

ON THE TYPICAL PROPERTIES OF INVERSE
PROBLEMS IN STATISTICAL MECHANICS



IACOPO MASTROMATTEO

A DISSERTATION

PRESENTED TO THE FACULTY OF SISSA

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

ADVISER: PROFESSOR MATTEO MARSILI

SEPTEMBER 2012

© Copyright by Iacopo Mastromatteo, 2012.

All rights reserved.

Abstract

In this work we consider the problem of extracting a set of interaction parameters from an high-dimensional dataset describing T independent configurations of a complex system composed of N binary units. This problem is formulated in the language of statistical mechanics as the problem of finding a family of couplings compatible with a corresponding set of empirical observables in the limit of large N . We focus on the typical properties of its solutions and highlight the possible spurious features which are associated with this regime (model condensation, degenerate representations of data, criticality of the inferred model). We present a class of models (complete models) for which the analytical solution of this inverse problem can be obtained, allowing us to characterize in this context the notion of stability and locality. We clarify the geometric interpretation of some of those aspects by using results of differential geometry, which provides means to quantify consistency, stability and criticality in the inverse problem. In order to provide simple illustrative examples of these concepts we finally apply these ideas to datasets describing two stochastic processes (simulated realizations of a Hawkes point-process and a set of time-series describing financial transactions in a real market) .

List of previously published work

1. MASTROMATTEO, I.
Beyond inverse Ising model: structure of the analytical solution for a class of inverse problems. *Arxiv preprint: arXiv:1209.1787* (2012).
2. BARATO A.C., MASTROMATTEO I., BARDOSCIA M. AND MARSILI M.
Impact of meta-order in the minority game. Submitted to *Quant. Finance*.
Arxiv preprint: arXiv:1112.3908 (2012).
3. MASTROMATTEO I., ZARINELLI E. AND MARSILI M.
Reconstruction of financial network for robust estimation of systemic risk. *J. Stat. Mech.* **P03011** (2011).
4. MASTROMATTEO I. AND MARSILI M.
On the criticality of inferred models. *J. Stat. Mech.* **P10012** (2011).
5. MASTROMATTEO I., MARSILI M. AND ZOI P.
Financial correlations at ultra-high frequency: theoretical models and empirical estimation. *Eur. Phys. J. B*, **80** (2) 243-253 (2011).

Chapter 2 has an introductory purpose, and contains mainly non-original work. Chapter 3 presents original results not yet published. Chapter 4 discusses the ideas behind article number 1, while Chapter 5 covers exhaustively the content of article number 4. The subject of the remaining manuscripts has not been included in this thesis.

Acknowledgements

I am truly indebted with my advisor M. Marsili for encouraging me to spend three years studying beautiful and challenging problems. I have been consistently borrowing his advices and profiting of his skills, which he never refused to share. I thank A.C. Barato, C. Battistin, M. Bardoscia, E. Zarinelli and P. Zoi, with whom I had the pleasure to collaborate with during the course of the PhD. I especially thank André for his efforts in coercing this thesis into an almost readable form. I acknowledge the former and the current members of M. Marsili group (M. Bardoscia, F. Caccioli, L. Caniparoli, L. Dall'Asta, G. De Luca, D. De Martino, G. Gori, G. Livan, P. Vivo), for all the interesting discussions we had and for the exceedingly long time we spent in the ICTP cafeteria. I thank my classmates A. De Luca and J. Viti (with whom I'm going to share another part of academic life), together with F. Buccheri, L. Foini, F. Mancarella and X. Yu. I thank M. Masip for his constant disposability and his wise advices. I would also like to thank all the persons from outside SISSA and ICTP with whom I had the opportunity to have valuable and stimulating interactions all over these years: M. Alava, F. Altarelli, E. Aurell, J.P. Bouchaud, A. Braunstein, S. Cocco, A. Codello, S. Franz, A. Kirman, F. Lillo, Y. Roudi, B. Tóth, R. Zecchina.

On a personal note, I would like to thank my parents for all their support. Finally, I thank Najada for the years we spent together, and for her decision to follow me during the forthcoming ones.

Contents

Abstract	iii
List of previously published work	iv
Acknowledgements	v
1 Introduction	1
2 Binary Inference	5
2.1 The direct problem	6
2.1.1 Statistical model	7
2.1.2 Entropy and Kullback-Leibler divergence	10
2.1.3 Observables	12
2.1.4 Small and large deviations	14
2.2 The inverse problem	15
2.2.1 Bayesian formulation	15
2.2.2 Maximum likelihood criteria	16
2.2.3 Statement of the inverse problem	17
2.2.4 Small and large deviations	19
2.2.5 Examples	21
2.3 The regularized inverse problem	24
2.3.1 Bayesian formulation	26
2.3.2 Two popular regularizers	26

2.3.3	Examples	29
3	High-dimensional inference	32
3.1	Computational limitations and approximate inference schemes	33
3.1.1	Boltzmann Learning	33
3.1.2	Mean field approaches for pairwise models	34
3.2	The large N , finite M regime	40
3.3	Fully-connected ferromagnet	42
3.3.1	The mean-field solution	43
3.3.2	Finite N corrections	45
3.4	Saddle-point approach to mean-field systems	48
3.4.1	Ergodicity breaking for a fully connected pairwise model	52
3.5	Disorder and heterogeneity: the regime of large N and large M	54
3.5.1	Self-averaging properties and inverse problem	55
4	Complete representations	57
4.1	Orthogonality and completeness	58
4.2	Inference on complete models	60
4.2.1	The complete inverse problem	60
4.2.2	Regularization of the complete inverse problem	65
4.2.3	Pairwise model on trees	75
4.2.4	One-dimensional periodic chain with arbitrary range couplings	78
4.3	Applications	81
4.3.1	Complete inverse problem	81
4.3.2	L-1 norm vs L-2 norm: emergence of state symmetry breaking	85
4.3.3	Pairwise model on a tree	88
4.3.4	One-dimensional periodic chain	91

5	Information geometry and criticality	93
5.1	Metric structure of the probability space	94
5.1.1	Fisher information as a metric	94
5.1.2	Sanov theorem and distinguishable distributions	97
5.1.3	Complexity measures and criticality	101
5.1.4	Examples	104
5.2	Inference of a non-equilibrium model	107
5.2.1	The Hawkes process	109
5.2.2	Trades in a financial market	114
5.3	Applications	116
5.3.1	Pairwise fully-connected model for Hawkes processes	117
5.3.2	Pairwise fully-connected model for NYSE trade events	121
6	Conclusion	127
A	Binary Inference	129
A.1	Maximum entropy principle	129
A.2	Concavity of the free energy	130
A.3	Small deviations of the empirical averages	131
A.4	Sanov theorem	131
A.5	Cramér-Rao bound	134
A.6	Convergence of the inferred couplings	135
B	High-dimensional inference	137
B.1	The fully-connected ferromagnet: saddle-point calculation	137
B.1.1	The leading contribution F_0	138
B.1.2	Transition line and metastability	140
B.1.3	Marginal polytope for a fully connected ferromagnet	142

C	Convex optimization	143
C.1	Differentiable target	143
C.1.1	Gradient descent algorithm	144
C.2	Non-differentiable target	146
C.2.1	Sub-gradient descent algorithm	148
D	Complete families	150
D.1	Rate of convergence for the complete inverse problem	150
D.2	Factorization property for tree-like models	151
D.3	Factorization property of the one-dimensional periodic chain	153
E	Geometry	156
E.1	Geodesics	156
E.2	Property of the maximum likelihood estimator	157
E.3	Expansion of the Kullback-Leibler divergence	159
E.4	Volume of indistinguishability	159
E.5	Estimation of the empirical observables for an Hawkes point process	161
	Bibliography	167
	Notation	174

Chapter 1

Introduction

The generations living during the last twenty or thirty years witnessed a huge scientific revolution which has been, essentially, technology driven. An impressive amount of computational power became cheaply available for people and institutions, while at the same time the quantity of data describing many aspects of our world started to grow in a seemingly unbound fashion: the human genome can be efficiently sequenced in some days [75, 88], the interactions among proteins in a human body can in principle be enumerated one-by-one [69], financial transactions are recorded with resolutions well below one second [1], the dynamics of networks of all kinds (social, economics, neural, biological) can be tracked in real-time. Parallel to this, the widely accepted scientific paradigm according to which it is necessary to ground reliable models on solid first principles started to crumble: promising results evidenced that it is possible to extract accurate statistical models from empirical datasets without even trying to guess what is their underlying structure, nor to characterize which input-output relations govern their behavior. Large datasets can be automatically and faithfully *compressed* in small sets of coefficients [31], their features can be *described* accurately with unsupervised algorithms, new data can be *predicted* with a given degree of accuracy on the basis of the older one (see for example [3]). Google

uses pattern recognition and Bayesian techniques to translate from one language to the other regardless of the formal rules of the grammar [2], and Netflix can predict how much you will rate a movie (one to five) with an error around 0.85 without knowing anything about you but a few of your former preferences [4]. The embarrassing success of this approach compels a basic epistemological question about modeling: does an approach based solely on statistical learning lead to any actual understanding? What does one learn about a system when processing data in this way?

This problem is particularly relevant when dealing with the task of high-dimensional inference, in which a typically large set of parameters is extracted from an even larger dataset of empirical observations. What meaning has to be associated with each of the many parameters extracted from data? Are there *combinations* of such numbers describing global, macroscopic features of the system? A prototypical example is provided by the study of networks of neurons, in which one would like to understand how the brain works (e.g., the presence of collective states of the network, the possibility to store and retrieve informations) by processing data describing the behavior of a huge set of elementary units (the neurons). This task can be thought of as a seemingly hopeless one: in a way it is similar to reverse-engineering how a laptop works by probing the electric signal propagating through its circuitry. A modern answer to this type of arguments is the idea that if data is sufficient and the inference algorithm is good enough, some of the actual features of the system will eventually be detected. In the case of a laptop, one can think to extract from data not only the wiring pattern a set of cables, but to detect collective features such as the fact that a computer is an essentially deterministic object (in contrast to biological networks, where fluctuations are essential), or that it possesses multiple collective states (say, switched-on, switched-off or sleepy).

Physics, and in particular statistical mechanics, has much to do with all of this picture for two main reasons. The first one is technical: while the high-dimensional limit is

a relatively new regime in the field of statistical inference, statistical mechanics has since long developed mathematical descriptions of systems composed by a very large (or better, infinite) number of interacting components [40]. Hence, mapping problems of statistical inference onto problems of statistical mechanics opens the way to a remarkable amount of mathematical machinery which can be used to solve quickly and accurately problems which become very complicated for large systems [45, 79]. This is even more true since the study of heterogeneous and glassy materials produced sophisticated tools (replica trick, cavity methods) suitable to study systems in which no apparent symmetry or regularity is present, as often found in data describing complex systems [56]. The second, and more philosophical, reason is that statistical mechanics is naturally built to explain collective behaviors on the basis of individual interactions. Just as the ideal gas can be understood by studying the aggregate behavior of many non-interacting particles, or the emergence of spontaneous magnetization can be derived by studying the interactions of single spins, statistical mechanics can be used to predict the collective behavior of biological, social and economic systems starting from a given set of rules describing the interaction of some fundamental units [30]. In 1904 Ludwig Boltzmann, almost a century before anyone could take him literally, anticipated that

“The wide perspectives opening up if we think of applying this science to the statistics of living beings, human society, sociology and so on, instead of only to mechanical bodies, can here only be hinted at in a few words.”

Hence, from the perspective of (large-scale) statistical learning, it is natural to use statistical mechanics methods to study the emergence of collective properties of a system once the microscopic interactions of the fundamental units have been reconstructed through a careful analysis of empirical data.

Unfortunately, even if one is able to do that, it is not always easy to understand how much of the inferred model faithfully describes the system: it is possible, and

it is often the case, that the procedure which is used to perform the data analysis influences so much the outcome that the actual properties of the system get lost along the way, and the inferred model shows a spurious behavior determined just by the fitting procedure. For example, models with binary interactions may describe very well systems in which the interaction is actually multi-body [39], just as critical models (strongly fluctuating statistical systems) may fit random set of observables much better than ordinary ones [53]. Noise itself may be fitted very well by sophisticated models, while non-stationary systems might be accurately described by using equilibrium distributions [84]. In all of these cases, it is important to develop quantitative tools which allow to distinguish between spurious features of the inferred model and genuine ones.

