48}$$

$$\omega_2 = \frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\sqrt{1+\frac{\hat{Q}_{2z}^2-\hat{Q}_{1z}^2}{(\hat{Q}_{1z}+\tilde{\lambda}_2)^2}}. \tag{D.49}$$

D.4 Dynamical system convergence analysis

In this section, we study the convergence of the linear recast of Oracle MLVAMP, by trying to satisfy the linear matrix inequality (LMI) (4.44). We focus on the smoothed problem (4.52). in particular, we will look at the regime where the additional regularization λ_2 is arbitrarily large, while $\tilde{\lambda}_2$ is fixed but non-zero. From bounds (4.48) and (4.49); we see that $\hat{Q}_{2x}, \hat{Q}_{1z}$ will grow with λ_2,; while $\hat{Q}_{2z}, \hat{Q}_{1x}$ remain finite. We write the corresponding linear matrix inequality (4.44) and expand the constraint term:

$$0 \succeq \begin{bmatrix} \mathbf{A}^T\mathbf{P}\mathbf{A} - \tau^2\mathbf{P} & \mathbf{A}^T\mathbf{P}\mathbf{B} \\ \mathbf{B}^T\mathbf{P}\mathbf{A} & \mathbf{B}^T\mathbf{P}\mathbf{B} \end{bmatrix} + \begin{bmatrix} \beta_0\mathbf{C}_0^T\mathbf{M}_0\mathbf{C}_0 + \beta_1\mathbf{C}_1^T\mathbf{M}_1\mathbf{C}_1 & \beta_0\mathbf{C}_0^T\mathbf{M}_0\mathbf{D}_0 + \beta_1\mathbf{C}_1^T\mathbf{M}_1\mathbf{D}_1 \\ \beta_0\mathbf{D}_0^T\mathbf{M}_0\mathbf{C}_0 + \beta_1\mathbf{D}_1^T\mathbf{M}_1\mathbf{C}_1 & \beta_0\mathbf{D}_0^T\mathbf{M}_0\mathbf{D}_0 + \beta_1\mathbf{D}_1^T\mathbf{M}_1\mathbf{D}_1 \end{bmatrix} \tag{D.50}$$

A little basic algebra shows that:

$$\mathbf{C}_0^T\mathbf{M}_0\mathbf{C}_0 = \begin{bmatrix} \mathbf{0}_{\mathbf{M}\times\mathbf{M}} & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \omega_0^2\mathbf{I}_{\mathbf{N}} \end{bmatrix} \qquad \mathbf{C}_1^T\mathbf{M}_1\mathbf{C}_1 = \begin{bmatrix} \omega_1^2\mathbf{W}_3^T\mathbf{W}_3 & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \mathbf{0}_{\mathbf{N}\times\mathbf{N}} \end{bmatrix} \tag{D.51}$$

$$\mathbf{C}_0^T\mathbf{M}_0\mathbf{D}_0 = \mathbf{0}_{(\mathbf{M}+\mathbf{N})\times(\mathbf{M}+\mathbf{N})} \qquad \mathbf{D}_0^T\mathbf{M}_0\mathbf{C}_0 = \mathbf{0}_{(\mathbf{M}+\mathbf{N})\times(\mathbf{M}+\mathbf{N})} \tag{D.52}$$

$$\mathbf{C}_1^T\mathbf{M}_1\mathbf{D}_1 = \begin{bmatrix} \mathbf{0}_{\mathbf{M}\times\mathbf{M}} & \omega_1^2\mathbf{W}_3^T\mathbf{W}_4 \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \mathbf{0}_{\mathbf{N}\times\mathbf{N}} \end{bmatrix} \qquad \mathbf{D}_1^T\mathbf{M}_1\mathbf{C}_1 = \begin{bmatrix} \mathbf{0}_{\mathbf{M}\times\mathbf{M}} & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \omega_1^2\mathbf{W}_4^T\mathbf{W}_3 & \mathbf{0}_{\mathbf{N}\times\mathbf{N}} \end{bmatrix} \tag{D.53}$$

$$\mathbf{D}_0^T\mathbf{M}_0\mathbf{D}_0 = \begin{bmatrix} \mathbf{0}_{\mathbf{M}\times\mathbf{M}} & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & -\mathbf{I}_{\mathbf{N}} \end{bmatrix} \qquad \mathbf{D}_1^T\mathbf{M}_1\mathbf{D}_1 = \begin{bmatrix} -\mathbf{I}_{\mathbf{M}} & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \omega_1^2\mathbf{W}_4^T\mathbf{W}_4 \end{bmatrix} \tag{D.54}$$

where all the matrices constituting the blocks have been defined in (4.31). This gives the following form for the constraint matrix, which is the last matrix in (D.50):

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2^T & \mathbf{H}_3 \end{bmatrix} \tag{D.55}$$

where

$$\mathbf{H}_1 = \begin{bmatrix} \beta_1\omega_1^2\mathbf{W}_3^T\mathbf{W}_3 & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \beta_0\omega_0^2\mathbf{I}_{\mathbf{N}} \end{bmatrix} \tag{D.56}$$

$$\mathbf{H}_2 = \begin{bmatrix} \mathbf{0}_{\mathbf{M}\times\mathbf{M}} & \beta_1\omega_1^2\mathbf{W}_3^T\mathbf{W}_4 \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \mathbf{0}_{\mathbf{N}\times\mathbf{N}} \end{bmatrix} \tag{D.57}$$

$$\mathbf{H}_3 = \begin{bmatrix} -\beta_1\mathbf{I}_{\mathbf{M}} & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & -\beta_0\mathbf{I}_{\mathbf{N}} + \beta_1\omega_1^2\mathbf{W}_4^T\mathbf{W}_4 \end{bmatrix} \tag{D.58}$$

thus the LMI becomes:

$$0 \succeq \begin{bmatrix} -\tau^2\mathbf{P} + \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2^T & \mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3 \end{bmatrix}. \tag{D.59}$$

We take $\mathbf{P}$ as block diagonal:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \mathbf{P}_2 \end{bmatrix} \tag{D.60}$$

where $\mathbf{P}_1 \in \mathbb{R}^{M\times M}$ and $\mathbf{P}_2 \in \mathbb{R}^{N\times N}$ are positive definite (no zero eigenvalues) and diagonalizable in the same basis as $\mathbf{F}^T\mathbf{F}$, which is also the eigenbasis of $\mathbf{W}_1, \mathbf{W}_3, \mathbf{W}_2^T\mathbf{W}_2, \mathbf{W}_4^T\mathbf{W}_4$. We have:

$$\mathbf{B}^T\mathbf{P}\mathbf{B} = \begin{bmatrix} \mathbf{P}_1 + \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 & \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1 \\ \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 & \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \end{bmatrix}. \tag{D.61}$$

We are then trying find the conditions for the following problem to be feasible with $0 < \tau < 1$:

$$\begin{bmatrix} \tau^2\mathbf{P} - \mathbf{H}_1 & -\mathbf{H}_2 \\ -\mathbf{H}_2^T & -(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \end{bmatrix} \succeq 0 \tag{D.62}$$

Schur's lemma then says that the strict version of (D.62), which we will consider, is equivalent [76] to:

$$-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0 \quad \text{and} \quad \tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0 \tag{D.63}$$

We want to verify these two conditions.

Conditions for $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0$

We want to derive the conditions for:

$$\begin{bmatrix} \beta_1\mathbf{I_M} - \mathbf{P}_1 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 & -\mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1 \\ -\mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 & \beta_0\mathbf{I_N} - \beta_1\omega_1^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \end{bmatrix} \succ 0. \tag{D.64}$$

Applying Schur's lemma again gives the equivalent problem:

$$\beta_0\mathbf{I_N} - \beta_1\omega_1^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \succ 0 \tag{D.65}$$

$$\begin{aligned} &\beta_1\mathbf{I_M} - \mathbf{P}_1 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 \\ &- \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1(\beta_0\mathbf{I_N} - \beta_1\omega_1^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1)^{-1}\mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 \succ 0. \end{aligned} \tag{D.66}$$

We start with (D.65). A sufficient condition for it to hold true is:

$$\beta_0 > \beta_1\omega_1^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) + \lambda_{max}(\mathbf{P}_2)\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1). \tag{D.67}$$

From appendix 4.4.5, we have:

$$\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1) \leqslant \left(\frac{\hat{Q}_{2x}}{\hat{Q}_{1x}}\right)^2 \max\left(\frac{|\hat{Q}_{1x} - \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\hat{Q}_{1x} - \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})}\right)^2 \tag{D.68}$$

$$\leqslant \max\left(\left(1 - \frac{\hat{Q}_{2z}}{\hat{Q}_{1x}}\lambda_{min}(\mathbf{F}^T\mathbf{F})\right)^2, \left(1 - \frac{\hat{Q}_{2z}}{\hat{Q}_{1x}}\lambda_{max}(\mathbf{F}^T\mathbf{F})\right)^2\right) \equiv b_1 \tag{D.69}$$

and

$$\begin{aligned} \omega_1^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) &\leqslant \left(\frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\right)^2\left(1 + \frac{(\hat{Q}_{2z})^2 - (\hat{Q}_{1z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2}\right)\left(\frac{\hat{Q}_{2x}}{\chi_z\hat{Q}_{1z}}\right)^2 \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F})}{(\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F}))^2} \\ &\leqslant \hat{Q}_{1z}\left(2\tilde{\lambda}_2 + \frac{\tilde{\lambda}_2^2}{\hat{Q}_{1z}} + \frac{(\hat{Q}_{2z})^2}{\hat{Q}_{1z}}\right)\left(\frac{\hat{Q}_{1z} + \hat{Q}_{2z}}{\hat{Q}_{2z}(\hat{Q}_{1z} + \tilde{\lambda}_2)}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F}). \end{aligned} \tag{D.70}$$

For arbitrarily large $\hat{Q}_{1z}$, the quantity $\left(2\tilde{\lambda}_2 + \frac{\tilde{\lambda}_2^2}{\hat{Q}_{1z}} + \frac{(\hat{Q}_{2z})^2}{\hat{Q}_{1z}}\right)\left(\frac{\hat{Q}_{1z}+\hat{Q}_{2z}}{\hat{Q}_{2z}(\hat{Q}_{1z}+\tilde{\lambda}_2)}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F})$ is trivially upperly above whatever the value of $\tilde{\lambda}_2, \hat{Q}_{2z}$. Let b_2 be such an upper bound independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$. The sufficient condition for (D.65) to hold thus becomes:

$$\beta_0 > \beta_1\hat{Q}_{1z}b_2 + \lambda_{max}(\mathbf{P}_2)b_1 \tag{D.71}$$

where b_1, b_2 are constants independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$.