The purpose of this work is precisely to inquire some of those aspects in the simpler setting in which we consider a statistical system consisting in a string of N binary variables, used to model T independently drawn configurations. We will show that, while the small N regime the problem of inference can be completely controlled (chapter 2), the large N regime becomes computationally intractable and non-trivial collective properties may emerge (chapter 3). Such features be observed independently of the data, and have to be associated uniquely with the properties of the model which is used to perform the inference, regardless of the system which one is trying to describe. In chapter 4 we will show under which conditions the problem of inferring a model is easy, showing in some cases its explicit solution. We will also evidence the limits of non-parametric inference, highlighting that for under-sampled systems correlations might be confused with genuine interactions. In chapter 5 we will provide a geometric interpretation for the problem of inference, showing a metric which can be used to meaningfully assess the collective phase of an inferred system. We will apply these ideas to two datasets, describing extensively the results of their analysis in the light of our approach.

Chapter 2

Binary Inference

In this chapter we will describe the problem of extracting information from empirical datasets describing a stationary system composed of a large number of interacting units. Interestingly, this problem has almost simultaneously received a great deal of attention from the literature of diverse communities (biology [77, 87], genetics [22], neuroscience [72, 76, 24], economy, finance [48, 59, 29], sociology). This can be traced back to two main reasons: first, it is now possible across many fields to analyze the synchronous activity of the components of a complex system (e.g., proteins in a cell, neurons in the brain, traders in a financial market) due to technological advantages either in the data acquisition procedures or in the experimental techniques used to probe the system. Secondly, data highly resolved in time is often available, which (beyond implying that finer time-scales can be explored) provides researchers with a large number of observations of the system. Defining as N the number of components of the system and as T the number of available samples, these last observations can be summarized by asserting that the limit of large N and large T can be accessed for a large number of complex systems. In this work we will restrict ourselves to the more specific case in which such systems are described by binary units, reminding to the reader that (i) most of what will be shown can be generalized to the case of

non-binary (Potts) or continuous variables [85] and (ii) the binary case already allows to describe in detail several systems [72, 76, 24]. In section 2.1 we describe the models that we consider, which usually go under the name of *exponential families* and are justified on the basis of the maximum entropy principle (appendix A.1), and state the *direct problem*, alias the calculation of the observables given the model. In section 2.2 we present the problem of inferring a model from data (the *inverse problem*) and characterize it as the Legendre-conjugated of the direct one. In section 2.3 we present the *regularization* techniques which can be used to cure the pathological behavior of some inverse problems and improve their generalizability. Although the results presented in this chapter are far from being original, we aim to show as transparently as possible the deep connections between information theory and statistical mechanics, emphasizing the strong analogy between direct and inverse problems.

2.1 The direct problem

We introduce in this section the *direct problem* – which deals with finding the observables associated with a given statistical model – as a preliminary step towards the formulation of an inference problem. This is the problem typically considered by statistical mechanics, hence we will adopt most of the terminology and the notation from this field. The main results that we will present are associated with the *free energy* – which we use in order to generate the averages and the covariances of the model – and to its relations with the notion of Kullback-Leibler divergence and the one of Shannon entropy. Finally, we will characterize the large and small deviation properties of the empirical averages of the observables under the model.

2.1.1 Statistical model

We consider a system of N binary spins $s = (s_1, \dots, s_N) \in \{-1, 1\}^N = \Omega$, indexed by $i \in V = \{1, \dots, N\}$. A *probability density* \mathbf{p} is defined as any positive function $\mathbf{p} : \Omega \rightarrow \mathbb{R}$ such that $\sum_s p(s) = 1$, while the space of all possible probability densities on Ω is denoted as $\mathcal{M}(\Omega)$. We also consider a families of real-valued functions $\phi : \Omega \rightarrow \mathbb{R}^{|\phi|}$ with components $\phi(s) = (\phi_1(s), \dots, \phi_{|\phi|}(s))$, which will be referred as *binary operators*, and are more commonly known in the literature of statistical learning as *sufficient statistics* or *potential functions* [85], and will be used in order to construct a probability density on the configuration space of the system.

Definition 2.1. Given a set of binary operators $\phi = \{\phi_\mu\}_{\mu=1}^M$ and a vector of real numbers $\mathbf{g} = \{g_\mu\}_{\mu=1}^M$ a *statistical model* is defined as the pair (ϕ, \mathbf{g}) . Its associated probability density $\mathbf{p} = (p_s)_{s \in \Omega}$ is given by

$$p(s) = \frac{1}{Z(\mathbf{g})} \exp \left(\sum_{\mu=1}^M g_\mu \phi_\mu(s) \right), \quad (2.1)$$

whereas the normalization constant $Z(\mathbf{g})$ is defined as

$$Z(\mathbf{g}) = \sum_s \exp \left(\sum_{\mu=1}^M g_\mu \phi_\mu(s) \right) \quad (2.2)$$

and is referred as the *partition function* of the model. The *free energy* $F(\mathbf{g})$ is defined as $F(\mathbf{g}) = -\log Z(\mathbf{g})$.

For conciseness, the identity operator will always be labeled as the zero operator $\phi_0(s) = 1$, in order to reabsorb the normalization constant $Z(\mathbf{g})$ into its conjugated coupling g_0 . The probability density will be written as $p(s) = p_s$, so that (2.1) will be compactly written as

$$p_s = \exp \left(\sum_{\mu=0}^M g_\mu \phi_{\mu,s} \right). \quad (2.3)$$

With these definitions, the coupling g_0 results equal to the free energy $g_0 = -\log Z(\mathbf{g}) = F(\mathbf{g})$. Given a family of operators ϕ , we also denote as $\mathcal{M}(\phi)$ the set of all the statistical models of the form (2.1) obtained by varying the coupling vector \mathbf{g} . Given the probability density (2.1) and a generic subset $\Gamma \subseteq V$ (which we call a *cluster*), we also define the *marginal* $p^\Gamma(s^\Gamma)$ as

$$p^\Gamma(s^\Gamma) = \sum_{s_i | i \notin \Gamma} p(s), \quad (2.4)$$

which expresses the probability to find spins belonging to the Γ in a given configuration once the degrees of freedom associated with spins outside such cluster have been integrated out (whereas $p^\emptyset = 1$ and $p^V(s) = p(s)$).

This construction will be used to study inference problems in which the M operators $\phi(s)$ are *a priori* known. We will disregard for the moment the issue of optimally selecting the most appropriate operators in order to describe a given set of data, an important problem known as *model selection*. Let us indeed remind that models of the form (2.1) can be justified on the basis of the maximum entropy principle, which will be stated in appendix A.1. The next notions which will be defined are the one of ensemble average and the one of susceptibility which will be extensively used throughout our discussion.

Definition 2.2. Given a statistical model (ϕ, \mathbf{g}) of the form (2.1), we define the *ensemble average* of an operator ϕ_μ as the quantity

$$\langle \phi_\mu \rangle = \sum_s \phi_{\mu,s} p_s, \quad (2.5)$$

while the *generalized susceptibility* matrix $\hat{\chi}$ is defined as the covariance matrix whose elements are given by

$$\chi_{\mu,\nu} = \langle \phi_\mu \phi_\nu \rangle - \langle \phi_\mu \rangle \langle \phi_\nu \rangle. \quad (2.6)$$

Beyond describing fluctuations around the ensemble average of the ϕ operators, the generalized susceptibility $\hat{\chi}$ is a fundamental object in the field of information theory [27], in whose context is more often referred as *Fisher information*, and is more commonly defined as

$$\chi_{\mu,\nu} = - \left\langle \frac{\partial^2 \log p_s}{\partial g_\mu \partial g_\nu} \right\rangle. \quad (2.7)$$

Its relevance in the field of information theory and statistical learning will later be elucidated by properties (2.23) and (2.24) which concern with the direct problem. Sanov thorem (2.35), Cramér-Rao bound (2.38), together with equations (2.36) and (2.37), clarify its role in the context of the inverse problem.

Proposition 2.1. *The free energy function enjoys the properties*

$$\langle \phi_\mu \rangle = - \frac{\partial F}{\partial g_\mu} \quad (2.8)$$

and

$$\chi_{\mu,\nu} = - \frac{\partial^2 F}{\partial g_\mu \partial g_\nu}, \quad (2.9)$$

thus it is the generating function of the averages and of the fluctuations of the operators ϕ_μ contained in the model.

Equation (2.9) implies that covariances $\chi_{\mu,\nu}$ are related to the response of the ensemble averages with respect to changes of the couplings through

$$\chi_{\mu,\nu} = \langle \phi_\mu \phi_\nu \rangle - \langle \phi_\mu \rangle \langle \phi_\nu \rangle = \frac{\partial \langle \phi_\mu \rangle}{\partial g_\nu}, \quad (2.10)$$

a relation known as *fluctuation-dissipation* relation, which is a direct consequence of the stationary nature of the probability distribution (2.1). Another fundamental property of the free energy function $F(\mathbf{g})$ is its *concavity*, which will later allow us to

relate the field of statistical inference with the one of convex optimization (appendix C). It can be shown (appendix A.2) that:

Proposition 2.2.

- The susceptibility matrix $\hat{\chi}$ is a positive semidefinite matrix, thus the free energy $F(\mathbf{g})$ is a concave function.
- If the family of operators ϕ is minimal (i.e. it doesn't exist a non-zero vector \mathbf{x} such that $\sum_{\mu} x_{\mu} \phi_{\mu,s}$ is constant in s), then the susceptibility matrix $\hat{\chi}$ is strictly positive definite and the free energy $F(\mathbf{g})$ is strictly concave.

Definition 2.3. Given a statistical model (ϕ, \mathbf{g}) of the form (2.1), the *direct problem* is defined as the calculation of the free energy $F(\mathbf{g})$, of the averages $\langle \phi \rangle$ and of the susceptibility matrix $\hat{\chi}$ as functions of the coupling vector \mathbf{g} .

2.1.2 Entropy and Kullback-Leibler divergence

In this section we will define the concept of Shannon entropy, which will be used as an information theoretic measure of the information content of a distribution.

Definition 2.4. Given a probability density \mathbf{p} , we define the Shannon entropy $S(\mathbf{p})$ as the function

$$S(\mathbf{p}) = - \sum_s p_s \log p_s \tag{2.11}$$

The quantity $S(\mathbf{p})$ measures the amount of disorder associated with the random variable s , and satisfies the following properties:

- $0 \leq S(\mathbf{p}) \leq \log |\Omega|$. In particular $S(\mathbf{p}) = 0$ for $p(s) = \delta_{s,s'}$ (when the variable s is maximally informative), while $S(\mathbf{p}) = \log |\Omega|$ for the flat case $p(s) = 1/|\Omega|$ (in which s is maximally undetermined).
- The function $S(\mathbf{p})$ is concave in \mathbf{p} .

They can be proven straightforwardly, as for example in [27]. Another information-theoretic notion which will be extensively used is the Kullback-Leibler divergence $D_{KL}(\mathbf{p}|\mathbf{q})$, which characterizes the distance between two probability distributions. Although it doesn't satisfy the symmetry condition nor the triangular inequality required to define a proper measure of distance, in chapter 5 we will show that indeed a rigorous concept of distance can be extracted by means of the Kullback-Leibler divergence.

Definition 2.5. Given a pair of probability densities \mathbf{p} and \mathbf{q} , the *Kullback-Leibler* divergence $D_{KL}(\mathbf{p}||\mathbf{q})$ is defined as

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_s p_s \log \frac{p_s}{q_s} \quad (2.12)$$

Such quantity enjoys the following properties:

- $D_{KL}(\mathbf{p}||\mathbf{q}) \geq 0$ for any pair of probability densities \mathbf{p}, \mathbf{q} .
- $D_{KL}(\mathbf{p}||\mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$.
- $D_{KL}(\mathbf{p}||\mathbf{q})$ is a convex function in both \mathbf{p} and \mathbf{q} .

These property justify the role played by the Kullback-Leibler divergence in information theory, and can be proven straightforwardly (see [27]). Notice indeed that given two statistical models (ϕ, \mathbf{g}) and (ϕ, \mathbf{g}') respectively associated with densities \mathbf{p} and \mathbf{p}' , the entropy and the Kullback-Leibler divergence can be written as

$$S(\mathbf{p}) = -F(\mathbf{g}) - \sum_{\mu=1}^M g_{\mu} \langle \phi_{\mu} \rangle_{\mathbf{g}} \quad (2.13)$$

$$D(\mathbf{p}||\mathbf{p}') = F(\mathbf{g}) - F(\mathbf{g}') + \sum_{\mu=1}^M (g_{\mu} - g'_{\mu}) \langle \phi_{\mu} \rangle_{\mathbf{g}} , \quad (2.14)$$

so that the concavity properties of $S(\mathbf{p})$ and $D_{KL}(\mathbf{p}||\mathbf{q})$ can be related to the ones of the free energy $F(\mathbf{g})$. These quantities will be relevant in order to characterize the large deviation properties both for the direct and of the inverse problem.

2.1.3 Observables

Throughout all our discussion, we will focus on the case in which T independent, identically distributed (i.i.d.) configurations of the system denoted as $\hat{\mathbf{s}} = \{s^{(t)}\}_{t=1}^T$ are observed. The joint probability of observing the dataset $\hat{\mathbf{s}}$ (also called *likelihood*) given a statistical model (ϕ, \mathbf{g}) is

$$P_T(\hat{\mathbf{s}}|\mathbf{g}) = \prod_{t=1}^T p(s^{(t)}) = \exp\left(T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu}\right) \quad (2.15)$$

where the quantities

$$\bar{\phi}_{\mu} = \frac{1}{T} \sum_{t=1}^T \phi_{\mu}(s^{(t)}) \quad (2.16)$$

are called *empirical averages*. It is worth remarking that $P_T(\hat{\mathbf{s}})$ depend on the observed configurations just through the empirical averages $\bar{\phi}$. We will denote averages over the measure $P_T(\hat{\mathbf{s}}|\mathbf{g})$ with the notation $\langle \dots \rangle_T$. We also define the *empirical frequencies* (also known as *type*) $\bar{\mathbf{p}}$ as the vector with components

$$\bar{p}_s = \frac{1}{T} \sum_{t=1}^T \delta_{s, s^{(t)}} , \quad (2.17)$$

which enjoys the following properties:

- It is positive and normalized ($\sum_s \bar{p}_s = 1$), thus it defines a probability density on Ω (i.e., $\bar{\mathbf{p}} \in \mathcal{M}(\Omega)$).
- The empirical averages $\bar{\phi}$ can be obtained as $\bar{\phi}_{\mu} = \sum_s \phi_{\mu, s} \bar{p}_s$.

- If the dataset $\hat{\mathbf{s}}$ is generated by a probability distribution \mathbf{p} , then $\bar{\mathbf{p}}$ is distributed according to the multinomial distribution

$$P_T(\bar{\mathbf{p}}|\mathbf{p}) = \left(\prod_s \frac{p_s^{T_s}}{T_s!} \right) T! \delta \left(T - \sum_s T_s \right), \quad (2.18)$$

where $T_s = T\bar{p}_s$. Its first and second momenta are

$$\langle \bar{p}_s \rangle_T = p_s \quad (2.19)$$

$$\langle \bar{p}_s \bar{p}_{s'} \rangle_T - \langle \bar{p}_s \rangle_T \langle \bar{p}_{s'} \rangle_T = \frac{1}{T} (\delta_{s,s'} p_s - p_s p_{s'}) . \quad (2.20)$$

Finally, given a collection of operators ϕ we will denote the set of all empirical averages $\bar{\phi}$ that are compatible with at least one probability density in Ω with

$$\mathcal{G}(\phi) = \left\{ \bar{\phi} \in \mathbb{R}^M \mid \exists \bar{\mathbf{p}} \in \mathcal{M}(\Omega) \text{ s.t. } \bar{\phi}_\mu = \sum_s \phi_{\mu,s} \bar{p}_s \forall \mu \right\}, \quad (2.21)$$

which is called in the literature *marginal polytope* [85]. It can be proven that (see for example [85]):

- $\mathcal{G}(\phi)$ is a convex set (i.e., given $\bar{\phi}, \bar{\phi}' \in \mathcal{G}(\phi)$, for any $\alpha \in [0, 1]$ also $\alpha \bar{\phi} + (1 - \alpha) \bar{\phi}' \in \mathcal{G}(\phi)$).
- $\mathcal{G}(\phi) = \text{conv}\{\phi(s) \in \mathbb{R}^M \mid s \in \Omega\}$, where $\text{conv}\{\cdot\}$ denotes the convex hull operation.
- $\mathcal{G}(\phi)$ is characterized by the Minkowski-Weyl theorem as a subset of \mathbb{R}^M identified by a finite set of inequalities. More formally, one can find a set of vectors $\{\mathbf{x}_a, y_a\}_{a=1}^d$ with d finite such that

$$\mathcal{G}(\phi) = \left\{ \phi \in \mathbb{R}^M \mid \sum_{\mu=1}^M x_{\mu,a} \bar{\phi}_\mu \geq y_a \forall a \in \{1, \dots, d\} \right\} \quad (2.22)$$

2.1.4 Small and large deviations

In the case of the direct problem it is natural to formulate the following questions:

1. What are the most likely values for the empirical averages $\bar{\phi}$?
2. How probable it is to find rare instances $\hat{\mathbf{s}}$?

The first question is relatively easy to answer, and characterizes the role of the generalized susceptibility in the direct problem as ruling the convergence of the empirical averages to the ensemble averages¹, as shown in the following and proven in appendix A.3.

Proposition 2.3. *Given a statistical model (ϕ, \mathbf{g}) , the empirical averages $\bar{\phi}$ satisfy the relations*

$$\langle \bar{\phi}_\mu \rangle_T = \langle \phi_\mu \rangle \tag{2.23}$$

$$\langle \bar{\phi}_\mu \bar{\phi}_\nu \rangle_T - \langle \bar{\phi}_\mu \rangle_T \langle \bar{\phi}_\nu \rangle_T = \frac{\chi_{\mu,\nu}}{T}. \tag{2.24}$$

The explicit form of the likelihood function (2.15) allows to answer exhaustively also to the second question.

Proposition 2.4. *Given a probability density \mathbf{p} defined by a statistical model (ϕ, \mathbf{g}) , the function $I_{\mathbf{p}}(\bar{\mathbf{p}}) = -\frac{1}{T} \log P_T(\bar{\mathbf{p}}|\mathbf{p}) = -F(\mathbf{g}) - \sum_{\mu=1}^M g_\mu \bar{\phi}_\mu$ is the large deviation function for the direct problem.*

This implies that the probability of observing dataset a generic $\hat{\mathbf{s}}$ decays exponentially in T , with a non-trivial rate function $I_{\mathbf{p}}(\bar{\mathbf{p}})$ determined by the empirical

¹In the framework that we are considering (i.i.d. sampling of configurations drawn by the same distribution) empirical averages always converge to ensemble averages with an error scaling as $1/\sqrt{T}$. Indeed it makes sense to model the case in which the probability measure \mathbf{p} breaks into *states*, so that for any finite length experiment, just samples belonging to the same state are observed. This is meant to model the phenomenon of ergodicity breaking, which we will comment about in section 3.4.

averages $\bar{\phi}$ only. Also notice that the large deviation function can be expressed entirely in terms of the entropy and the Kullback-Leibler divergence as

$$I_{\mathbf{p}}(\bar{\mathbf{p}}) = D_{KL}(\bar{\mathbf{p}}|\mathbf{p}) + S(\bar{\mathbf{p}}) . \quad (2.25)$$

2.2 The inverse problem

In this section we introduce the *inverse problem* of extracting a coupling vector \mathbf{g}^* given a set of operators ϕ and a vector of empirical averages $\bar{\phi}$. We will present this problem as dual with respect to the direct one, showing that just as the knowledge of the free energy $F(\mathbf{g})$ completely solves the direct problem, the Legendre transform of $F(\mathbf{g})$ denoted as $S(\bar{\phi})$ and characterized as the Shannon entropy, analogously controls the inverse one.

2.2.1 Bayesian formulation

We will be interested in calculating the set of couplings \mathbf{g}^* which best describes a given set of data $\hat{\mathbf{s}}$ of length T within the statistical model (ϕ, \mathbf{g}) . Bayes theorem provides a mathematical framework in which the problem can be rigorously stated, by connecting the likelihood function $P_T(\hat{\mathbf{s}}|\mathbf{g})$ described in section 2.1.3 to the *posterior* of the model $P_T(\mathbf{g}|\hat{\mathbf{s}})$, which specifies the probability that the data $\hat{\mathbf{s}}$ has been generated by model \mathbf{g} . Bayes theorem states in fact that

$$P_T(\mathbf{g}|\hat{\mathbf{s}}) \propto P_T(\hat{\mathbf{s}}|\mathbf{g})P_0(\mathbf{g}) , \quad (2.26)$$

where $P_0(\mathbf{g})$ is known as the *prior*, and quantifies the amount of information which is *a priori* available about the model by penalizing or enhancing the probability of models specified by \mathbf{g} by an amount $P_0(\mathbf{g})$. Bayes theorem also links the concept of prior to the one of regularization which will be discussed in section 2.3, but for

the moment we will consider the prior $P_0(\mathbf{g})$ to be uniform (i.e. a \mathbf{g} -independent constant), so that it can be reabsorbed into the pre factor of equation (2.26). In this case finding the best model to describe the empirical averages may mean:

- Finding the point in the space of couplings \mathbf{g} in which the function $P_T(\hat{\mathbf{s}}|\mathbf{g})$ is maximum (*maximum likelihood* approach).
- Finding the region of the space of couplings in which such probability is high (*Bayesian* approach).

These two approaches lead to very similar results in the case in which the likelihood function is strictly concave, as one can prove by means of large deviation theory (see section 2.2.4 and appendix A.6). Roughly speaking, when the number of observations T is large, the posterior $P_T(\mathbf{g}|\hat{\mathbf{s}})$ concentrates around the maximum likelihood parameter, being the rate of convergence fixed by the stability matrix of the maximum and the number of samples T . Hence we will later define as the inverse problem the characterization of the maximum likelihood parameters and of their linear stability, disregarding the detailed shape of the function $P_T(\mathbf{g}|\hat{\mathbf{s}})$.

2.2.2 Maximum likelihood criteria

The maximum likelihood criteria requires to find the maximum of the likelihood function $P_T(\hat{\mathbf{s}}|\mathbf{g})$, whose solution is obtained by differentiation of equation (2.15) with respect to the couplings g_μ , and reads for each μ

$$\langle \phi_\mu \rangle = \bar{\phi}_\mu, \tag{2.27}$$

a condition which will be referred as *momentum matching* condition. Thus, the best parameters \mathbf{g}^* describing a set of data $\hat{\mathbf{s}}$ under the model (2.1) in absence of prior are the ones for which the ensemble averages of the model are matched with the empirical ones.

Remark 2.1. *It is easy to see that the matching condition (2.27) can alternatively be obtained by minimizing the Kullback-Leibler divergence $D_{KL}(\bar{\mathbf{p}}|\mathbf{p})$ between the probability distribution defined by the empirical frequencies $\bar{\mathbf{p}}$ and the probability density \mathbf{p} defined by the statistical model (ϕ, \mathbf{g}) .*

2.2.3 Statement of the inverse problem

The concavity properties of the likelihood function (or equivalently, of the free energy $F(\mathbf{g})$), allow for a characterization of the problem of inferring the maximum likelihood parameters \mathbf{g}^* given data $\hat{\mathbf{s}}$ in terms of a Legendre transform of $F(\mathbf{g})$.