We now turn to (D.66). A sufficient condition for it to hold is:

$$\beta_1 > \lambda_{max}(\mathbf{P}_1) + \lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2)\lambda_{max}(\mathbf{P}_2) + \frac{(\lambda_{max}(\mathbf{P}_2))^2\lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2)\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1)}{\beta_0 - \beta_1\omega_1^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) - \lambda_{max}(\mathbf{P}_2)\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1)} \tag{D.72}$$

Note that condition (D.65) ensures that the denominator in (D.72) is non-zero. We then have:

$$\lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2) \leqslant \left(\frac{\hat{Q}_{2z}}{\chi_x\hat{Q}_{1x}}\right)^2 \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F})}{(\hat{Q}_{2x} + \hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})^2} \tag{D.73}$$

$$\leqslant \left(\frac{\hat{Q}_{2z}(1 + \frac{\hat{Q}_{1x}}{\hat{Q}_{2x}})}{\hat{Q}_{1x}}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F}) \tag{D.74}$$

This quantity can be upperly bounded by a constant independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$ for arbitrarily large $\hat{Q}_{2x}$, that we call b_3 be such a constant. Then a sufficient condition for condition (D.66) to hold is:

$$\beta_1 > \lambda_{max}(\mathbf{P}_1) + b_3\lambda_{max}(\mathbf{P}_2) + \frac{b_1b_3(\lambda_{max}(\mathbf{P}_2))^2}{\beta_0 - \beta_1\hat{Q}_{1z}b_2 - \lambda_{max}(\mathbf{P}_2)b_1}. \tag{D.75}$$

We see that β_0 must scale linearly with $\hat{Q}_{1z}$ which is one of the parameters that is made arbitrarily large since it grows with λ_2. Then β_0 also needs to become arbitrarily large for the conditions to hold. We choose $\beta_0 = 2\beta_1\hat{Q}_{1z}b_2 + \lambda_{max}(\mathbf{P}_2)b_1$ for the rest of the proof. Condition (D.71) is then automatically verified, and β_1 needs to be chosen according to condition (D.75), which becomes:

$$\beta_1 > \lambda_{max}(\mathbf{P}_1) + b_3\lambda_{max}(\mathbf{P}_2) + \frac{b_1b_3\lambda_{max}^2(\mathbf{P}_2)}{\beta_1\hat{Q}_{1z}b_2} \tag{D.76}$$

This obviously has a bounded solution for large values of $\hat{Q}_{1z}$. We now turn to the second part of (D.63).

Conditions for $\tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0$

We need to study the term $-\mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T$ (we study it with the $-$ sign since the middle matrix is negative definite from conditions (D.65,D.66) which are now verified). As we will see, because of the form of $\mathbf{H}_2$, we do not need to explicitly compute the whole inverse. Let

$$\mathbf{Z} = -(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \\ \mathbf{Z}_2^T & \mathbf{Z}_3 \end{bmatrix}$$

where $\mathbf{Z}$ is divided into blocks of the same size as those of $(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)$. We then have:

$$-\mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T = \mathbf{H}_2\mathbf{Z}\mathbf{H}_2^T \tag{D.77}$$

$$= \begin{bmatrix} \beta_1^2\omega_1^4\mathbf{W}_3^T\mathbf{W}_4\mathbf{Z}_3\mathbf{W}_4^T\mathbf{W}_3 & \mathbf{0}_{\mathbf{M}\times\mathbf{N}} \\ \mathbf{0}_{\mathbf{N}\times\mathbf{M}} & \mathbf{0}_{\mathbf{N}\times\mathbf{N}} \end{bmatrix}. \tag{D.78}$$

We thus only need to characterize the lower right block of $\mathbf{Z}$. It is easy to see that conditions (D.65) and (D.66) also enforce that both the Schur complements associated with the upper left and lower right blocks of $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)$ are invertible, thus giving the following form for $\mathbf{Z}_3$

using the block matrix inversion lemma [76]:

$$\mathbf{Z}_3 = (\beta_0 \mathbf{I_N} - \beta_1 \omega_1^2 \mathbf{W}_4^T \mathbf{W}_4 \tag{D.79}$$

$$- \mathbf{W}_1^T \mathbf{P}_2 \mathbf{W}_1 - \mathbf{W}_1^T \mathbf{P}_2 \mathbf{W}_2 (\beta_1 \mathbf{I_M} - \mathbf{P}_1 - \mathbf{W}_2^T \mathbf{P}_2 \mathbf{W}_2)^{-1} \mathbf{W}_2^T \mathbf{P}_2 \mathbf{W}_1)^{-1}. \tag{D.80}$$

We obtain the following upper bound on the largest eigenvalue of $\mathbf{Z}_3$:

$$\lambda_{max}(\mathbf{Z}_3) \leqslant \frac{1}{\beta_0 - \beta_1 \hat{Q}_{1z} b_2 - \lambda_{max}(\mathbf{P}_2) b_1 - \frac{b_1 b_3 \lambda_{max}^2(\mathbf{P}_2)}{\beta_1 - \lambda_{max}(\mathbf{P}_1) - b_2 \lambda_{max}(\mathbf{P}_2)}}, \tag{D.81}$$

and using the prescription $\beta_0 = 2\beta_1 \hat{Q}_{1z} b_2 + \lambda_{max}(\mathbf{P}_1) b_1$, we get:

$$\lambda_{max}(\mathbf{Z}_3) = \frac{1}{\beta_1 \hat{Q}_{1z} b_2 - \frac{b_1 b_3 \lambda_{max}^2(\mathbf{P}_2)}{\beta_1 - \lambda_{max}(\mathbf{P}_1) - b_2 \lambda_{max}(\mathbf{P}_2)}} \leqslant \frac{b_4}{\hat{Q}_{1z}} \tag{D.82}$$

where b_4 is a constant independent of the arbitrarily large parameters $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$. Thus $\lambda_{max}(\mathbf{Z}_3)$ can be made arbitrarily small by making λ_2 arbitrarily large.

We now want to find conditions for $\tau^2 \mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2 (\mathbf{B}^T \mathbf{P} \mathbf{B} + \mathbf{H}_3)^{-1} \mathbf{H}_2^T \succ 0$ which is equivalent to:

$$\begin{bmatrix} \tau^2 \mathbf{P}_1 - \beta_1 \omega_1^2 \mathbf{W}_3^T \mathbf{W}_3 - \beta_1^2 \omega_1^4 \mathbf{W}_3^T \mathbf{W}_4 \mathbf{Z}_3 \mathbf{W}_4^T \mathbf{W}_3 & \mathbf{0_{M \times N}} \\ \mathbf{0_{N \times M}} & \tau^2 \mathbf{P}_2 - \beta_0 \omega_0^2 \mathbf{I_N} \end{bmatrix}. \tag{D.83}$$

This involves a block diagonal matrix, we only need to check that separate blocks are positive-definite. We start with the upper left block, for which a sufficient condition is:

$$\tau^2 \lambda_{min}(\mathbf{P}_1) - \beta_1 \omega_1^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) - \beta_1^2 \omega_1^4 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \lambda_{max}(\mathbf{W}_4^T \mathbf{W}_4) \lambda_{max}(\mathbf{Z}_3) > 0 \tag{D.84}$$

Using the bounds from appendix 4.4.5, we have:

$$\omega_1^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \leqslant \left(\frac{\hat{Q}_{1z}}{\hat{Q}_{2z}} \right)^2 \left(1 + \frac{(\hat{Q}_{2z})^2 - (\hat{Q}_{1z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2} \right) \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \tag{D.85}$$

$$\leqslant \left(1 + \frac{(\hat{Q}_{2z})^2 - (\hat{Q}_{1z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2} \right) \max \left(\frac{|\hat{Q}_{2x} - \hat{Q}_{1z} \lambda_{min}(\mathbf{F}^T \mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z} \lambda_{min}(\mathbf{F}^T \mathbf{F})}, \frac{|\hat{Q}_{2x} - \hat{Q}_{1z} \lambda_{max}(\mathbf{F}^T \mathbf{F})|}{\hat{Q}_{2x} + \hat{Q}_{2z} \lambda_{max}(\mathbf{F}^T \mathbf{F})} \right)^2 \tag{D.86}$$

$$\leqslant \frac{2\tilde{\lambda}_2 \hat{Q}_{1z} + \tilde{\lambda}_2^2 + (\hat{Q}_{2z})^2}{(\hat{Q}_{1z} + \tilde{\lambda}_2)^2} \max \left(\left(1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{min}(\mathbf{F}^T \mathbf{F}) \right)^2, \left(1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{max}(\mathbf{F}^T \mathbf{F}) \right)^2 \right) \tag{D.87}$$

$$\leqslant \frac{1}{\hat{Q}_{1z}} (2\tilde{\lambda}_2 + \frac{(\tilde{\lambda}_2^2 + (\hat{Q}_{2z})^2)}{\hat{Q}_{1z}}) \max \left(\left(1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{min}(\mathbf{F}^T \mathbf{F}) \right)^2, \left(1 - \frac{\hat{Q}_{1z}}{\hat{Q}_{2x}} \lambda_{max}(\mathbf{F}^T \mathbf{F}) \right)^2 \right) \tag{D.88}$$

Thus there exists a constant b_5, independent of $\lambda_2, \hat{Q}_{1z}, \hat{Q}_{2x}$ such that, for sufficiently large $\hat{Q}_{1z}$:

$$\omega_1^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \leqslant \frac{b_5}{\hat{Q}_{1z}}. \tag{D.89}$$

Remember that we had:

$$\omega_1^2 \lambda_{max}(\mathbf{W}_3^T \mathbf{W}_3) \leqslant \hat{Q}_{1z} b_2, \tag{D.90}$$

which gives the following sufficient condition for the upper left block in (D.83):

$$\tau^2\lambda_{min}(\mathbf{P}_1) - \beta_1\frac{b_5}{\hat{Q}_{1z}} - \beta_1^2\frac{b_2b_5b_4}{\hat{Q}_{1z}} > 0. \tag{D.91}$$

A sufficient condition for the lower right block in (D.83) then reads:

$$\tau^2\lambda_{min}(\mathbf{P}_2) - \beta_0\omega_0^2 > 0, \tag{D.92}$$

where we have:

$$\beta_0\omega_0^2 = \left(\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\right)^2\left(1+\frac{(\hat{Q}_{2x})^2-(\hat{Q}_{1x})^2}{(\hat{Q}_{1x}+\lambda_2)^2}\right)(2\beta_1\hat{Q}_{1z}b_2+\lambda_{max}(\mathbf{P}_2)b_1) \tag{D.93}$$

$$= \frac{1}{\hat{Q}_{2x}}(\hat{Q}_{1x})^2\left(1+\frac{(\hat{Q}_{2x})^2-(\hat{Q}_{1x})^2}{(\hat{Q}_{1x}+\lambda_2)^2}\right)\left(2\beta_1\frac{\hat{Q}_{1z}}{\hat{Q}_{2x}}b_2+\lambda_{max}(\mathbf{P}_2)\frac{b_1}{\hat{Q}_{2x}}\right) \tag{D.94}$$