Definition 2.6. Given a minimal set of operators ϕ and a set of empirical averages $\bar{\phi}$, the function $S(\bar{\phi})$ is defined as the Legendre transform

$$-S(\bar{\phi}) = \max_{\mathbf{g}} \left(\sum_{\mu=1}^M g_{\mu} \bar{\phi}_{\mu} + F(\mathbf{g}) \right). \quad (2.28)$$

We denote with \mathbf{g}^* the (only) value of \mathbf{g} maximizing equation (2.28). Such quantity satisfies

$$\bar{\phi}_{\mu} = - \left. \frac{\partial F(\mathbf{g})}{\partial g_{\mu}} \right|_{\mathbf{g}=\mathbf{g}^*}. \quad (2.29)$$

By construction the statistical model (ϕ, \mathbf{g}^*) verifies the matching condition (2.27).

By considering the Shannon entropy $S(\mathbf{p}) = - \sum_s p_s \log p_s$ and by plugging probability density \mathbf{p}^* inside its definition, one can see that it holds

$$S(\mathbf{p}^*) = - \sum_{\mu=1}^M g_{\mu}^* \bar{\phi}_{\mu} - F(\mathbf{g}^*) = S(\bar{\phi}), \quad (2.30)$$

which characterizes the Legendre transformation (2.28) of the free energy $F(\mathbf{g})$: $S(\bar{\phi})$ is the Shannon entropy of the distribution expressed as a function of the empirical averages.

Remark 2.2. *The existence of a solution $\mathbf{g}^*(\bar{\phi})$ to the minimization problem defining the entropy $S(\bar{\phi})$ is guaranteed by a general result stating that given any operator set ϕ defining a marginal polytope $\mathcal{G}(\phi)$, the empirical averages $\bar{\phi}_\mu = \sum_s \phi_{\mu,s} \bar{p}_s$ can be matched by ensemble averages $\langle \phi^* \rangle$ associated with the statistical model (ϕ, \mathbf{g}^*) , with $\mathbf{g}^* \in (\mathbb{R} \cup \{-\infty, +\infty\})^M$. The interested reader is referred to [85] for the mathematical details.*

Proposition 2.5. *By differentiation of equation (2.28) one finds that*

$$-\frac{\partial S}{\partial \bar{\phi}_\mu} = g_\mu^*, \quad (2.31)$$

while by applying the chain rule to the equation $\delta_{\mu,\nu} = \partial g_\mu / \partial g_\nu$ one finds that

$$-\frac{\partial^2 S}{\partial \bar{\phi}_\mu \partial \bar{\phi}_\nu} = \chi_{\mu,\nu}^{-1}. \quad (2.32)$$

Equations (2.31) and (2.32) are analogous to equations (2.8) and (2.9) which relate to the direct problem. Just as the free energy $F(\mathbf{g})$ generates averages and susceptibilities in the direct problem, the entropy $S(\bar{\phi})$ is the generating function for the inverse one. Hence, an inference problem can be solved by explicitly computing the Shannon entropy $S(\bar{\phi})$ and finding its maximum (either analytically or numerically).

Definition 2.7. The problem of determining the entropy $S(\bar{\phi})$, the inferred couplings \mathbf{g}^* and the inverse susceptibility $\hat{\chi}^{-1}$ as functions of the averages $\bar{\phi}$ will be referred as the *inverse problem*.

2.2.4 Small and large deviations

Two questions analogous to the ones formulated in section 2.1.4 in the case of the direct problem can be formulated for the inverse problem, namely: (i) what are the most likely values for the inferred coupling \mathbf{g}^* obtained by a dataset $\hat{\mathbf{s}}$ of length T ? and (ii) how likely it is that such dataset has been generated by a model very different from the maximum likelihood one? In order to answer to those two questions we need to consider the large deviation function for the inverse problem. This can be obtained by noting that in absence of a prior, Bayes theorem and equation (2.25) imply that

$$P_T(\mathbf{p}|\bar{\mathbf{p}}) \propto P_T(\bar{\mathbf{p}}|\mathbf{p}) = e^{-T(D_{KL}(\bar{\mathbf{p}}|\mathbf{p})+S(\bar{\mathbf{p}}))} \propto e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})} \quad (2.33)$$

so that we can prove the following proposition.

Proposition 2.6. *Given a vector of empirical frequencies $\bar{\mathbf{p}}$, the large deviation function for the inverse problem $I_{\bar{\mathbf{p}}}(\mathbf{p}) \propto -\frac{1}{T} \log P_T(\mathbf{p}|\bar{\mathbf{p}})$ is given by the Kullback-Leibler divergence*

$$I_{\bar{\mathbf{p}}}(\mathbf{p}) = D_{KL}(\bar{\mathbf{p}}|\mathbf{p}) . \quad (2.34)$$

This implies that the probability for data $\bar{\mathbf{p}}$ to be generated by any model \mathbf{p} decays exponentially fast in T with a rate function given by the large deviation function $D_{KL}(\bar{\mathbf{p}}|\mathbf{p})$. This result can be seen as a particular case of a more general theorem, which is known as Sanov theorem and whose proof can be found in appendix A.4.²

Theorem 2.1. *Consider a statistical model defined by a probability distribution \mathbf{p} , and a (compact) set of probability densities $\mathcal{M} \subseteq \mathcal{M}(\Omega)$. Then if $\bar{\mathbf{p}}$ is a vector of*

²We won't adopt the informal version of the theorem often found in literature (see for example [54]), which doesn't require the introduction of the set \mathcal{M}' . In such form the theorem is not valid when, for any value of T , \mathcal{M} has empty intersection with the set of realizable empirical frequencies, as the probability for any point in \mathcal{M} to be realized is strictly zero regardless of T .

empirical frequencies sampled from the distribution $P_T(\hat{\mathbf{s}}|\mathbf{p})$, it holds that

$$\lim_{\delta \rightarrow 0} \lim_{T \rightarrow \infty} -\frac{1}{T} \log \text{Prob}(\bar{\mathbf{p}} \in \mathcal{M}') = D_{KL}(\mathbf{q}^*|\mathbf{p}), \quad (2.35)$$

where $\mathbf{q}^* = \arg \min_{\mathbf{q} \in \mathcal{M}} D_{KL}(\mathbf{q}|\mathbf{p})$ and \mathcal{M}' is the compact set $\mathcal{M}' = \{\mathbf{p} + \delta \mathbf{p} = \mathbf{p}' \in \mathcal{M}(\Omega) \mid \mathbf{p} \in \mathcal{M}, \delta \mathbf{p} \in [-\delta, \delta]^{|\Omega|}\}$.

Building on these results, we can provide an answer for our first question and find out what are the most likely distributions \mathbf{p} having generated data $\bar{\mathbf{p}}$. In particular, it is possible to expand the Kullback-Leibler divergence around its minimum and perform a saddle-point estimation, obtaining the following result.

Proposition 2.7. *Consider a generic dataset $\hat{\mathbf{s}}$ defining the empirical distribution $\bar{\mathbf{p}} \in \mathcal{M}(\Omega)$. Then, given a family of operators ϕ , the posterior probability (with uniform prior) $P_T(\mathbf{g}|\bar{\mathbf{p}}) \propto P_T(\bar{\mathbf{p}}|\mathbf{g})$ defines a probability measure on space $\mathcal{M}(\phi)$, parametrized by the coupling vector \mathbf{g} which defines the statistical model \mathbf{p} . The averages and the covariances under this measure are given in the large T limit by*

$$\frac{\int d\mathbf{g} g_\mu e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})}}{\int d\mathbf{g} e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})}} \xrightarrow{T \rightarrow \infty} g_\mu^* \quad (2.36)$$

$$\frac{\int d\mathbf{g} g_\mu g_\nu e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})}}{\int d\mathbf{g} e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})}} - g_\mu^* g_\nu^* \xrightarrow{T \rightarrow \infty} \frac{\chi_{\mu,\nu}^{-1}}{T}. \quad (2.37)$$

where \mathbf{g}^* is the maximum likelihood estimator of \mathbf{g} and $\hat{\chi}^{-1}$ is the inverse of the Fisher information matrix calculated in \mathbf{g}^* .

This result (proved in appendix A.6) characterizes the inverse of the generalized susceptibility as the matrix quantifying the speed in T at which the probability measure on the inferred couplings concentrates around the maximum likelihood estimate. The centrality of this matrix in the inverse problem is also provided by a rigorous bound that can be proven for the covariance of any unbiased estimator, and known as

Cramér-Rao bound. From this perspective, $\chi_{\mu,\nu}^{-1}$, can be seen as establishing a bound to the maximum rate of convergence for the estimator of a coupling.

Theorem 2.2. *Consider a statistical model (ϕ, \mathbf{g}) with $F(\mathbf{g})$ strictly concave and an unbiased estimator of the couplings \mathbf{g}^* (i.e., such that $\langle g_\mu^* \rangle_T = g_\mu$). Then the covariance matrix of \mathbf{g}^* under the measure $\langle \dots \rangle_T$ satisfies*

$$\langle (\mathbf{g}^* - \mathbf{g})(\mathbf{g}^* - \mathbf{g})^T \rangle_T \succeq \frac{\hat{\mathbf{X}}^{-1}}{T} \quad (2.38)$$

where with $\hat{\mathbf{X}} \succeq \hat{\mathbf{Y}}$ we indicate that the matrix $\hat{\mathbf{X}} - \hat{\mathbf{Y}}$ is positive semidefinite.

The proof of this theorem is presented in the appendix A.5.

2.2.5 Examples

Independent spins model

The simplest model of the form (2.1) which can be considered is of the form

$$p(s) = \frac{1}{Z(\mathbf{h})} \exp \left(\sum_{i \in V} h_i s_i \right) \quad (2.39)$$

and will be called *independent spin model*. The model contains N operators of the form $\{\phi_{\{i\}}(s) = s_i\}_{i \in V}$ (called in the following *magnetizations*), whose conjugated couplings are denoted as $g_{\{i\}} = h_i$ (and referred as *external fields*). The empirical magnetizations will be denoted as $\bar{s}_i = m_i$. The direct problem can be solved by evaluating the partition function of the model, so that the free energy $F(\mathbf{h})$ results

$$F(\mathbf{h}) = -N \log 2 - \sum_{i \in V} \log \cosh h_i . \quad (2.40)$$

The ensemble averages and generalized susceptibilities can be obtained by differentiation, and are given by

$$m_i = \tanh h_i \quad (2.41)$$

$$\chi_{i,j} = \frac{\delta_{i,j}}{\cosh^2 h_i} \quad (2.42)$$

The inverse problem is also easily solvable, as the Legendre transformation of $F(h)$ can explicitly be computed, and the entropy results

$$S(\mathbf{m}) = - \sum_{i \in V} \left(\frac{1+m_i}{2} \log \frac{1+m_i}{2} + \frac{1-m_i}{2} \log \frac{1-m_i}{2} \right) \quad (2.43)$$

while by differentiation one finds

$$h_i^* = \operatorname{arctanh} m_i \quad (2.44)$$

$$\chi_{i,j}^{-1} = \frac{\delta_{i,j}}{1-m_i^2} \quad (2.45)$$

The additivity both of the entropy and of the free energy, which are crucial in order to solve the model, descend directly by the independence of $p(s)$, which can be written as a product of single spin marginals

$$p(s) = \prod_{i \in V} p^{\{i\}}(s_i). \quad (2.46)$$

Notice that the existence of the solution is guaranteed for any \mathbf{m} in the hypercube $[-1, 1]^N$, while its uniqueness is enforced by the minimality of the operator set $\{s_i\}_{i=1}^N$ (which is additionally an orthogonal set in the sense that will be defined in (4.1)). As expected, for $m_i = \pm 1$, the estimator h_i^* is divergent, so that $h_i^*(m_i = \pm 1) = \pm \infty$.

The pairwise model

The next model that will be presented is known in a large variety of fields with different names (Ising model in physics, graphical model in the field of statistical learning), and is defined by the probability density

$$p(s) = \frac{1}{Z(\mathbf{h}, \hat{\mathbf{J}})} \exp \left(\sum_{i \in V} h_i s_i + \sum_{(i,j) \in E} J_{ij} s_i s_j \right), \quad (2.47)$$

where E is a given set of *edges*, that is, a given subset of $\{(i, j) \in V \times V \mid i < j\}$. While in statistical mechanics it has been extensively used since 1925 as a prototypical model to study magnetic materials [41, 15], it has deserved a special interest in the field of statistical learning as it is the simplest model which is able to capture the correlation structure of a given dataset³. The operator content of this model is a set of N magnetizations, conjugated to their corresponding external fields (as in section 2.2.5), and a set of $|E| \leq \frac{N(N-1)}{2}$ operators $\{\phi_{\{i,j\}}(s_i, s_j) = s_i s_j\}_{(i,j) \in E}$ conjugated to a set of *pairwise couplings* $g_{\{i,j\}} = J_{ij}$. We will call *empirical correlations* the averages $\overline{s_i s_j} = c_{ij}$.