We remind the reader that $\hat{Q}_{1z}, \hat{Q}_{2x}$ grow linearly with λ_2. Thus the dominant scaling at large λ_2 is (exchanging $\hat{Q}_{2x}$ with $\hat{Q}_{1z}$ up to a constant):

$$\beta_0\omega_0^2 \leqslant \frac{b_6}{\hat{Q}_{1z}}, \tag{D.95}$$

where b_6 is a constant independent of the arbitrarily large quantities. The final condition becomes:

$$\tau^2\lambda_{min}(\mathbf{P}_1) - \beta_1\frac{b_5}{\hat{Q}_{1z}} - \beta_1^2\frac{b_2b_5b_4}{\hat{Q}_{1z}} > 0 \tag{D.96}$$

$$\tau^2\lambda_{min}(\mathbf{P}_2) - \frac{b_6}{\hat{Q}_{1z}} > 0 \tag{D.97}$$

where we want $\tau < 1$. We now choose $\tau^2 = \tilde{c}/\hat{Q}_{1z}$ with a constant $\tilde{c}$ independent of $\lambda_2, \hat{Q}_{1z}, \hat{Q}_{2x}$ that verifies $\tilde{c} > \max\left(\frac{\beta_1 b_5+\beta_1^2 b_2 b_5 b_4}{\lambda_{min}(\mathbf{P}_1)}, \frac{b_6}{\lambda_{min}(\mathbf{P}_2)}\right)$, such that:

$$\frac{\tilde{c}}{\hat{Q}_{1z}}\lambda_{min}(\mathbf{P}_1) - \beta_1\frac{b_5}{\hat{Q}_{1z}} - \beta_1^2\frac{b_2b_5b_4}{\hat{Q}_{1z}} > 0 \tag{D.98}$$

$$\frac{\tilde{c}}{\hat{Q}_{1z}}\lambda_{min}(\mathbf{P}_2) - \frac{b_6}{\hat{Q}_{1z}} > 0. \tag{D.99}$$

Since β_1 is bounded for large values of $\hat{Q}_{1z}$, and the b_i and c are constants independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$, one can then just enforce $\tilde{c} < \hat{Q}_{1z}$ if λ_2 is large enough. We then obtain $\tau < 1$ and a linear convergence rate of $\sqrt{c/\lambda_2}$, taking $c \equiv \tilde{c}\lambda_2/\hat{Q}_{1z}$. This rate being strictly smaller than 1, it ensures convergence. We see that the eigenvalues of the matrix $\mathbf{P}$ are of little importance as long as they are non-vanishing. We choose $\mathbf{P}$ as the identity. Inequality (4.55) thus holds, and our linear system converges, also proving convergence of Oracle MLVAMP in the considered regime of regularization.

Appendix E

Replica computation of the ground state energy for perceptrons for a Gaussian i.i.d. matrix

E.1 Notations and problem setting

In this section, we present the replica computation of for generalized linear models, corresponding to the hypothesis class $\mathcal{F}_\varphi$ from (5.32), i.e.

$$\mathcal{F}_\varphi \equiv \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^N \mapsto \{-1, 1\} \\ \mathbf{x} \mapsto \varphi\left(\frac{1}{\sqrt{N}} \mathbf{w}^T \mathbf{x}\right) \end{cases} , \mathbf{w} \in \mathbb{R}^N \right\} . \tag{E.1}$$

We gather the examples into a data matrix $\{\mathbf{x}_1^T, \dots, \mathbf{x}_M^T\} = \mathbf{X} \in \mathbb{R}^{M \times N}$, where each element is identically and independently sampled from $P_x(\mathbf{x}) = \mathcal{N}(0, 1)$, and corresponding labels $\mathbf{y} = (y_1, ..., y_M)$ are drawn randomly from $P_y(.)$. We consider for the moment a generic prior distribution $\mathbf{w} \sim P_w(.)$ that factorizes, and a component-wise activation function $\varphi(.)$. The cost function of a given sample is $V(y_\mu | z_\mu) = \mathbb{1}\left[y_\mu \neq \varphi(z_\mu)\right]$ where $z_\mu \equiv \frac{1}{\sqrt{N}} \mathbf{w}^T \mathbf{x}_\mu$. It returns 0 if the the estimator classifies the example correctly, and 1 otherwise. Finally, we define the constraint function at inverse temperature β, that depends explicitly on the Hamiltonian (5.24):

$$\mathcal{I}(\mathbf{y}|\mathbf{z}, \beta) \equiv \prod_{\mu=1}^{M} e^{-\beta V(y_\mu | z_\mu)} = e^{-\beta \mathcal{H}(\{\mathbf{y}, \mathbf{X}\}, \mathbf{w})}. \tag{E.2}$$

Note that the constraint function converges at zero temperature to a hard constraint function $\mathcal{I}(\mathbf{y}|\mathbf{z}, \beta) \underset{\beta \to \infty}{\longrightarrow} \prod_{\mu=1}^{M} \mathbb{1}\left[V(y_\mu | z_\mu) = 0\right]$.To compute the replicated partition function, we introduce $n \in \mathbb{N}$ replicas of the system, and resort to the replica trick (5.37). Assuming there exists an analytical continuation for $n \to 0$ and that we can revert limits, the averaged free energy Φ of the initial system becomes:

$$\Phi(\alpha, \beta) = -\lim_{n \to 0} \left[\lim_{N \to \infty} \frac{1}{N\beta} \frac{\partial \log \mathbb{E}_{\mathbf{y}, \mathbf{X}} \left[\mathcal{Z}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta)^n\right]}{\partial n} \right] , \tag{E.3}$$

where the replicated partition function average reads

$$\mathbb{E}_{\mathbf{y}, \mathbf{X}} \left[\mathcal{Z}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta)^n\right] = \int_{\mathbb{R}^M} \mathrm{d}\mathbf{y} P_y(\mathbf{y}) \int_{\mathbb{R}^{M \times N}} \mathrm{d}\mathbf{X} P_x(\mathbf{X}) \mathcal{Z}(\{\mathbf{y}, \mathbf{X}\}, \alpha, \beta)^n \tag{E.4}$$

$$= \int_{\mathbb{R}^M} \mathrm{d}\mathbf{y} P_y(\mathbf{y}) \int_{\mathbb{R}^{M \times N}} \mathrm{d}\mathbf{X} P_x(\mathbf{X}) \prod_{a=1}^{n} \int_{\mathbb{R}^N} \mathrm{d}P_w(\mathbf{w}^a) \prod_{\mu=1}^{M} \int \mathrm{d}z_\mu^a \mathcal{I}(y_\mu | z_\mu^a, \beta) \, \delta\left(z_\mu^a - \frac{1}{\sqrt{N}} \mathbf{w}^{aT} \mathbf{x}_\mu\right) . \tag{E.5}$$

E.2 Average over the Gaussian i.i.d. data matrix

As the data matrix is taken (Gaussian) i.i.d., we have for $i,j \in [\![1;N]\!]$ and $\mu,\nu \in [\![1;M]\!]$ the expectation $\mathbb{E}_{\mathbf{X}}[x_{\mu i}x_{\nu j}] = \delta_{\mu\nu}\delta_{ij}$. Hence $z_\mu^a = \frac{1}{\sqrt{N}}\sum_{i=1}^N x_{\mu i}w_i^a$ is the sum of i.i.d. random variables, and the central limit theorem guarantees that in the large size limit $N \to \infty$, $z_\mu^a \sim \mathcal{N}\left(\mathbb{E}_{\mathbf{X}}[z_\mu^a], \mathbb{E}_{\mathbf{X}}[z_\mu^a z_\mu^b]\right)$, with the two first moments given by

$$\mathbb{E}_{\mathbf{X}}[z_\mu^a] = \frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{E}_{\mathbf{X}}[x_{\mu i}]w_i^a = 0 \tag{E.6}$$

$$\mathbb{E}_{\mathbf{X}}[z_\mu^a z_\nu^b] = \frac{1}{N}\sum_{i,j}\mathbb{E}_{\mathbf{X}}[x_{\mu i}x_{\nu j}]w_i^a w_j^b = \frac{1}{N}\sum_{i,j}\delta_{\mu\nu}\delta_{ij}w_i^a w_j^b = \left(\frac{1}{N}\sum_{i=1}^N w_i^a w_i^b\right)\delta_{\mu\nu}\,. \tag{E.7}$$

In the following, we introduce the overlap matrix of size $n \times n$ defined by its elements:

$$Q_{ab} = \frac{\mathbf{w}^{aT}\mathbf{w}^b}{N} \tag{E.8}$$

and the vectors $\tilde{\mathbf{z}}_\mu \equiv (z_\mu^a)_{a=1,\dots,n}$ and $\tilde{\mathbf{w}}_i \equiv (w_i^a)_{a=1,\dots,n} \in \mathbb{R}^n$. From the above calculation, $\tilde{\mathbf{z}}_\mu$ follows a multivariate Gaussian distribution $\tilde{\mathbf{z}}_\mu \sim P_{\tilde{z}}(\tilde{\mathbf{z}},\mathbf{Q}) \overset{\mathrm{d}}{=} \mathcal{N}(\mathbf{0}_n,\mathbf{Q})$ where $\overset{\mathrm{d}}{=}$ stands for equality in distribution, and $P_{\tilde{w}}(\tilde{\mathbf{w}}_i) = \prod_{a=1}^n P_w(\tilde{w}_i^a)$. We do a change of variable, using the Fourier representation of the δ-Dirac function, bringing in a new matrix $\hat{\mathbf{Q}}$ of size $n \times n$:

$$1 = \int_{\mathbb{R}^{n\times n}} \mathrm{d}\mathbf{Q} \prod_{a\le b}\delta\left(NQ_{ab} - \sum_{i=1}^N w_i^a w_i^b\right) \tag{E.9}$$

$$\propto \int_{\mathbb{R}^{n\times n}} \mathrm{d}\mathbf{Q}\int_{\mathbb{R}^{n\times n}} \mathrm{d}\hat{\mathbf{Q}}\exp\left(-\frac{N}{2}\mathrm{Tr}\mathbf{Q}\hat{\mathbf{Q}}\right)\exp\left\{\frac{1}{2}\sum_{i=1}^N \tilde{\mathbf{w}}_i^T\hat{\mathbf{Q}}\tilde{\mathbf{w}}_i\right\}. \tag{E.10}$$