Remark 2.3. *This direct problem for the pairwise model is hard to solve in the general case for even moderate values of N , in the sense that the calculation of the partition function $Z(\mathbf{g})$ is a problem which is known to belong to the #P-complete class [43, 42]. Only for some subclasses of this general problem an exact, analytical solution for the partition function can be obtained (e.g., regular lattices, trees) and evaluated in polynomial time, while in general just approximate solutions can be obtained in polynomial time [42]. Another possible approach consists in finding approximated expressions for the partition function $Z(\mathbf{h}, \hat{\mathbf{J}})$ which are proven to converge in the*

³This can be shown via the maximum entropy principle, which is presented and thoroughly commented in appendix A.1.

limit of large system size or weak interaction to the exact result for the free energy of the model (mean-field approximations).

In the next sections we will introduce specific versions of model (2.47) for which we will be able to solve the inverse problem, namely the fully connected ferromagnet (section 3.3) and the pairwise tree (section 4.2.3).

2.3 The regularized inverse problem

The inverse problem described in section 2.2 may appear extremely easy to solve due to the concavity of the free energy $F(\mathbf{g})$. The optimization of concave functions is usually very easy because fast algorithms such as gradient ascents can find in short time a maximizer (if any) for $F(\mathbf{g})$ (appendix C). Despite that, there are several cases in which this procedure may be problematic, so that the function $F(\mathbf{g})$ is often replaced by a modified function $F(\mathbf{g}) - H_0(\mathbf{g})$ which enforces a better behavior for the inverse problem. In this case the function $H_0(\mathbf{g})$ is called a regularizer. In a Bayesian setting, regularization can be understood as an injection of *a priori* information about a statistical model. Indeed the issue of regularization is a topic of fundamental importance in the field of statistical inference well beyond the need of enforcing mathematical tractability of the model. In particular it can be used to deal with these cases:

- **Divergencies:** Regularization can cure divergencies, by removing infinite couplings. A solution to any inverse problem is guaranteed to exist for any set of empirical averages $\bar{\phi} \in \mathcal{G}(\phi)$, but such solution may be located at the boundary of the coupling space, in whose case one or more couplings are divergent. Penalizing large couplings with a regularizer ensures that the inferred couplings attain a finite value. This is often the case for neurobiological or protein data

and can be related to undersampling, as motivated in sections 4.2.1, 4.2.3 and 4.2.4 [25, 72, 87, 24, 26].

- **Uniqueness:** Regularization can enforce uniqueness for the solution of the inverse problem, by removing the zero models of the $\hat{\chi}$ matrix. Such modes can arise if the family ϕ is not minimal (appendix A.2), or can be linked to the large N limit (chapter 3).
- **Generalizability:** Regularization can be used to improve generalizability of a statistical model in the case of under sampling: if the inferred probability has a much smaller entropy with respect to the true one, an inferred model is likely not to be predictive. A compromise between faithfulness to the data and simplicity of the model can nevertheless be achieved by penalizing the complexity of the model with a regularization term, which is expected to lift the entropy of the inferred model. The balance between over and under fitting can be heuristically evaluated by using cross-validation methods (e.g., by using one half of the data to calibrate the model and by computing the likelihood of the other half) or by using a complexity measure for the inferred model (such as the Akaike information criterium [7] or the Bayesian information criterion [73]), in order to tune the regularizer to a correct value (see also section 5.1.3).
- **Model selection:** Finally, regularization can be used as a tool to perform model selection. In the case in which data are distributed according to a specific, unknown, statistical model, it is possible to perform inference by using a more general distribution which is likely to contain (or to be very close) to the true one. By adding a suitable regularizing term (such as an L-1 or L-0 norm) it is sometimes possible to recover the original model as a particular sub-class of a more general distribution. For example, this has been used in the context of graph reconstruction, where models defined by specific topologies have

been successfully selected by a regularizer out of the space of all possible graph structures [63, 86].

2.3.1 Bayesian formulation

Consider an empirical dataset $\hat{\mathbf{s}}$ and a model defined by a set of operators ϕ . Then the posterior of the model can be written as in (2.26), in which it is $P_T(\mathbf{g}|\hat{\mathbf{s}}) \propto P_T(\hat{\mathbf{s}}|\mathbf{g})P_0(\mathbf{g})$, so that the problem of inference can be reformulated as the minimization of the function

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -\log P_T(\hat{\mathbf{s}}|\mathbf{g}) - \log P_0(\mathbf{g}) = -T \sum_{\mu=0}^M g_\mu \bar{\phi}_\mu - \log P_0(\mathbf{g}) . \quad (2.48)$$

Definition 2.8. Given a statistical model (ϕ, \mathbf{g}) and a positive prior function $P_0(\mathbf{g})$ we define a *regularizer* as the function $H_0(\mathbf{g}) = -\log P_0(\mathbf{g})$.

Notice that due to convexity of the $\hat{\chi}$ matrix, if the regularizer $H_0(\mathbf{g})$ is (strictly) convex, also $H(\mathbf{g}|\hat{\mathbf{s}})$ is (strictly) convex. Hence, the introduction of a strictly convex prior can be used to remove zero modes from the $\hat{\chi}$ matrix thus enforcing a unique solution for the inverse problem. In our analysis we will restrict to the case of convex regularizers. Also notice that if $H_0(\mathbf{g}) = +\infty$ when any component of \mathbf{g}^* is divergent, the solution of the inverse problem is confined to a finite region of the coupling space.

2.3.2 Two popular regularizers

We present two known regularization schemes, with the purpose of providing simple examples of convex regularizers, while showing at the same time two widely used regularization mechanisms. The details about the properties and the implementation of the algorithms used to solve these regularized problems are reminded in appendix C.

L-2 regularization

Given a statistical model (ϕ, \mathbf{g}) , a set of empirical frequencies $\bar{\phi}$ and a vector β such that it is component-wise $\beta_\mu > 0$, we consider the minimization problem

$$H(\mathbf{g}) = -T \sum_{\mu=0}^M g_\mu \bar{\phi}_\mu + \sum_{\mu=1}^M \frac{\beta_\mu}{2} g_\mu^2, \quad (2.49)$$

which we call the *L-2 regularized* inverse problem. This choice for $H_0(\mathbf{g})$ enforces strict concavity of the problem and finiteness of the values of \mathbf{g}^* , which should satisfy the set of equations

$$\bar{\phi}_\mu - \langle \phi_\mu \rangle - \frac{\beta_\mu}{T} g_\mu = 0. \quad (2.50)$$

This regularization corresponds to the Gaussian prior $P_0(\mathbf{g}) \propto \exp\left(-\sum_{\mu} \frac{\beta_\mu}{2} g_\mu^2\right)$. Notice also that the regularizer is differentiable, so that a solution of this problem can be addressed efficiently by using techniques such as the ones described in the first part of appendix C. As in the non-regularized case, the main computational limitation consists in calculating the gradient of the minus log-likelihood function $-\log P_T(\bar{\phi}|\mathbf{g})$, which requires the knowledge of the averages $\langle \phi \rangle$ as functions of the coupling vector \mathbf{g} . This regularization procedure is typically used to remove infinite values arising in the solution of the non-regularized inverse problem.

L-1 regularization

We also present the *L-1 regularized* inverse problem, which is defined by the minimization problem

$$H(\mathbf{g}) = -T \sum_{\mu=0}^M g_\mu \bar{\phi}_\mu + \sum_{\mu=1}^M \beta_\mu |g_\mu|, \quad (2.51)$$

corresponding to the choice of an exponentially decaying prior $\exp\left(-\sum_{\mu} \beta_\mu |g_\mu|\right) \propto P_0(\mathbf{g})$ for the coupling vector \mathbf{g} . Analogously to the L-2 case, this regularizer is convex and enforces a finite value for the inferred couplings \mathbf{g}^* . Unlike that, this

regularizer is non-differentiable. This introduces some difficulties in the solution of the minimization problem, as shown in the second part of appendix C, where it is shown that the inferred coupling vector should satisfy the equation

$$0 \in \bar{\phi}_\mu - \langle \phi_\mu \rangle - \beta_\mu \text{sgn}(g_\mu) , \quad (2.52)$$

where $\text{sgn}(x)$ is the set-valued function defined by equation (C.11). The main interest in this regularizer arises from its efficacy as a feature-selector, as it is able to provide sparse solutions for \mathbf{g}^* , i.e., to put exactly to zero some components of the inferred couplings vector. Despite being first used in the field of *compressed-sensing*, in which the use of the L-1 regularizer has been exploited to solve underconstrained sets of linear equations [31], this regularized has been successfully applied in the field of binary inference (also called *logistic regression*), in which it has been useful to reconstruct the structure of an interaction network of the form (2.47) [63, 86] and even in more general cases dealing with non-binary interaction [71].

Remark 2.4. *The two regularizers presented so far are special cases of the L-p regularization scheme, which is associated with the choice $H_0(\mathbf{g}) \propto \sum_\mu \beta_\mu \|g_\mu\|_p$, where $\|x\|_p = |x|^p$ is the L-p norm of x . Notice that the L-p regularizer is convex (hence leading to computationally tractable minimization problems) for $p \geq 1$, and is strictly so for $p > 1$. In particular, the L-1 regularizer can be seen as the simple (alias, convex) regularizer that is closer to the L-0 one, which is associated with the problem of minimizing the number of inferred parameters for a fixed value of the posterior, a criterium which one would think to use in order to minimize the complexity of the inferred model.*

2.3.3 Examples

Independent spins model

Consider the model defined by the probability density (2.39). Then we can consider the regularized inverse problems in which one tries to minimize

$$H(\mathbf{h}|\hat{\mathbf{s}}) = -T \sum_{i \in V} (h_i m_i - N \log 2 - \log \cosh h_i) + H_0(\mathbf{h}), \quad (2.53)$$

in the two cases $H_0(\mathbf{h}) = \sum_i \frac{\beta_i}{2} h_i^2$ and $H_0(\mathbf{h}) = \sum_i \beta_i |h_i|$ corresponding respectively to the L-2 and L-1 norm. In the first case the (decoupled) set of equations which has to be solved in order to find the vector \mathbf{h} is

$$m_i - \tanh(h_i) = \frac{\beta_i}{T} h_i, \quad (2.54)$$

whose graphical solution is depicted in figure 2.1. Such plot and equation (2.54) also

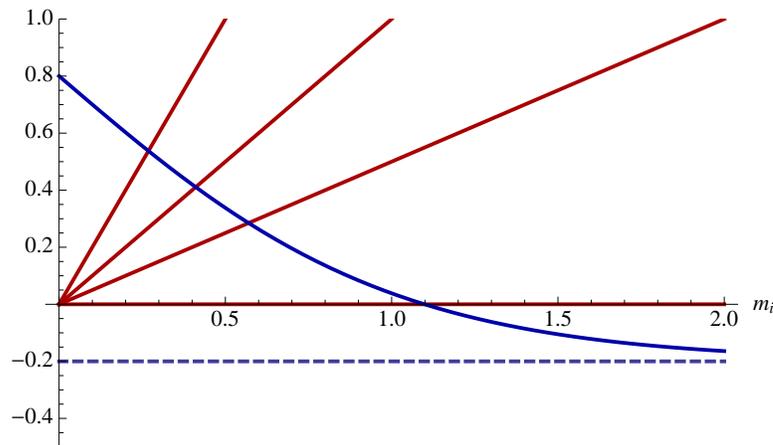


Figure 2.1: Graphical solution of equation (2.54) yielding the inferred field h_i^* for the L-2 regularized independent spin model. The blue curve displays the quantity $m_i - \tanh h_i$ in the case $m_i = 0.8$, while the red ones show the product $\beta_i h_i / T$ for $\beta_i/T = 0.5, 1, 2$. The dashed line plotted for reference corresponds to the line $m_i - 1$.

show that the inferred couplings h_i attain a finite value for any of $-1 \leq m_i \leq 1$ and $0 \leq \beta_i < \infty$. In the case of the L-1 norm, one has to solve the decoupled set of

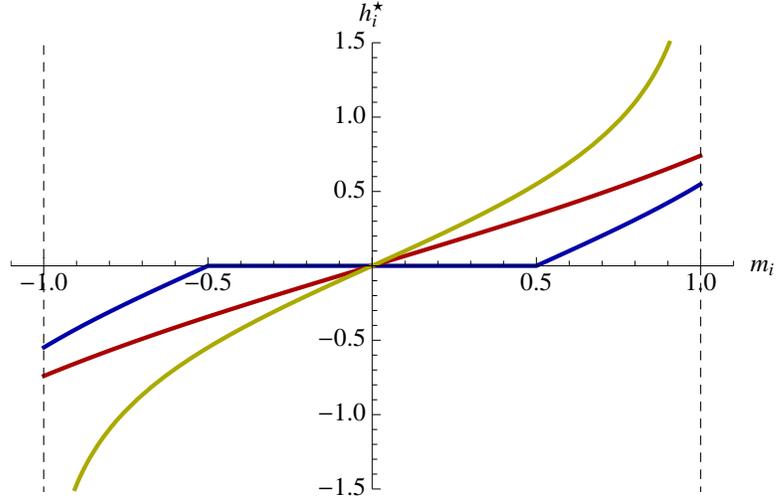


Figure 2.2: Solution for the inferred field h_i^* for the L-1 (blue line) and L-2 (red line) regularized independent spin model as a function of the empirical magnetization for $\beta_i/T = 0.5$. The solution for the non-regularized problem is also plotted for comparison (yellow line).

equations

$$m_i - \tanh h_i = \frac{\beta_i}{T} \operatorname{sgn}(h_i), \quad (2.55)$$

whose solution is

$$h_i = \begin{cases} 0 & \text{if } \frac{\beta_i}{T} > |m_i| \\ \operatorname{arctanh} \left[m_i - \frac{\beta_i}{T} \operatorname{sign}(m_i) \right] & \text{if } \frac{\beta_i}{T} \leq |m_i| \end{cases}. \quad (2.56)$$

The solution for h_i in the two cases for a specific value of β_i is plotted in figure 2.2.

Notice that:

- Both regularizations schemes produce a finite h_i in the case $|m_i| = 1$.
- The zero-field solutions of the L-1 regularized problem can be seen as arising from a complexity related criteria, stating that *operators which do not add enough descriptive power to the model should be suppressed by the assignment of zero weight to their conjugated coupling*. In this example the notion of “enough

descriptive power” is quantified through the comparison of β_i against the directional derivative of the log-likelihood $\partial_{h_i} \log P_T(m_i \geq 0 | h_i)|_{h_i \rightarrow 0^\pm} = Tm_i$.

Despite its trivial solution, we have chosen to present this problem as it shows with simplicity some basic features of the L-1 and L-2 regularizers which are retained even in more complicated scenarios.

Chapter 3

High-dimensional inference

Rigorous results in information theory – such as the ones presented in section 2.2.4 – are able to provide both qualitative and quantitative understanding of the inverse problem in the regime of finite N and large T , the case most of the literature on statistical learning deals with, while computational techniques such as the ones described in appendix C provide efficient means to find its solution. Nevertheless, recent technological advances in several fields (such as biology, neuroscience, finance, economy) are pushing the fields of statistical learning towards a less trivial regime, in which both N and T are large, with a given relation among system size and number of samples keeping fixed their scaling. The reason for this change of perspective is that it is now possible for several complex systems to record a large number of data samples describing simultaneous the activity of the many microscopic constituents [22, 72, 76, 24, 48, 59, 29]. The question that naturally arises in this case is whether it makes sense to consider a model with a large (possibly very large) number of parameters, if the data available is also very large. The answer is non-trivial, and requires the addition of some degree of complexity to the problem of inference. The first problem which has to be addressed (section 3.1) is of purely technical nature, and deals with the problem of finding the minimum of a convex function when its

gradient is computationally intractable. Then, we will describe some interesting conceptual problems which arise when considering the large N limit. For simplicity, we will consider initially the problem in which both N and T are large, but the number of inferred parameters M is finite (section 3.2). Discussing the case in which M scales with N as well will require the introduction of the notion of *disorder*, which we will briefly comment about in section 3.5.

3.1 Computational limitations and approximate inference schemes

In appendix C we show how it is possible to construct algorithms which are guaranteed to find a minimum (if any) for a convex function. Then the solution of the inverse problem can be written as a minimization problem over a convex function $H(\mathbf{g})$ of the form

$$H(\mathbf{g}) = -\log P_T(\bar{\mathbf{s}}|\mathbf{g}) - \log P_0(\mathbf{g}) , \quad (3.1)$$

that problem is in principle solved. Indeed, the problem which often arises in many practical cases is that the naive minimization of this function can be extremely slow, and *ad-hoc* techniques have to be implemented in order to overcome this problem.

3.1.1 Boltzmann Learning

One of the most intuitive algorithms to solve the inverse problem is provided by the *Boltzmann learning* procedure [6], which consists in the application of algorithm C.1.1 to the inverse problem described in section 2.2. In that case, the minimization procedure of $H(\mathbf{g})$ consists in constructing a succession $\{\mathbf{g}^{(k)}\}_{k=1}^K$ of the form

$$\mathbf{g}^{(k+1)} = \mathbf{g}^{(k)} - \epsilon_k \nabla H(\mathbf{g}^{(k)}) \quad (3.2)$$

where $\{\epsilon_k\}_{k=1}^K$ is a *schedule* satisfying the set of conditions (C.4) which enforce the convergence of $\mathbf{g}^{(k)}$ to the minimum (if any) \mathbf{g}^* . Indeed the computation of each of the $\mathbf{g}^{(k)}$ requires the evaluation of $H(\mathbf{g}^{(k)})$ and the calculation of a gradient of the form

$$\nabla H(\mathbf{g}^{(k)}) = T (\langle \phi_\mu \rangle_{\mathbf{g}^{(k)}} - \bar{\phi}_\mu) + \frac{\partial}{\partial g_\mu} H_0(\mathbf{g}^{(k)}) . \quad (3.3)$$

The calculation of the gradient (or the sub-gradient) of $H(\mathbf{g})$ requires evaluating the ensemble averages of the operators ϕ , which is a computationally challenging task if N is even moderately large. This is true even when the function H and the ensemble averages $\langle \phi \rangle$ are not computed via direct enumeration (which would in principle entail a summation over 2^N states for each of the M operators plus the identity), and are instead calculated with Monte Carlo methods. The number of iterations required to calculate each of the gradients and the function H with a controlled precision is in fact typically fast growing in N , being the quality of the approximation and the time computational power required to obtain it dependent on the algorithm which is adopted to compute the averages (see for example [6, 54, 50, 47]). Summarizing:

- Boltzmann learning is able to solve with arbitrary precision any inverse problem.
- The computational power required to solve the inverse problem through the Boltzmann learning procedure with a given degree of accuracy (i.e. $H(\mathbf{g}^{(K)}) - H(\mathbf{g}^*)$ smaller than a fixed ϵ) grows *fast* in N .

3.1.2 Mean field approaches for pairwise models

An alternative approach to the Boltzmann learning procedure can be constructed by adopting so-called *mean-field* techniques, which allow to obtain efficient approximations for the free energy $F(\mathbf{g})$ and the averages $\langle \phi \rangle$ of a statistical model. Such techniques are suitable for systems whose partition function can be quickly, although approximately, evaluated with a precision which either increases with the system size

N or decreases with the magnitude of the interactions, so that in many practical applications the difference between the approximated observables and the exact ones is very small [68, 67]. For pairwise models of the form (2.47), mean-field approximations are well-known since long time in statistical physics. In particular we will consider approaches in which the free energy of the model (2.47) is expanded in a series around a non-interacting or a weakly correlated model (naive mean field, TAP approximation, Sessak-Monasson approximation), or obtained by assuming a factorization property of the probability distribution in terms of one and two body marginals (Bethe approximation). We will briefly describe these approximate inference schemes without providing explicit derivations, supplying the interested reader with the necessary references.

In order to motivate the mean-field approach, we first state the result [62].

Proposition 3.1. *Consider a pairwise model of the form*

$$p(s) = \frac{1}{Z(\mathbf{h}, \beta \hat{\mathbf{J}})} \exp \left(\sum_{i \in V} h_i s_i + \beta \sum_{(i,j) \in E} J_{ij} s_i s_j \right), \quad (3.4)$$

where $\beta > 0$ is an expansion parameter. Then its free energy can be written as

$$F_\beta(\mathbf{h}, \mathbf{J}) = \sum_{n=0}^{\infty} \beta^n \frac{\partial^n F_\beta}{\partial \beta^n} \quad (3.5)$$

where the terms $\frac{\partial^n F_\beta}{\partial \beta^n}$ are functions such that: (i) depend only on the couplings J_{ij} and the ensemble magnetizations $\langle s_i \rangle$ (ii) for $n \geq 1$ the n -th term involves n -th powers of J_{ij} (iii) the ensemble magnetizations satisfy the self-consistency equations

$$\frac{\partial F_\beta(\mathbf{h}, \beta \hat{\mathbf{J}})}{\partial \langle s_i \rangle} = 0. \quad (3.6)$$

Leaving aside the problem of convergence of the series (3.5), the free energy for a generic pairwise model can in principle be obtained by setting $\beta = 1$ in the above expansion.

- **Naive mean field:** The naive mean field approximation can be obtained by truncating the series (3.5) for $n = 2$, thus obtaining the expression

$$F_{nMF}(\mathbf{h}, \mathbf{J}) = \sum_i \left[\frac{1 + \langle s_i \rangle}{2} \log \frac{1 + \langle s_i \rangle}{2} + \frac{1 - \langle s_i \rangle}{2} \log \frac{1 - \langle s_i \rangle}{2} \right] - \sum_{i \in V} h_i \langle s_i \rangle - \sum_{(i,j) \in E} J_{ij} \langle s_i \rangle \langle s_j \rangle, \quad (3.7)$$

while the self-consistency equations become

$$\langle s_i \rangle = \tanh \left(\sum_{(i,j) \in E} J_{ij} \langle s_j \rangle + h_i \right). \quad (3.8)$$

The solution of the inverse problem within this inference scheme can be obtained by inserting the momentum matching condition $\langle s_i \rangle = m_i$ in the previous expression, yielding a first set of relations among \mathbf{h}^* , $\hat{\mathbf{J}}^*$ and \mathbf{m} . Matching the correlations c_{ij} with the ensemble averages $\langle s_i s_j \rangle$ requires instead the use of linear response theory¹ [45], which can be used to prove that

$$\chi_{\{i\},\{j\}} = \frac{\partial \langle s_i \rangle}{\partial h_j} = c_{ij} - m_i m_j. \quad (3.9)$$

¹Nor by using this inference scheme, nor by using TAP approximation one is able to enforce the momentum matching condition for the correlations without resorting to linear response. This is due to the decorrelation property of the mean-field approximation, which will be thoroughly commented for a simpler model in section 3.3.