The replicated partition function factorizes and becomes an integral over the matrix parameters $\mathbf{Q}$ and $\hat{\mathbf{Q}}$, that can be evaluated using a Laplace method in the $N \to \infty$ limit,

$$\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^n\right] \propto \int \mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}} e^{N\Phi^{(n)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)} \underset{N\to\infty}{\simeq} e^{N\,\mathrm{Extr}_{\mathbf{Q},\hat{\mathbf{Q}}}\left\{\Phi^{(n)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)\right\}} \tag{E.11}$$

where

$$\begin{cases} \Phi^{(n)}\left(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta\right) \equiv -\frac{1}{2}\mathrm{Tr}\mathbf{Q}\hat{\mathbf{Q}} + \log\Psi_{\mathrm{w}}^{(n)}(\hat{\mathbf{Q}}) + \alpha\log\Psi_{\mathrm{out}}^{(n)}(\mathbf{Q},\beta)\,, \\ \Psi_{\mathrm{w}}^{(n)}(\hat{\mathbf{Q}}) = \int_{\mathbb{R}^n}\mathrm{d}\tilde{w}P_{\tilde{w}}(\tilde{\mathbf{w}})e^{\frac{1}{2}\tilde{\mathbf{w}}^T\hat{\mathbf{Q}}\tilde{\mathbf{w}}}\,, \\ \Psi_{\mathrm{out}}^{(n)}(\mathbf{Q},\beta) = \int \mathrm{d}y P_y\left(y\right)\int_{\mathbb{R}^n}\mathrm{d}\tilde{z}P_{\tilde{z}}(\tilde{\mathbf{z}},\mathbf{Q})\mathcal{I}(y|\tilde{\mathbf{z}},\beta). \end{cases} \tag{E.12}$$

Finally, using (E.3) and switching the two limits $n \to 0$ and $N \to \infty$, the quenched free energy Φ simplifies as an extremization problem

$$\Phi(\alpha,\beta) = -\frac{1}{\beta}\underset{\mathbf{Q},\hat{\mathbf{Q}}}{\mathrm{Extr}}\left\{\lim_{n\to 0}\frac{\partial\Phi^{(n)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)}{\partial n}\right\}, \tag{E.13}$$

over general symmetric matrices $\mathbf{Q}$ and $\hat{\mathbf{Q}}$. Until now, we have written the main steps without picking an ansatz on overlap parameters, we will now see how the calculation unfolds for each choice.

Choosing an ansatz The simplest and commonly used ansatz yield the following form for the overlap matrix:

- Replica Symmetry (RS) ansatz: $\mathbf{Q}^{(\mathrm{RS})} = (Q - q_0)\mathbf{I}_n + q_0\mathbf{J}_n$
- 1-Step Replica Symmetry Breaking (1RSB) ansatz:
  $\mathbf{Q}^{(\mathrm{1RSB})} = (Q - q_1)\mathbf{I}_n + (q_1 - q_0)\mathbf{I}_{n/m_0} \otimes \mathbf{J}_{m_0} + q_0\mathbf{J}_n$,
- 2-Step Replica Symmetry Breaking (2RSB) ansatz:
  $\mathbf{Q}^{(\mathrm{2RSB})} = (Q - q_2)\,\mathbf{I}_n + (q_2 - q_1)\,\mathbf{I}_{n/m_1} \otimes \mathbf{J}_{m_1} + (q_1 - q_0)\,\mathbf{I}_{n/m_0} \otimes \mathbf{J}_{m_0} + q_0\mathbf{J}_n$

where $\mathbf{I}_k$ is the identity matrix of size k, and $\mathbf{J}_k$ is the matrix of size k full of ones. Q is always the self-overlap, while overlaps q_0, q_1, q_2 show up successively through steps of symmetry breaking, and each time express different levels of overlaps within and between clusters of solutions. Some intuition on these ansatz was given in 1.1.4. Plugging these ansatz in (E.13), and taking the derivative followed by the $n \to 0$ limit, optimizing over the space of matrices boil down to a much simpler optimization problem over a few scalar order parameters.

E.3 RS free energy for an i.i.d. data matrix

Let us compute the functional $\Phi^{(n)}(\mathbf{Q}, \hat{\mathbf{Q}}, \alpha, \beta)$ appearing in the free energy eq. (E.13) in the RS ansatz. The latter assumes that all replica remain equivalent with a common overlap $q_0 = \frac{1}{N}\sum_{i=1}^{N} w_i^a w_i^b$ for $a \neq b$ and a norm $Q = \frac{1}{N}\sum_{i=1}^{N} w_i^a w_i^a$, leading to the following expressions for matrices $\mathbf{Q}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{n\times n}$:

$$\mathbf{Q}^{(\mathrm{RS})} = \begin{pmatrix} Q & q_0 & \dots & q_0 \\ q_0 & Q & \dots & \dots \\ \dots & \dots & \dots & q_0 \\ q_0 & \dots & q_0 & Q \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{Q}}^{(\mathrm{RS})} = \begin{pmatrix} \hat{Q} & \hat{q}_0 & \dots & \hat{q}_0 \\ \hat{q}_0 & \hat{Q} & \dots & \dots \\ \dots & \dots & \dots & \hat{q}_0 \\ \hat{q}_0 & \dots & \hat{q}_0 & \hat{Q} \end{pmatrix} . \tag{E.14}$$

Let us compute separately the terms involved in the functional $\Phi^{(n)}(\mathbf{Q}, \hat{\mathbf{Q}}, \alpha, \beta)$ in (E.12): the first is a trace term, the second $\Psi_{\mathrm{w}}^{(n)}$ depends on the weights prior, and the third term $\Psi_{\mathrm{out}}^{(n)}$ depends on the constraint $\mathcal{I}(y|z)$ (E.2).

i) Trace term

The trace term can be easily computed and takes the form

$$\frac{1}{2}\mathrm{Tr}\mathbf{Q}\hat{\mathbf{Q}}\Big|_{\mathrm{RS}} = \frac{1}{2}\left(nQ\hat{Q} + n(n-1)q_0\hat{q}_0\right) . \tag{E.15}$$

ii) Prior integral

Evaluated at the RS fixed point, and using the Hubbard-Stratonovich transform (or equivalently a Gaussian identity, which sounds less classy, see footnote 1), the prior integral can be further simplified

$$\begin{aligned}\Psi_{\mathrm{w}}^{(n)}(\hat{\mathbf{Q}})\Big|_{\mathrm{RS}} &= \int \mathrm{d}\tilde{w} P_{\tilde{w}}(\tilde{\mathbf{w}}) e^{\frac{1}{2}\tilde{\mathbf{w}}^T \hat{\mathbf{Q}}_{\mathrm{RS}} \tilde{\mathbf{w}}} = \int \mathrm{d}\tilde{w} P_{\tilde{w}}(\tilde{\mathbf{w}}) \exp\left(\frac{(\hat{Q} - \hat{q}_0)}{2}\sum_{a=1}^{n}(\tilde{w}^a)^2\right) \exp\left(\hat{q}_0\left(\sum_{a=1}^{n}\tilde{w}^a\right)^2\right) \\ &= \int \mathrm{D}\xi_0 \left[\int \mathrm{d}w P_w(w) \exp\left(\frac{(\hat{Q} - \hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right]^n . \end{aligned} \tag{E.16}$$

iii) Constraint integral

Recall the vector $\tilde{\mathbf{z}} \sim P_{\tilde{z}}(\tilde{\mathbf{z}}, \mathbf{Q}) \overset{\mathrm{d}}{=} \mathcal{N}(\mathbf{0}_n, \mathbf{Q})$ follows a Gaussian distribution with zero mean and covariance matrix $\mathbf{Q}$. In the RS ansatz, the covariance can be rewritten as a linear combination

of the identity $\mathbf{I}_n$ and $\mathbf{J}_n$: $Q^{(\text{2RSB})} = (Q - q_0)\mathbf{I}_n + q_0\mathbf{J}_n$, which allows to split the variable z^a into two Gaussian parts

$$z^a = \sqrt{q_0}\xi_0 + \sqrt{Q - q_0}u^a$$

with $\xi_0 \sim \mathcal{N}(0,1)$ and $\forall a,\ u_a \sim \mathcal{N}(0,1)$. The constraint integral then reads:

$$\begin{aligned}\Psi_{\text{out}}^{(n)}(\mathbf{Q},\beta)\Big|_{\text{RS}} &= \int \mathrm{d}y P_y(y) \int_{\mathbb{R}^n} \mathrm{d}\tilde{z} P_{\tilde{z}}(\tilde{\mathbf{z}},\mathbf{Q})\mathcal{I}(y|\tilde{\mathbf{z}},\beta) \\ &= \int \mathrm{d}y P_y(y) \int \mathrm{D}\xi_0 \int \prod_{a=1}^{n} \mathrm{D}u^a \mathcal{I}\left(y|\sqrt{q_0}\xi_0 + \sqrt{Q-q_0}u^a,\beta\right) \\ &= \int \mathrm{d}y P_y(y) \int \mathrm{D}\xi_0 \left[\int \mathrm{D}z \mathcal{I}\left(y|\sqrt{q_0}\xi_0 + \sqrt{Q-q_0}z,\beta\right)\right]^n .\end{aligned} \tag{E.17}$$

Putting all pieces together, the functional $\Phi^{(n)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)$ taken at the RS fixed point has an explicit formula and dependency in n:

$$\begin{aligned}\Phi^{(n)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)\Big|_{\text{RS}} \underset{n\to 0}{\simeq} &-\frac{1}{2}\left(nQ\hat{Q} + n(n-1)q_0\hat{q}_0\right) \\ &+ n\int \mathrm{D}\xi_0 \log\left(\int \mathrm{d}w P_w(w) \exp\left(\frac{(\hat{Q}-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right) \\ &+ n\alpha \int \mathrm{d}y P_y(y) \int \mathrm{D}\xi_0 \log\left(\int \mathrm{D}z \mathcal{I}\left(y|\sqrt{q_0}\xi_0 + \sqrt{Q-q_0}z,\beta\right)\right).\end{aligned} \tag{E.18}$$

RS free energy Taking the derivative with respect to n and the $n \to 0$ limit, the RS free energy has a simple expression

$$\Phi^{(\text{RS})}(\alpha,\beta) = -\frac{1}{\beta}\underset{q_0,\hat{q}_0}{\mathbf{extr}}\left\{-\frac{1}{2}Q\hat{Q} + \frac{1}{2}q_0\hat{q}_0 + \Psi_{\text{w}}^{(\text{RS})}(\hat{q}_0) + \alpha\Psi_{\text{out}}^{(\text{RS})}(q_0,\beta)\right\}, \tag{E.19}$$

$$\Psi_{\text{w}}^{(\text{RS})}(\hat{q}_0) \equiv \mathbb{E}_{\xi_0}\log\mathbb{E}_w\left[\exp\left(\frac{(\hat{Q}-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right], \tag{E.20}$$

$$\Psi_{\text{out}}^{(\text{RS})}(q_0,\beta) \equiv \mathbb{E}_y\mathbb{E}_{\xi_0}\log\mathbb{E}_z\left[\mathcal{I}\left(y|\sqrt{Q-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right], \tag{E.21}$$

where $\xi_0, z \sim \mathcal{N}(0,1)$, $w \sim P_w(.), y \sim P_y(.)$ and $Q = \hat{Q} = 1$ in the case where $\frac{1}{N}\|\mathbf{w}\|_2^2 = 1$.

Simplification in the spherical case In the spherical/Gaussian case, $\Psi_{\text{w}}^{(n)}(\hat{\mathbf{Q}})$ in eq. (E.12) can be directly integrated as

$$\Psi_{\text{w}}^{(n)}(\hat{\mathbf{Q}}) = \int_{\|\tilde{\mathbf{w}}\|_2^2 = N} \mathrm{d}\tilde{\mathbf{w}} e^{\frac{1}{2}\tilde{\mathbf{w}}^T\hat{\mathbf{Q}}\tilde{\mathbf{w}}} = -\frac{1}{2}\log\det\left(2\pi(\mathbf{I}_n + \hat{\mathbf{Q}})\right) \tag{E.22}$$

Besides, taking the derivative of (E.12) with respect to $\hat{\mathbf{Q}}$, we find $\mathbf{Q}^{-1} = (\mathbf{I}_n + \hat{\mathbf{Q}})$. Finally, we get rid of $\hat{\mathbf{Q}}$ to reach

$$\Phi^{(n)}(\mathbf{Q},\alpha,\beta) \equiv \frac{1}{2}\log\det(2\pi\mathbf{Q}) + \alpha\log\Psi_{\text{out}}^{(n)}(\mathbf{Q},\beta). \tag{E.23}$$

The above determinant reads in the RS ansatz

$$\frac{1}{2}\det(\mathbf{Q})\Big|_{\text{RS}} \simeq \frac{n}{2}\left(\log(1-q_0) + \frac{q_0}{1+(n-1)q_0} + \mathcal{O}(n)\right). \tag{E.24}$$

Finally, the RS free energy in this simple case becomes

$$\Phi^{(\mathrm{RS})}(\alpha,\beta) = -\frac{1}{\beta}\underset{q_0}{\mathrm{Extr}}\left\{\frac{1}{2(1-q_0)} + \frac{1}{2}\log(2\pi) + \frac{1}{2}\log(1-q_0) + \alpha\Psi_{\mathrm{out}}^{(\mathrm{RS})}(q_0,\beta)\right\}, \tag{E.25}$$

with $\Psi_{\mathrm{out}}^{(\mathrm{RS})}$ defined in (E.21).

E.4 1RSB free energy for an i.i.d. data matrix

The free energy (E.13) can also be evaluated at the simplest non trivial fixed point, thanks to the one-step Replica Symmetry Breaking ansatz (1RSB). Instead of assuming that replicas are equivalent, it states that the symmetry between replica is broken and that replicas are clustered in different states, with inner overlap q_1 and outer overlap q_0. Translating this analytically, the matrices can be expressed as function of the Parisi parameter m_0:

$$\begin{aligned}\mathbf{Q}^{(\mathrm{1RSB})} &= q_0\mathbf{J}_n + (q_1-q_0)\,\mathbf{I}_{\frac{n}{m_0}}\otimes\mathbf{J}_{m_0} + (Q-q_1)\,\mathbf{I}_n\\ \hat{\mathbf{Q}}^{(\mathrm{1RSB})} &= \hat{q}_0\mathbf{J}_n + (\hat{q}_1-\hat{q}_0)\,\mathbf{I}_{\frac{n}{m_0}}\otimes\mathbf{J}_{m_0} + \left(\hat{Q}-\hat{q}_1\right)\mathbf{I}_n\,.\end{aligned} \tag{E.26}$$

i) Trace term

Again, the trace term can be easily computed

$$\frac{1}{2}\mathrm{Tr}\mathbf{Q}\hat{\mathbf{Q}}\Big|_{\mathrm{1RSB}} = \frac{1}{2}\left(nQ\hat{Q} + n(m_0-1)q_1\hat{q}_1 + n(n-m_0)q_0\hat{q}_0\right). \tag{E.27}$$

ii) Prior integral

Separating replicas with different overlaps q_0, q_1, the prior integral can be written, using Hubbard-Stratonovich transforms to decouple replicas, as

$$\begin{aligned}\Psi_{\mathrm{w}}^{(n)}(\hat{\mathbf{Q}})\Big|_{\mathrm{1RSB}} &= \int \mathrm{d}\tilde{\mathbf{w}}P_{\tilde{w}}(\tilde{\mathbf{w}})e^{\frac{(\hat{Q}-\hat{q}_1)}{2}\sum_{a=1}^{n}(\tilde{w}^a)^2+\frac{(\hat{q}_1-\hat{q}_0)}{2}\sum_{k=1}^{\frac{n}{m_0}}\sum_{a,b=(k-1)m_0+1}^{km_0}\tilde{w}^a\tilde{w}^b+\frac{\hat{q}_0}{2}\left(\sum_{a=1}^{n}\tilde{w}^a\right)^2}\\ &= \int \mathrm{D}\xi_0\left[\int \mathrm{D}\xi_1\left[\int \mathrm{d}wP_w(w)\exp\left(\frac{(\hat{Q}-\hat{q}_1)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1-\hat{q}_0}\xi_1\right)w\right)\right]^{m_0}\right]^{\frac{n}{m_0}}\end{aligned} \tag{E.28}$$

with $\xi_0, \xi_1 \sim \mathcal{N}(0,1)$.

iii) Constraint integral

Again, the vector $\tilde{\mathbf{z}} \sim P_{\tilde{z}} \overset{\mathrm{d}}{=} \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{(\mathrm{1RSB})}\right)$ is distributed as a Gaussian vector with zero mean and covariance

$$\mathbf{Q}^{(\mathrm{1RSB})} = q_0\mathbf{J}_n + (q_1-q_0)\,\mathbf{I}_{\frac{n}{m_0}}\otimes\mathbf{J}_{m_0} + (Q-q_1)\,\mathbf{I}_n.$$

The Gaussian vector of covariance $\mathbf{Q}^{(\mathrm{1RSB})}$ can be decomposed in a sum of normal Gaussian vectors $\xi_0 \sim \mathcal{N}(0,1)$, $\xi_k \sim \mathcal{N}(0,1)\ \forall k \in [\![1; \frac{n}{m_0}]\!]$ and $u_a \sim \mathcal{N}(0,1)\ \forall a \in [\![(k-1)m_0+1; km_0]\!]$:

$$z^a = \sqrt{q_0}t_0 + \sqrt{q_1-q_0}t_k + \sqrt{Q-q_1}u_a\,.$$

Finally the constraint integral reads

$$\begin{aligned}
&\Psi_{\text{out}}^{(n)}(Q,\beta)\Big|_{\text{1RSB}} \\
&= \int \mathrm{d}y P_y(y) \int \mathrm{D}\xi_0 \int \prod_{k=1}^{n/m_0} \mathrm{D}\xi_k \int \prod_{a=(k-1)m_0+1}^{km_0} \mathrm{D}u_a \mathcal{I}\left(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_k + \sqrt{Q-q_1}u_a, \beta\right) \\
&= \int \mathrm{d}y P_y(y) \int \mathrm{D}\xi_0 \left[\int \mathrm{D}\xi_1 \left[\int \mathrm{D}z \mathcal{I}\left(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{Q-q_1}z, \beta\right)\right]^{m_0}\right]^{\frac{n}{m_0}}. \qquad \text{(E.29)}
\end{aligned}$$

Gathering previous computations (E.27, E.28, E.29), the functional $\Phi^{(n)}$ evaluated at the 1RSB fixed point reads:

$$\begin{aligned}
&\Phi^{(n)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)\Big|_{\text{1RSB}} \\
&\underset{n\to 0}{\simeq} -\frac{1}{2}\left(nQ\hat{Q} + n(m_0-1)q_1\hat{q}_1 + n(n-m_0)q_0\hat{q}_0\right) \qquad \text{(E.30)} \\
&+ \frac{n}{m_0}\int \mathrm{D}\xi_0 \log\left(\int \mathrm{D}\xi_1 \left[\int \mathrm{d}w P_w(w) \exp\left(\frac{(\hat{Q}-\hat{q}_1)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1-\hat{q}_0}\xi_1\right)w\right)\right]^{m_0}\right) \\
&+ \alpha\frac{n}{m_0}\int \mathrm{d}y P_y(y) \int \mathrm{D}\xi_0 \log\left(\int \mathrm{D}\xi_1 \left[\int \mathrm{D}z \mathcal{I}\left(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{Q-q_1}z, \beta\right)\right]^{m_0}\right).
\end{aligned}$$

1RSB free energy The free energy for the 1RSB ansatz reads:

$$\begin{aligned}
\Phi^{(\text{1RSB})}(\alpha,\beta) &= -\frac{1}{\beta}\underset{\mathbf{q},\hat{\mathbf{q}},m_0}{\mathbf{extr}}\left\{\frac{1}{2}\left(q_1\hat{q}_1 - Q\hat{Q}\right) + \frac{m_0}{2}\left(q_0\hat{q}_0 - q_1\hat{q}_1\right)\right. \\
&\qquad \left. + \Psi_{\text{w}}^{(\text{1RSB})}(\hat{\mathbf{q}}) + \alpha\Psi_{\text{out}}^{(\text{1RSB})}(\mathbf{q},\beta)\right\} \qquad \text{(E.31)} \\
\Psi_{\text{w}}^{(\text{1RSB})}(\hat{\mathbf{q}}) &\equiv \frac{1}{m_0}\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_w\left[\exp\left(\frac{(\hat{Q}-\hat{q}_1)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1-\hat{q}_0}\xi_1\right)w\right)\right]^{m_0}\right) \\
\Psi_{\text{out}}^{(\text{1RSB})}(\mathbf{q},\beta) &\equiv \frac{1}{m_0}\mathbb{E}_y\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_z\left[\mathcal{I}(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{Q-q_1}z, \beta)\right]^{m_0}\right)
\end{aligned}$$

where $\mathbf{q} = (q_0, q_1)$, $\hat{\mathbf{q}} = (\hat{q}_0, \hat{q}_1)$, $\xi_0, \xi_1, z \sim \mathcal{N}(0,1)$, $w \sim P_w(.)$, $y \sim P_y(.)$ and $Q = \hat{Q} = 1$.