Putting those informations together, one finds that

$$(\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T)_{ij}^{-1} = \frac{\delta_{ij}}{1 - m_i^2} - J_{ij}^* \quad (3.10)$$

$$h_i^* = \operatorname{atanh}(m_i) - \sum_{i < j} J_{ij}^* m_j \quad (3.11)$$

- **TAP approximation:** The Thouless-Anderson-Palmer (TAP) approximation can be obtained by considering an additional term in the expansion (3.5), often denoted as *Onsager reaction* [81], leading to the expression for the free energy

$$\begin{aligned} F_{TAP}(\mathbf{h}, \mathbf{J}) &= \sum_i \left[\frac{1 + \langle s_i \rangle}{2} \log \frac{1 + \langle s_i \rangle}{2} + \frac{1 - \langle s_i \rangle}{2} \log \frac{1 - \langle s_i \rangle}{2} \right] \quad (3.12) \\ &- \sum_{i \in V} h_i \langle s_i \rangle - \sum_{(i,j) \in E} J_{ij} \langle s_i \rangle \langle s_j \rangle - \frac{1}{2} \sum_{(i,j) \in E} J_{ij}^2 (1 - m_i^2)(1 - m_j^2), \end{aligned}$$

and the self-consistency relation²

$$\langle s_i \rangle = \tanh \left(\sum_{(i,j) \in E} J_{ij} [\langle s_j \rangle - J_{ij}(1 - \langle s_j \rangle^2) \langle s_i \rangle] + h_i \right). \quad (3.13)$$

Also in this case, in order to apply this approximation to the inverse problem [79], one has to use the momentum matching conditions together with linear response theory, leading to the expression [64]

$$(\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T)_{ij}^{-1} = \left[\frac{1}{1 - m_i^2} + \sum_{k \in V} J_{ik}^* (1 - m_k^2) \right] \delta_{ij} - J_{ij}^* - 2J_{ij}^{*2} m_i m_j \quad (3.14)$$

$$h_i^* = \operatorname{atanh}(m_i) - \sum_{i < j} J_{ij}^* [m_j - J_{ij}^* (1 - m_j^2) m_i]. \quad (3.15)$$

²Notice that the potential emergence of multiple solutions of equation (3.13) is a known feature of several pairwise models, and is generally associated with the emergence of an instability linked with the presence of a glassy phase [8].

While the expansion (3.5) is a series for $F(\mathbf{h}, \hat{\mathbf{J}})$, and is hence associated with the direct problem, it is also possible to find an analogous expansion for the entropy $S(\mathbf{m}, \hat{\mathbf{c}})$ due to Sessak and Monasson which is more naturally associated with the inverse problem [74].

Proposition 3.2. *Given a pairwise model of the form (2.47), the entropy $S(\mathbf{m}, \hat{\mathbf{c}})$ can be expanded as*

$$S(\mathbf{m}, \beta \delta \hat{\mathbf{c}}) = \sum_{n=0}^{\infty} \beta^n \frac{\partial^n S(\mathbf{m}, \beta \delta \hat{\mathbf{c}})}{\partial \beta^n} \quad (3.16)$$

where $\beta > 0$ is a parameter controlling the expansion and $\delta \hat{\mathbf{c}} = \hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$. One can see that (i) the terms $\frac{S(\mathbf{m}, \beta \delta \hat{\mathbf{c}})}{\partial \beta^n}$ depend upon \mathbf{m} and $\delta \hat{\mathbf{c}}$, (ii) for $n \geq 1$ the n -th term of the expansion contains powers of the connected correlation $c_{ij} - m_i m_j$ of order n .

By setting $\beta = 1$, it is also possible to use such an expansion to construct a mean field approximation: the terms in (3.16) can be constructed explicitly through a recursion relation, and each of those can be represented by a diagram, converting the series (3.16) into a diagrammatic expansion.

- **Sessak-Monasson expansion** An infinite number of terms of the expansion (3.16) (which are associated with *loop* diagrams and *two-spin* diagrams) are analytically resummed in [74], where it is found that their contribution leads to

$$\begin{aligned} J_{ij}^* &= \delta_{ij}(1 - m_i^2) - (\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T)_{ij}^{-1} \\ &+ \frac{1}{4} \log \left[\frac{(1 + m_i + m_j + c_{ij})(1 - m_i - m_j + c_{ij})}{(1 + m_i - m_j - c_{ij})(1 - m_i + m_j - c_{ij})} \right] \\ &- \frac{c_{ij} - m_i m_j}{(1 - m_i^2)(1 - m_j^2) - (c_{ij} - m_i m_j)^2}, \end{aligned} \quad (3.17)$$

which is commonly referred as the Sessak-Monasson approximation.

Notice that the expansion (3.16) automatically leads to a series expansion for the external fields and the couplings by using relation (2.31) and exploiting the linearity

of the derivative, without the need of resorting to linear response theory.

A different type of approximation is the so-called Bethe approximation, in which the free energy is written as

$$\begin{aligned}
 F_{BA}(\mathbf{h}, \mathbf{J}) = & - \sum_{(i,j) \in E} p^{\{i,j\}}(m_i, m_j, c_{ij}) \log p^{\{i,j\}}(\langle s_i \rangle, \langle s_j \rangle, \langle s_i s_j \rangle) \\
 & - \sum_{i \in V} (1 - |\partial i|) p^{\{i\}}(\langle s_i \rangle) \log p^{\{i\}}(\langle s_i \rangle) \\
 & - \sum_{i \in V} h_i \langle s_i \rangle - \sum_{(i,j) \in E} J_{ij} \langle s_i s_j \rangle,
 \end{aligned} \tag{3.18}$$

where $\partial i = \{(i, j) \in E\}$ and the averages $\langle s_i \rangle$ and $\langle s_i s_j \rangle$ are self-consistently chosen in order to minimize (3.18). This approximate expression is exact whenever the probability distribution $p(s)$ can be written as a product of one and two body marginals, which is true in the case of trees (see section 4.2.3 and appendix D.2). Notice that for generic systems, the self-consistence equations are not guaranteed to yield a unique, stable solution, being the solutions to the minimization conditions associated with fixed points of the so-called Belief-Propagation (BP) algorithm for constraint satisfaction problems [54]. The expression for the averages obtained by using the free-energy (3.18) is given by [64]

$$\langle s_i \rangle = \tanh \left[h_i + \sum_{j|(i,j) \in \partial i} \operatorname{atanh} [\tanh(J_{ij}) f(\langle s_i \rangle, \langle s_j \rangle, \tanh J_{ij})] \right], \tag{3.19}$$

where

$$f(m_1, m_2, t) = \frac{1 - t^2 - \sqrt{(1 - t^2)^2 - 4t(m_1 - m_2 t)(m_2 - m_1 t)}}{2t(m_2 - m_1 t)}. \tag{3.20}$$

- **Bethe approximation** The use of linear response theory together with equation (3.20) allows to find a solution of the inverse problem in Bethe approxima-

tion, yielding

$$\begin{aligned}
 J_{ij} &= \operatorname{atanh} \left[m_i m_j - \frac{1}{2 \left(\widehat{\delta \mathbf{c}}^{-1} \right)_{ij}} \sqrt{1 + 4(1 - m_i^2)(1 - m_j^2) \left(\widehat{\delta \mathbf{c}}^{-1} \right)_{ij}^2} \right. \\
 &+ \frac{1}{\left(\widehat{\delta \mathbf{c}}^{-1} \right)_{ij}} \left(\frac{1}{4} - m_i m_j \left(\widehat{\delta \mathbf{c}}^{-1} \right)_{ij} \sqrt{1 + 4(1 - m_i^2)(1 - m_j^2) \left(\widehat{\delta \mathbf{c}}^{-1} \right)_{ij}^2} \right. \\
 &\left. \left. + \left(2m_i^2 m_j^2 - m_i^2 - m_j^2 \right) \left(\widehat{\delta \mathbf{c}}^{-1} \right)_{ij}^2 \right)^{1/2} \right] \\
 h_i &= \operatorname{atanh}(m_i) - \sum_{j \in V} \operatorname{atanh} [\tanh(J_{ij}) f(m_i, m_j, \tanh(J_{ij}))] \quad (3.22)
 \end{aligned}$$

where $\widehat{\delta \mathbf{c}} = \hat{\mathbf{c}} - \mathbf{m} \mathbf{m}^T$. Notice that this equation describes the fixed point solution of the susceptibility propagation algorithm (SusProp) [55] without the need of numerically iterating the algorithm itself [64].

Remark 3.1. *The techniques described above have been extensively used in order to solve the inverse problem for the pairwise model. Indeed no general result for the quality of these approximations is rigorously known, thus it is worth remarking that (i) several approximations have been tested on synthetic and experimental data (see for example [24, 68, 67, 52, 64, 12, 26]) in order to check their performance and (ii) those approximations describe the correct expression of the free energy for some specific models. In particular the free energy (3.7) is the exact free energy for the (either homogeneous or heterogeneous) Curie-Weiss model in the limit of large N , (3.12) is the correct free energy for the Sherrington-Kirkpatrick model [62] and the Bethe approximation is exact for loop-less graphs (appendix D.2).*

3.2 The large N , finite M regime

We will be interested in sketching some features of the inverse problem which arise for large values of N (a regime known in statistical mechanics as the *thermodynamic*

limit), and in commenting about their role in the solution of an inference problem such as the one described in section 2.2. In particular we will consider the following issues:

- **Loss of concavity:** A model defined by a strictly concave free-energy $F(\mathbf{g})$ may develop null-modes associated with the matrix $\hat{\chi}$. This implies that the solution of the inverse problem may lose its uniqueness or, more precisely, large regions of the space $\mathcal{M}(\phi)$ might be associated with similar sets of empirical averages $\bar{\phi}$.
- **Model condensation:** Models undergoing a so-called *second order phase transition* display a divergence of one or more components of the generalized susceptibility matrix $\hat{\chi}$. This indicates that large portions of the marginal polytope $\mathcal{G}(\phi)$ can be described by slightly shifting the values of \mathbf{g} around the *critical point* in which $\hat{\chi}$ diverges. More generally, even for non-critical points finite regions of the space of the empirical averages can be mapped by the inverse problem onto sets of apparently vanishing measure of the space $\mathcal{M}(\phi)$. We call this behavior *model condensation*, a phenomenon which will be discussed in great detail in chapter 5.
- **Ergodicity breaking:** The probability measure \mathbf{p} may break in a set of P states, each of them characterized by a different probability density $\mathbf{p}^{(\alpha)}$ (with $\mathbf{p} = \sum_{\alpha=1}^P q_{\alpha} \mathbf{p}^{(\alpha)}$ and $\sum_{\alpha=1}^P q_{\alpha} = 1$). If this is the case, empirical averages produced with a finite amount of data $T \ll |\Omega|$ by any realistic dynamics concentrate according to the measure $\mathbf{p}^{(\alpha)}$ rather than the full measure \mathbf{p} . Then, equation (2.23) fails to hold and the sampled averages are no longer representative of the global probability measure. Hence, the notion of ergodicity breaking deals with the direct problem more than with the inverse one, as it relates to the problem of the convergence of the averages $\bar{\phi}$ to the empirical ensemble

averages $\langle \phi \rangle$. As the discussion of this phenomenon will require the addition of some structure to the direct problem, we will briefly comment its role in section 3.4.

Those features are expected to be universal, i.e., present in several models in the limit $N \rightarrow \infty$ limit. Nevertheless, we will just study a single model known as the fully connected ferromagnet, and try to underline the characteristics which are expected to generalize also to other type of models.

3.3 Fully-connected ferromagnet

We want to illustrate some of the features described above by discussing a completely solvable model. Such model is a particular case of the pairwise model (2.47), and is also known as the Curie-Weiss model of magnetism. It has been used as a prototypical model to study the emergence of a spontaneous magnetization in ferromagnetic materials, as it is one of the simplest statistical models which are able to describe a thermodynamic *phase transition* between a non-ordered phase and an ordered one.

Definition 3.1. Consider the pair of operators $\phi = \left(\sum_i s_i, \frac{1}{N} \sum_{i < j} s_i s_j \right)$, and the statistical model (ϕ, \mathbf{g}) defined by $\mathbf{g} = (h, J)$, so that its associated probability density is given by

$$p(s) = \frac{1}{Z(h, J)} \exp \left(\frac{J}{N} \sum_{i < j} s_i s_j + h \sum_i s_i \right). \quad (3.23)$$

We call this model a *fully connected ferromagnet*. As for the pairwise model, we will write $m = \frac{1}{N} \sum_i \bar{s}_i$ and $c = \frac{2}{N(N-1)} \sum_{i < j} \bar{s}_i \bar{s}_j$.