In the spherical case, equation (E.23) remains valid. We can compute the overlap matrix determinant in the 1RSB ansatz:

$$\begin{aligned}
\det \mathbf{Q}|_{\text{1RSB}} &= \left(nq_0 + m_0(q_1-q_0) + (1-q_1)\right) \times (1-q_1)^{n-\frac{n}{m_0}} \times \left(m_0(q_1-q_0) + (1-q_1)\right)^{\frac{n}{m_0}-1} \\
\log\det \mathbf{Q}|_{\text{1RSB}} &\simeq n\left(\frac{m_0-1}{m_0}\log(1-q_1) + \frac{\log\left(m_0(q_1-q_0)+(1-q_1)\right)}{m_0} + \frac{q_0}{m_0(q_1-q_0)+(1-q_1)}\right)
\end{aligned}$$

Using the above expression for the determinant and the simplified replica potential (E.23) we obtain

$$\begin{aligned}
\Phi^{(\text{1RSB})}(\alpha,\beta) = -\frac{1}{\beta}\underset{q_0,q_1,m_0}{\text{Extr}}\left\{\frac{1}{2}\log(2\pi) + \frac{m_0-1}{2m_0}\log(1-q_1) + \frac{1}{2m_0}\log\left(m_0(q_1-q_0)+(1-q_1)\right)\right. \\
\left. + \frac{q_0}{2\left(m_0(q_1-q_0)+(1-q_1)\right)} + \alpha\Psi_{\text{out}}^{(\text{1RSB})}(\mathbf{q},\beta)\right\}. \qquad \text{(E.32)}
\end{aligned}$$

E.5 2RSB free energy for an i.i.d. data matrix

Analogously, the 2RSB ansatz for $m_1 < m_0 < n$ reads

$$\mathbf{Q}^{(\mathrm{2RSB})} = q_0\mathbf{J}_n + (q_1 - q_0)\,\mathbf{I}_{\frac{n}{m_0}} \otimes \mathbf{J}_{m_0} + (q_2 - q_1)\,\mathbf{I}_{\frac{n}{m_1}} \otimes \mathbf{J}_{m_1} + (Q - q_2)\,\mathbf{I}_n \tag{E.33}$$

$$\hat{\mathbf{Q}}^{(\mathrm{2RSB})} = \hat{q}_0\mathbf{J}_n + (\hat{q}_1 - \hat{q}_0)\,\mathbf{I}_{\frac{n}{m_0}} \otimes \mathbf{J}_{m_0} + (\hat{q}_2 - \hat{q}_1)\,\mathbf{I}_{\frac{n}{m_1}} \otimes \mathbf{J}_{m_1} + \left(\hat{Q} - \hat{q}_2\right)\mathbf{I}_n \tag{E.34}$$

and the above computation of the free energy generalizes to

$$\Phi^{(\mathrm{2RSB})}(\alpha,\beta) = -\frac{1}{\beta}\underset{\mathbf{q},\hat{\mathbf{q}},x0,m_1}{\mathrm{Extr}}\left\{\frac{1}{2}\left(q_2\hat{q}_2 - Q\hat{Q}\right) + \frac{m_1}{2}\left(q_1\hat{q}_1 - q_2\hat{q}_2\right) + \frac{m_0}{2}\left(q_0\hat{q}_0 - q_1\hat{q}_1\right)\right.$$
$$\left. + \Psi_{\mathrm{w}}^{(\mathrm{2RSB})}(\hat{\mathbf{q}}) + \alpha\Psi_{\mathrm{out}}^{(\mathrm{2RSB})}(\mathbf{q},\beta)\right\} \tag{E.35}$$

$$\Psi_{\mathrm{w}}^{(\mathrm{2RSB})}(\hat{\mathbf{q}}) \equiv \frac{1}{m_1}\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\left[\mathbb{E}_{\xi_2}\mathbb{E}_w\left[e^{\frac{(\hat{Q}-\hat{q}_2)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1-\hat{q}_0}\xi_1 + \sqrt{\hat{q}_2-\hat{q}_1}\xi_2\right)w}\right]^{m_1}\right]^{\frac{m_0}{m_1}}\right)$$

$$\Psi_{\mathrm{out}}^{(\mathrm{2RSB})}(\mathbf{q},\beta) \equiv \frac{1}{m_1}\mathbb{E}_y\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\left[\mathbb{E}_{\xi_2}\mathbb{E}_z\left[\mathcal{I}(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{q_2-q_1}\xi_2 + \sqrt{Q-q_2}z,\beta)\right]^{m_1}\right]^{\frac{m_0}{m_1}}\right).$$

E.6 Ground state energies - Spherical case

We focus on the particular case of the spherical perceptron with weights $\mathbf{w} \in \mathbb{R}^N$ such that $\|\mathbf{w}\|_2^2 = N$.

E.6.1 RS ground state energy $e_{\mathrm{gs}}^{(\mathrm{RS})}$

To compute the ground state energy, we first need to take both limits $q_0 \to 1$ and $\beta \to \infty$, keeping the product $\chi = \beta(Q - q_0)$ finite [98, 50, 180] inside (E.21), which reads:

$$\Psi_{\mathrm{out}}^{(\mathrm{RS})}(q_0,\beta) \equiv \mathbb{E}_y\mathbb{E}_{\xi_0}\log\mathbb{E}_z\left[\mathcal{I}\left(y|\sqrt{Q-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right]$$
$$\underset{(q_0,\beta)\to(1,\infty)}{\simeq} -\frac{1}{2}\log(2\pi(Q-q_0)) - \beta\int \mathrm{d}P_y(y)\int \mathrm{D}\xi_0 \min_{\xi_0,z}\left[V(y|z) + \frac{(z-\xi_0)^2}{2\chi}\right]$$

that finally leads, taking limits $q_0 \to 1$, $\beta \to \infty$ in (E.25), to the RS ground state energy

$$e_{\mathrm{gs}}^{(\mathrm{RS})} = \underset{\chi}{\mathrm{Extr}}\left\{-\frac{1}{2\chi} + \alpha\mathbb{E}_{y,\xi_0}\min_z\left[V(y|z) + \frac{(z-\xi_0)^2}{2\chi}\right]\right\}. \tag{E.36}$$

Application to the step-perceptron Taking the step function $V(y|z) = \theta(\kappa - z)$ with $\kappa > 0$ a robustness parameter and $P_y(y) = \delta(y-1)$ leads to the expression [58]:

$$e_{\mathrm{gs}}^{(\mathrm{RS})} = \underset{\chi}{\mathrm{Extr}}\left\{-\frac{1}{2\chi} + \alpha\left(\int_{-\infty}^{\kappa-\sqrt{2\chi}}\mathrm{D}\xi + \int_{\kappa-\sqrt{2\chi}}^{\kappa}\mathrm{D}\xi\frac{(\xi-\kappa)^2}{2\chi}\right)\right\}. \tag{E.37}$$

E.6.2 1RSB ground state energy $e_{\mathrm{gs}}^{(\mathrm{1RSB})}$

To compute the ground state energy in the 1RSB ansatz, we first need to take limits $q_1 \to 1$ with $\beta \to \infty$ and $m_0 \to 0$, keeping the products $\chi \equiv \beta(Q - q_1)$ and $\omega_0 \equiv m_0\beta$ finite [180], with

$\Delta q = 1 - q_0$. Recall

$$\begin{aligned}\Psi_{\text{out}}^{(1\text{RSB})}(\mathbf{q},\beta) &\equiv \frac{1}{m_0}\mathbb{E}_y\mathbb{E}_{\xi_0}\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_z\left[\mathcal{I}(y|\sqrt{q_0}\xi_0+\sqrt{q_1-q_0}\xi_1+\sqrt{Q-q_1}z,\beta)\right]^{m_0}\right)\\ &= \frac{1}{m_0}\int \mathrm{d}yP_y(y)\int \mathrm{D}\xi_0\log\int \mathrm{D}\xi_1\left(\int dz\mathcal{N}_z\left(\sqrt{q_0}\xi_0+\sqrt{q_1-q_0}\xi_1,1-q_1\right)e^{-\beta V(y|z)}\right)^{m_0} \qquad \text{(E.38)}\\ &\simeq \frac{1}{m_0}\int \mathrm{d}yP_y(y)\int \mathrm{D}\xi_0\log\int \mathrm{D}\xi_1 e^{-m_0\beta\min_z\left[V(y|z)+\frac{1}{2\beta(1-q_1)}\left(z-\sqrt{q_0}\xi_0-\sqrt{q_1-q_0}\xi_1\right)^2\right]}\end{aligned}$$

Taking $q_1 \to 1$ with $\beta \to \infty$ and $m_0 \to 0$ in (E.32), and defining $\Omega_0 = \frac{\omega_0}{\chi}$, we obtain the 1RSB ground state energy

$$\begin{aligned}e_{\text{gs}}^{(1\text{RSB})} = \underset{\chi,\Omega_0,q_0}{\text{Extr}}\Bigg\{&\frac{1}{2\Omega_0\chi}\log\left(1+\Omega_0\Delta q\right)+\frac{q_0}{2\chi\left(1+\Omega_0\Delta q\right)} \qquad \text{(E.39)}\\ &+\frac{\alpha}{\chi\Omega_0}\mathbb{E}_{\xi_0}\log\mathbb{E}_{\xi_1}e^{-\Omega_0\chi\min_z\left[V(y|z)+\frac{1}{2\chi}\left(z-\sqrt{q_0}\xi_0-\sqrt{\Delta q}\xi_1\right)^2\right]}\Bigg\}.\end{aligned}$$

E.6.3 2RSB ground state energy $e_{\text{gs}}^{(2\text{RSB})}$

Taking $q_2 \to 1$ with $\beta \to \infty$, we define $\Omega_0 = \frac{m_0\beta}{\chi}, \Omega_1 = \frac{m_1\beta}{\chi}$, we obtain in the same fashion the 2RSB ground state energy of the spherical perceptron:

$$\begin{aligned}e_{\text{gs}}^{(2\text{RSB})} = \underset{\chi,\Omega_1,\Omega_0,q_1,q_0,}{\text{Extr}}\Bigg\{&\frac{q_0}{2\chi(1+\Omega_1(1-q_1)+\Omega_0(q_1-q_0)}+\frac{1}{2\Omega_1\chi}\log(1+\Omega_1(1-q_1)) \qquad \text{(E.40)}\\ &+\frac{\alpha}{\chi\Omega_0}\mathbb{E}_{\xi_0}\log\mathbb{E}_{\xi_1}\left[\mathbb{E}_{\xi_2}e^{-\Omega_1\chi\min_z\left[V(y|z)+\frac{1}{2\chi}\left(z-\sqrt{q_0}\xi_0-\sqrt{q_1-q_0}\xi_1-\sqrt{1-q_1}\xi_2\right)^2\right]}\right]^{\Omega_0/\Omega_1}\\ &\frac{1}{2\Omega_0\chi}\log\left(1+\frac{\Omega_0(q_1-q_0)}{1+\Omega_1(1-q_1)}\right)\Bigg\}\end{aligned}$$

Note that taking $q_1 = q_0, m_0 = m_1$ recovers the 1RSB expression.

Bibliography

[1] A. Abbara, B. Aubin, F. Krzakala, and L. Zdeborová. Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning. volume 107 of *Proceedings of Machine Learning Research*, pages 27–54. PMLR, 2020.

[2] A. Abbara, A. Baker, F. Krzakala, and L. Zdeborová. On the universality of noiseless linear estimation with respect to the measurement matrix. *Journal of Physics A: Mathematical and Theoretical*, 53(16):164001, 2020.

[3] A. Abbara, Y. Kabashima, T. Obuchi, and Y. Xu. Learning performance in inverse Ising problems with sparse teacher couplings. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(7):073402, 2020.

[4] M. S. Advani, A. M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 2020.

[5] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[6] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[7] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NeurIPS'18, pages 3223–3234. 2018.

[8] B. Aubin, W. Perkins, and L. Zdeborová. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 52(29):294003, 2019.

[9] E. Aurell and M. Ekeberg. Inverse Ising inference using all the data. *Phys. Rev. Lett.*, 108:090201, 2012.

[10] D. Azagra, J. Ferrera, F. López-Mesas, and Y. Rangel. Smooth approximation of lipschitz functions on riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 326(2):1370–1378, 2007.

[11] L. Bachschmid-Romano and M. Opper. Learning of couplings for random asymmetric kinetic Ising models revisited: random correlation matrices and learning curves. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(9):P09016, 2015.

[12] L. Bachschmid-Romano and M. Opper. A statistical physics approach to learning curves for the inverse Ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(6):063406, 2017.

[13] J. Barbier, M. Dia, N. Macris, and F. Krzakala. The mutual information in random linear estimation. In *54th Annual Allerton Conference on Communication, Control, and Computing*, pages 625–632, 2016.

[14] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[15] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[16] J. Barbier, N. Macris, A. Maillard, and F. Krzakala. The mutual information in random linear estimation beyond i.i.d. matrices. In *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018.

[17] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[18] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[19] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[20] M. Bayati and A. Montanari. The lasso risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[21] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[22] J. Berg. Statistical mechanics of the inverse Ising problem and the optimal objective function. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(8):083402, 2017.

[23] R. Berthier, A. Montanari, and P.-M. Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

[24] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011.

[26] T. Broderick, M. Dudik, G. Tkacik, R. E. Schapire, and W. Bialek. Faster solutions of the inverse pairwise Ising problem. *arXiv preprint arXiv:0712.2437*, 2007.

[27] B. Cakmak, O. Winther, and B. H. Fleury. S-amp: Approximate message passing for general matrix ensembles. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pages 192–196. IEEE, 2014.

[28] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[29] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[30] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[31] B. Çakmak, O. Winther, and B. H. Fleury. S-amp: Approximate message passing for general matrix ensembles. In *Information Theory Workshop (ITW), 2014 IEEE*, pages 192–196. IEEE, 2014.

[32] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[33] A. Chervonenkis and V. Vapnik. Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data. *Automation and Remote Control*, 32:207–217, 1971.

[34] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204, 2015.

[35] S. Cocco and R. Monasson. Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys. Rev. Lett.*, 106:090601, 2011.

[36] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[37] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

[38] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[39] J. R. de Almeida and D. J. Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 1978.

[40] A. Decelle and F. Ricci-Tersenghi. Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of Ising models. *Physical review letters*, 112(7):070603, 2014.

[41] J. Ding and N. Sun. Capacity lower bound for the Ising perceptron. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 816–827. ACM, 2019.

[42] D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[43] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[44] D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[45] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[46] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, 2005.

[47] A. Eddington. *The Nature of the Physical World*. Gifford lectures. Macmillan, 1928.

[48] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[49] A. Engel and C. Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[50] R. Erichsen and W. K. Thuemann. Optimal storage of a neural network model: a replica symmetry-breaking solution. *Journal of Physics A: Mathematical and General*, 26(2):L61–L68, 1993.

[51] R. Feynman. *Statistical Mechanics: A Set Of Lectures*. Advanced Books Classics. Avalon Publishing, 1998.

[52] A. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter. Expectation consistent approximate inference: Generalizations and convergence. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 190–194. IEEE, 2016.

[53] A. K. Fletcher, P. Pandit, S. Rangan, S. Sarkar, and P. Schniter. Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis. In *Advances in Neural Information Processing Systems*, pages 7440–7449, 2018.

[54] A. K. Fletcher, S. Rangan, and P. Schniter. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1884–1888. IEEE, 2018.

[55] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[56] M. Gabrié. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020.

[57] E. Gardner. The space of interations in neural network models. *J. Phys. A*, 21:257, 1988.

[58] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[59] E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983–1994, 1989.

[60] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[61] C. Gerbelot, A. Abbara, and F. Krzakala. Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices. volume 125 of *Proceedings of Machine Learning Research*, pages 1682–1713. PMLR, 2020.

[62] C. Gerbelot, A. Abbara, and F. Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula). *arXiv preprint arXiv:2006.06581*, 2020.

[63] J. W. Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge Library Collection - Mathematics. Cambridge University Press, 2010.

[64] P. Giselsson and S. Boyd. Linear convergence and metric selection for douglas-rachford splitting and admm. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2016.

[65] R. Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

[66] R. Gribonval, V. Cevher, and M. E. Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, 2012.

[67] F. Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

[68] A. Guionnet. Asymptotics of Harish-Chandra-Itzykson-Zuber integrals and of Schur polynomials. In *Large Random Matrices: Lectures on Macroscopic Asymptotics*, pages 211–216. Springer, 2009.

[69] D. Guo and S. Verdú. Randomly spread cdma: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, 2005.

[70] L. S. Hamilton, J. Sohl-Dickstein, A. G. Huth, V. M. Carels, K. Deisseroth, and S. Bao. Optogenetic activation of an inhibitory network enhances feedforward functional connectivity in auditory cortex. *Neuron*, 80(4):1066–1076, 2013.

[71] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[72] D. O. Hebb. *The organization of behavior: a neuropsychological theory*. Science Editions, 1962.

[73] J. Hertz and H. Schwarze. Generalization in large committee machines. *Physica A: Statistical Mechanics and its Applications*, 200(1-4):563–569, 1993.

[74] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[75] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[76] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[77] A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

[78] Y. Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. In *Journal of Physics: Conference Series*, volume 95, page 012001. IOP Publishing, 2008.

[79] Y. Kabashima and M. Vehkapera. Signal recovery using expectation consistent approximation for linear observations. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 226–230. IEEE, 2014.

[80] Y. Kabashima, T. Wadayama, and T. Tanaka. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

[81] H. J. Kappen and F. d. B. Rodríguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156, 1998.

[82] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE journal of selected topics in signal processing*, 1(4):606–617, 2007.

[83] S. G. Krantz and H. R. Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

[84] S. G. Krantz and H. R. Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2012.

[85] W. Krauth and M. Mézard. Storage capacity of memory networks with binary coupling. *J. Phys (France)*, 50:3057–3066, 1989.

[86] W. Krauth and M. Opper. Critical storage capacity of the J=±1 neural network. *Journal of Physics A: Mathematical and General*, 22(11):L519, 1989.

[87] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

[88] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[89] J. Kurzweil. On approximation in real Banach spaces. *Studia Mathematica*, 14(2):214–231, 1954.

[90] J. Le Guillou and J. Zinn-Justin. Critical exponents from field theory. *Physical Review B*, 21(9):3976, 1980.

[91] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[92] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[93] Z. Liao and R. Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, pages 3069–3077, 2018.

[94] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov. Optimal structure and parameter learning of Ising models. *Science advances*, 4(3):e1700791, 2018.

[95] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[96] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing MRI. *IEEE signal processing magazine*, 25(2):72–82, 2008.

[97] J. Ma and L. Ping. Orthogonal amp. *IEEE Access*, 5:2020–2033, 2017.

[98] P. Majer, A. Engel, and A. Zippelius. Perceptrons above saturation. *Journal of Physics A: Mathematical and General*, 26(24):7405–7416, 1993.

[99] S. Mandelbrojt. Classes of infinitely differentiable functions. *Rice Institute Pamphlet-Rice University Studies*, 29(1), 1942.

[100] A. Manoel, F. Krzakala, G. Varoquaux, B. Thirion, and L. Zdeborová. Approximate message-passing for convex optimization with non-separable penalties. *arXiv preprint arXiv:1809.06304*, 2018.

[101] D. W. Marquardt and R. D. Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

[102] P. Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

[103] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[104] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181–2190, 1989.

[105] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

[106] M. Mézard, G. Parisi, and M. Virasoro. Sk model: The replica solution without replicas. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, 9(2):232, 1987.

[107] M. Mézard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond*. Lecture Notes in Physics Series. World Scientific, 1987.

[108] M. Mézard and J. Sakellariou. Exact mean-field inference in asymmetric kinetic Ising systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(07):L07001, 2011.

[109] T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, 2001.

[110] T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[111] G. J. Mitchison and R. M. Durbin. Bounds on the learning capacity of some multi-layer networks. *Biological Cybernetics*, 60(5):345–365, 1989.

[112] P. P. Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation. *CoRR*, abs/1906.03667, 2019.

[113] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[114] H. Monajemi, S. Jafarpour, M. Gavish, , and D. L. Donoho. Deterministic matrices matching the compressed sensing phase transitions of Gaussian random matrices. *Proceedings of the National Academy of Sciences*, 110(4):1181–1186, 2013.

[115] R. Monasson and R. Zecchina. Weight space structure and internal representations: A direct approach to learning and generalization in multilayer neural networks. *Physical Review Letters*, 75(12):2432–2435, 1995.

[116] H. C. Nguyen and J. Berg. Bethe–Peierls approximation and the inverse Ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03004, 2012.

[117] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan. A general analysis of the convergence of admm. In *International Conference on Machine Learning*, pages 343–352, 2015.

[118] M. Opper. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pages 922–925, 1995.

[119] M. Opper and W. Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

[120] M. Opper and D. Saad. *Advanced Mean Field Methods : Theory and Practice*. Neural information processing series. MIT Press, 2001.

[121] M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177, 2005.

[122] N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[123] G. Parisi. *Physical Letters*, (73A):203, 1979.

[124] G. Parisi. The order parameter for spin glasses: a function on the interval 0-1. *Journal of Physics A: Mathematical and General*, 13(3):1101, 1980.

[125] D. W. Peaceman and H. H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for industrial and Applied Mathematics*, 3(1):28–41, 1955.

[126] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[127] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

[128] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[129] M. Ramezanali, P. P. Mitra, and A. M. Sengupta. Mean field analysis of sparse reconstruction with correlated variables. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1267–1271, 2016.

[130] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

[131] S. Rangan, V. Goyal, and A. K. Fletcher. Asymptotic analysis of map estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems*, pages 1545–1553, 2009.

[132] S. Rangan, P. Schniter, and A. K. Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

[133] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[134] G. Reeves and H. D. Pfister. The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 665–669, 2016.

[135] F. Ricci-Tersenghi. The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08015, 2012.

[136] Y. Roudi and J. Hertz. Mean field theory for nonequilibrium network reconstruction. *Phys. Rev. Lett.*, 106:048702, 2011.

[137] C. Rush, A. Greig, and R. Venkataramanan. Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Transactions on Information Theory*, 63(3):1476–1500, 2017.

[138] C. Rush and R. Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.

[139] N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[140] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[141] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[142] K. Schwab. *The Fourth Industrial Revolution*. Crown Publishing Group, USA, 2017.

[143] H. Schwarze. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21):5781, 1993.

[144] V. Sessak and R. Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42(5):055001, 2009.

[145] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[146] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[147] K. Sharp and F. Matschinsky. Translation of Ludwig Boltzmann's paper "On the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium" 1909. *Entropy*, 17:1971–2009, 2015.

[148] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, 1975.

[149] T. Shinzato and Y. Kabashima. Perceptron capacity revisited: Classification ability for correlated patterns. *Journal of Physics A: Mathematical and Theoretical*, 41(32), 2008.

[150] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.*, 26(32):8254–8266, 2006.

[151] P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558, 2019.

[152] T. Takahashi and Y. Kabashima. Macroscopic analysis of vector approximate message passing in a model mismatch setting. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1403–1408, 2020.

[153] K. Takeda, S. Uda, and Y. Kabashima. Analysis of CDMA systems that are characterized by eigenvalue spectrum. *EPL (Europhysics Letters)*, 76(6):1193, 2006.

[154] M. Talagrand. *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer Science & Business Media, 2003.

[155] M. Talagrand. The Parisi formula. *Annals of mathematics*, pages 221–263, 2006.

[156] T. Tanaka. Mean-field theory of Boltzmann machine learning. *Physical Review E*, 58(2):2302, 1998.

[157] T. Tanaka. Statistical mechanics of CDMA multiuser demodulation. *EPL (Europhysics Letters)*, 54(4):540, 2001.

[158] T. Tanaka. Asymptotics of Harish-Chandra-Itzykson-Zuber integrals and free probability theory. *Journal of Physics: Conference Series*, 95:012002, 2008.

[159] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. L. Smith, H. Patel, A. Prieto, D. Petrusca, M. I. Grivich, A. Sher, P. Hottowy, W. Dabrowski, A. M. Litke, and J. M. Beggs. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.*, 28(2):505–518, 2008.

[160] G. Tavoni, U. Ferrari, F. P. Battaglia, S. Cocco, and R. Monasson. Inferred model of the prefrontal cortex activity unveils cell assemblies and memory replay. *bioRxiv*, 2015.

[161] Y. Terada, T. Obuchi, T. Isomura, and Y. Kabashima. Objective and efficient inference for couplings in neuronal networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 4976–4985, Red Hook, NY, USA, 2018. Curran Associates Inc.

[162] Y. Terada, T. Obuchi, T. Isomura, and Y. Kabashima. Objective and efficient inference for couplings in neuronal networks. In *Advances in Neural Information Processing Systems 31*, pages 4971–4980. Curran Associates, Inc., 2018.

[163] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[164] C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[165] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[166] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry. Searching for collective behavior in a large network of sensory neurons. *PLoS Comput. Biol.*, 10(1):e1003408, 2014.

[167] Y. Tsaig and D. L. Donoho. Extensions of compressed sensing. *Signal processing*, 86(3):549–571, 2006.

[168] A. M. Tulino, G. Caire, S. Verdu, and S. Shamai. Support recovery with sparsely sampled free random matrices. *IEEE Transactions on Information Theory*, 59(7):4243–4271, 2013.

[169] A. M. Tulino, S. Verdú, et al. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182, 2004.

[170] R. Urbanczik. Storage capacity of the fully-connected committee machine. *Journal of Physics A: Mathematical and General*, 30(11):L387, 1997.

[171] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[172] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[173] M. Vuffray, S. Misra, and A. Y. Lokhov. Efficient learning of discrete graphical models. *CoRR*, abs/1902.00600, 2019.

[174] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[175] P. Walters. *An introduction to ergodic theory*, volume 79. Springer Science & Business Media, 2000.

[176] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda. A pairwise maximum entropy model accurately describes resting-state human brain networks. *Nat. Commun.*, 4(May 2012):1370, 2013.

[177] T. Watanabe, N. Masuda, F. Megumi, R. Kanai, and G. Rees. Energy landscape and dynamics of brain activity during human bistable perception. *Nat. Commun.*, 5, 2014.

[178] T. L. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

[179] A. B. Watson. Image compression using the discrete cosine transform. *Mathematica journal*, 4(1):81, 1994.

[180] W. Whyte and D. Sherrington. Replica-symmetry breaking in perceptrons. *Journal of Physics A: Mathematical and General*, 29(12):3063–3073, 1996.

[181] Y. Xiong, C. Kwon, and J.-H. Oh. The storage capacity of a fully-connected committee machine. In *Advances in Neural Information Processing Systems*, pages 378–384, 1998.

[182] Y. Xu, S. Puranen, J. Corander, and Y. Kabashima. Inverse finite-size scaling for high-dimensional significance analysis. *Physical Review E*, 97(6):062112, 2018.

[183] J. S. Yedidia, W. T. Freeman, and Y. Weiss. *Understanding Belief Propagation and Its Generalizations*, page 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

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

[185] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *ICLR 2017, preprint arXiv:1611.03530*, 2017.

[186] J. Zhu and D. Baron. Performance regions in compressed sensing from noisy measurements. In *2013 47th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2013.

[187] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

RÉSUMÉ

Au cours de la dernière décennie, les techniques d'apprentissage automatique ont connu de formidables progrès, et donnent lieu à de nombreuses applications. Elles sont néanmoins très difficiles à analyser, car elles impliquent l'utilisation de réseaux de neurones profonds sur des données réelles, régis par un nombre énorme de paramètres. La physique statistique s'est depuis longtemps attaquée à l'étude des réseaux de neurones et des problèmes d'inférence, ce dès les années 80, se penchant d'abord sur des modèles simplifiés avec données aléatoires. Les algorithmes actuels gagnant en performance, un renouveau d'intérêt a secoué la communauté physique, qui s'est de nouveau attablée à leur étude ; s'efforçant de fournir des piliers de compréhension théorique solide à travers des problèmes synthétiques qui décrivent les cas les plus probables. Les physiciens emploient en particulier des méthodes heuristiques développées dans le domaine des verres de spin, telle la méthode des répliques. Dans cette thèse, nous approchons plusieurs problèmes à travers un formalisme d'inférence Bayesienne. Un premier résultat physique est obtenu pour le problème inverse d'Ising, dans le cas d'un réseau enseignant à poids épars - en grande partie nuls. Nous nous tournons ensuite vers l'acquisition comprimée, et observons que plusieurs classes de matrices structurées partagent les mêmes transitions dans le cas d'une reconstruction sans bruit, et nous expliquons ce phénomène pour les matrices invariantes par rotation à droite. Nous exploitons le lien entre physique statistique et algorithmes de passage de messages pour démontrer la formule des répliques qui caractérise la performance optimale de reconstruction pour une régression linéaire avec pénalité convexe. Nous étendons ce résultat au modèle linéaire généralisé, qui incorpore des non-linéarités et décrit un réseau de neurones à deux couches. Ces deux résultats concernent des matrices invariantes par rotation, dépassant ainsi l'hypothèse commune de données identiquement et indépendemment distribuées, et permettant d'incorporer des corrélations entre données. Enfin, nous montrons que la complexité de Rademacher, qui fournit un encadrement de l'écart de généralisation dans le pire des cas pour des problèmes de classification binaire, est intimement liée à l'énergie libre fondamentale du problème physique correspondant, et peut être calculée dans certains cas.

MOTS CLÉS

Physique statistique, systèmes désordonnés, apprentissage automatique, méthode des répliques, algorithmes de passage de messages, optimisation convexe.

ABSTRACT

In the last decade, machine learning techniques have achieved tremendous progresses and yield many applications. However, they are very hard to analyze, due to the huge number of parameters involved in deep networks dealing with real-world data. Statistical physics have a long-standing tradition of studying neural networks and inference problems, starting in the 80s, that initially focused on simplified models with random data. As algorithms recently became more efficient, the physics community witnessed a renewal of interest in the topic, attempting to provide solid theoretical foundations by studying synthetic problems and describing the typical, most probable case. In particular, physicists use heuristic methods developed in the spin glass field, such as the replica method. In this thesis, we approach several problems through a Bayesian inference setting. A first physical result is derived for the inverse Ising problem with sparse teacher weights. We then consider compressed sensing and observe that several classes of structured matrices share the same phase transitions, in the noiseless reconstruction case, and provide an explanation in the case of right rotationally invariant matrices. We build on the link between statistical physics and message passing algorithms, to prove the replica result characterizing the optimal reconstruction performance for linear regression with convex penalty. We also extend this result to the generalized linear model, which adds non-linearity and describes a two-layer neural network. Both results tackle the case of rotationally invariant data matrices, which goes beyond the usual assumption of identically and independently distributed data, and allows for correlated patterns. Finally, we show that the Rademacher complexity, that provides a worst-case bound for the generalization gap in classification problems, shares a deep connection with the ground state free energy of the corresponding physical problem, and can be computed in some settings.

KEYWORDS

Statistical physics, disordered systems, machine learning, replica method, message passing algorithms, convex optimization.