Due to symmetry, we will consider without loss of generality the model in the region $h \geq 0$. The free energy of the model $F(h, J)$ can be calculated in the large N

limit using a saddle-point approximation, and can be written as

$$F(h, J) \xrightarrow{N \rightarrow \infty} \frac{J}{2} + F_0(h, J) + F_{fluct}(h, J) + F_{trans}(h, J), \quad (3.24)$$

where $F_0(h, J)$ is the leading term of the saddle point expansion, $F_{fluct}(h, J)$ describes the Gaussian fluctuations around the saddle point solution and $F_{trans}(h, J)$ accounts for the presence of multiple solutions (the details of the expansion and the definition of the terms can be found in appendix B.1). Due to linearity of the derivative, it is possible to solve the direct problem taking into account the contributions of those terms separately. The phenomenology of the model is well-known, and can be roughly described keeping into account only the term $F_0(h, J)$. In particular for low values of J the direct problem has only one stable solution (*paramagnetic phase*), while for high values of J two stable solutions for the empirical averages emerge (*ferromagnetic phase*). In the case $h = 0$ the two regimes are separated by a phase transition in which the fluctuations of the average magnetization diverge.

3.3.1 The mean-field solution

The solution of the direct problem considering only $F_0(h, J)$ will be called *mean-field* solution. Notice that due to the scaling $F_0(h, J) \propto N$, for large values of N this contribution dominates the free energy $F(h, J)$.

Proposition 3.3. *For all $i \neq j$ the mean-field solution for the fully connected ferromagnet is :*

$$\left\langle \sum_i s_i \right\rangle_0 = N m_{s.p.}(h, J) \quad (3.25)$$

$$\left\langle \frac{1}{N} \sum_{i < j} s_i s_j \right\rangle_0 = N \frac{m_{s.p.}^2(h, J)}{2} \quad (3.26)$$

while the susceptibility matrix is given by

$$\chi_0 = N \chi_{s.p.} \begin{pmatrix} 1 & m_{s.p.} \\ m_{s.p.} & m_{s.p.}^2 \end{pmatrix}, \quad (3.27)$$

where $m_{s.p.}$ is the absolute minimum of the function $f_{h,J}(m) = \frac{1+m}{2} \log \frac{1+m}{2} + \frac{1-m}{2} \log \frac{1-m}{2} - \frac{Jm^2}{2} - hm$ and $\chi_{s.p.} = \partial m_{s.p.} / \partial h$.

Remark 3.2. *It is easy to check that the mean-field solution describes independent spins. In fact equations (3.25) and (3.26) imply that for large N and $i \neq j$*

$$\langle s_i \rangle^2 = \langle s_i s_j \rangle. \quad (3.28)$$

This fact is a consequence of the pathological behavior of the mean-field solution of this model. In particular this implies that the inverse problem has a solution just along the line $(m, c) = (m, m^2)$, while it is easy to see (appendix B.1.3) that for a generic distribution $\bar{\mathbf{p}} \in \mathcal{M}(\Omega)$ the set of all possible empirical averages (i.e., the marginal polytope associated with the fully connected ferromagnet) is

$$\mathcal{G}(\phi) = \left\{ (m, c) \in \mathbb{R}^2 \mid m \in [-1, 1] \wedge c \in \left[\frac{m^2 - 1/N}{1 - 1/N}, 1 \right] \right\} \quad (3.29)$$

This implies the following fact concerning the inverse problem.

Proposition 3.4. *The inverse problem for the fully connected ferromagnet has a mean-field solution if and only if $(m, c) = (m, m^2)$. In that case, the entropy is given by*

$$S(m, m^2) = N \left(\frac{1+m}{2} \log \frac{1+m}{2} + \frac{1-m}{2} \log \frac{1-m}{2} \right) \quad (3.30)$$

while the couplings belong to the space

$$h^* = \operatorname{arctanh} m - \delta J m \quad (3.31)$$

$$J^* = \delta J \quad (3.32)$$

restricted to the region in which $\operatorname{sign}(m) = \operatorname{sign}(h)$. Finally, the inverse susceptibility matrix is divergent.

This last fact can be understood by checking that the matrix χ_0 has eigenvalue decomposition $N \left(0, \frac{1-m_{s,p}^4}{1-J+Jm_{s,p}^2} \right)$. In particular, the null eigenvalue has eigenvector $(-m, 1)$ which indicates that the mean field solution of the direct problem is invariant under the change of couplings

$$(h, J) \rightarrow (h - \delta J m_{s,p}, J + \delta J) . \quad (3.33)$$

Thus, the inverse problem maps all the points belonging to the one-dimensional region (m, m^2) on the two-dimensional plane (h, J) . This apparently contradicts the remark in section 2.2 about the existence of solutions to the inverse problem for any point belonging to the marginal polytope $\mathcal{G}(\phi)$. Indeed, we will show in the next section that keeping properly into account the presence of the $h = 0, J > 1$ line allows to understand this discrepancy. Interestingly, the two-dimensional region $\mathcal{G}(\phi) \setminus \{(Nm, \frac{N-1}{2}m^2) \mid m \in [-1, 1]\}$ is mapped on such one-dimensional line.

3.3.2 Finite N corrections

Keeping into account the terms F_{fluct} and F_{trans} allows to describe the transition from the finite N regime to the mean-field one. In particular, the Gaussian fluctuations around the mean-field solution extend the region in which the inverse problem is solvable to a strip of finite width in the space $\mathcal{G}(\phi)$.

Proposition 3.5. *Given $(Nm, \frac{N-1}{2}c) \in \mathcal{G}(\phi)$, the inverse problem for a fully connected ferromagnet described by the terms F_0 and F_{fluct} of equation (3.24) has solution if and only if $c = m^2 + \frac{\delta c}{N}$ with δc finite, and reads³*

$$h = \operatorname{arctanh} m - Jm \quad (3.35)$$

$$J = \frac{\delta c}{(1 - m^2)(1 - m^2 + \delta c)}. \quad (3.36)$$

Proof. This can easily be proved by keeping into account the contributions to the averages $\langle \dots \rangle_0, \langle \dots \rangle_{fluct}$ shown in appendix B.1 and imposing $m_{s.p.} = m + \delta m/N$, $c = m^2 + \delta c/N$ in the momentum matching condition. \square

The null eigenvalue of the matrix $\hat{\chi}_0$ is lifted to a finite value, as one can see that

$$\det(\hat{\chi}_0 + \hat{\chi}_{fluct}) = N \frac{\chi_{s.p.}^3}{2} > 0, \quad (3.37)$$

and is of order N (instead of N^2 as could be expected on the basis of the scaling of the leading term χ_0). Summarizing, data with small connected correlations (i.e., $c - m^2 \sim 1/N$) are described by a fully connected model with finite h . Conversely, it must hold that the whole space $\mathcal{G}(\phi)$ stripped of the quasi-one dimensional region $(Nm, \frac{2}{N-1}m^2 + \delta c)$ is mapped on the region of the (J, h) plane in which $J > 1$ and $h \sim 1/N$. To show this, we consider the approximation in which the only relevant terms of the free energy $F(h, J)$ are $F(h, J) = F_0(h, J) + F_{trans}(h, J)$.

Proposition 3.6. *The inverse problem for the fully connected ferromagnet described by the terms $F_0 + F_{trans}$ has solution for any point $(m, c) \in \mathcal{G}(\phi)$ excluding the region*

³In the literature concerning the so-called inverse Ising model, this result is typically derived by differentiating the relation

$$\operatorname{arctanh} m_i = h_i + \frac{1}{N} \sum_k J_{ik} m_k \quad (3.34)$$

with respect to m_j , and by recognizing that through linear response one can write $(\partial h / \partial m)_{ij}^{-1} = c_{ij} - m_i m_j \approx \delta c_{ij} / N$ [45, 68].

$c - m^2 \sim 1/N$. The points (h^*, J^*) satisfy the equations

$$m = m_{s.p.} - (h\chi_{s.p.} + m_{s.p.})[1 - \tanh(Nhm_{s.p.})] \quad (3.38)$$

$$c = m_{s.p.}^2 + hm_{s.p.}\chi_{s.p.}[1 - \tanh(Nhm_{s.p.})] \quad (3.39)$$

$$m_{s.p.} = \tanh(Jm_{s.p.} + h). \quad (3.40)$$

Also in this case one can show that in the limit $h \ll N$, the null mode of $\hat{\chi}_0$ is lifted due to

$$\det(\hat{\chi}_0 + \hat{\chi}_{trans}) \xrightarrow{N \rightarrow \infty} N^3 \chi_{s.p.} m_{s.p.}^4 \operatorname{sech}(hNm_{s.p.}). \quad (3.41)$$

Finally, one can draw the following conclusion, which despite being a trivial consequence of what shown above, shows that the $N \rightarrow \infty$ limit can lead to counter-intuitive results.

Remark 3.3. Consider the solution of the inverse problem for a fully connected ferromagnet and a point $\bar{\phi} = (Nm, \frac{N-1}{2}c)$ drawn from the space of empirical averages $\mathcal{G}(\phi)$ with uniform measure. Then for any $\epsilon > 0$, $J^*(\bar{\phi}) > 1$ and $h^*(\bar{\phi}) \in [-\epsilon, \epsilon]$ with probability $P \xrightarrow{N \rightarrow \infty} 1$.

This simple example shows some of the features discussed above concerning the limit of large N , namely:

1. The free energy loses (strict) concavity, as one has $\det \hat{\chi} \xrightarrow{N \rightarrow \infty} \det \hat{\chi}_0 = 0$. This indicates that some directions in the coupling space cannot be discriminated. In this example, when N is large, interactions are no longer distinguishable from external fields due to the presence of an eigenvector $(-m, 1)$ associated with the null eigenvalue.
2. Model condensation takes place, as all the region $\mathcal{G}(\phi)$ but a set of null measure is mapped on a one-dimensional strip. This will be better elucidated in chapter

5, where we will be able to quantify the density of models contained in a finite region of the space (h, J) .

3.4 Saddle-point approach to mean-field systems

In this section we generalize the procedure employed in the case of the fully connected ferromagnet to the case in which a saddle-point approach is used to solve the direct problem for a generic system. In particular, we consider a statistical model (ϕ, \mathbf{g}) with partition function

$$Z(\mathbf{g}) = \sum_{s \in \Omega} \exp \left(\sum_{\mu=1}^M g_{\mu} \phi_{\mu,s} \right), \quad (3.42)$$

and suppose that the operators $\phi_{\mu,s}$ can be written as functions of a small set of parameters $\boldsymbol{\psi}(s) = (\psi_1(s), \dots, \psi_A(s))$, so that for any μ one has $\phi_{\mu}(s) = \phi_{\mu}[\boldsymbol{\psi}(s)]$. Then it is possible to write

$$\begin{aligned} Z(\mathbf{g}) &= \int d\boldsymbol{\psi} \sum_s \exp \left(\sum_{\mu=1}^M g_{\mu} \phi_{\mu}(\boldsymbol{\psi}) \right) \delta(\boldsymbol{\psi} - \boldsymbol{\psi}(s)) \\ &= \int d\boldsymbol{\psi} \exp \left(\sum_{\mu=1}^M g_{\mu} \phi_{\mu}(\boldsymbol{\psi}) + \Sigma(\boldsymbol{\psi}) \right), \end{aligned} \quad (3.43)$$

where $e^{\Sigma(\boldsymbol{\psi})} = \sum_s \delta(\boldsymbol{\psi} - \boldsymbol{\psi}(s))$, and $\Sigma(\boldsymbol{\psi})$ is often referred as *entropy* for the value of the order parameter $\boldsymbol{\psi}$. For many statistical models, one has that the limit

$$f(\boldsymbol{\psi}) = \lim_{N \rightarrow \infty} \frac{1}{N} \left(- \sum_{\mu=1}^M g_{\mu} \phi_{\mu}(\boldsymbol{\psi}) - \Sigma(\boldsymbol{\psi}) \right) \quad (3.44)$$

is finite, and $f(\boldsymbol{\psi})$ is often called (intensive) *free-energy* for the value of the order parameter $\boldsymbol{\psi}$. In this case, one can exploit a saddle-point approximation to evaluate

the partition function $Z(\mathbf{g})$ at large N . It results

$$Z(\mathbf{g}) \xrightarrow{N \rightarrow \infty} e^{-Nf(\boldsymbol{\psi}^*)} \sqrt{\frac{2\pi}{N \det(f^{(2)}(\boldsymbol{\psi}^*))}}. \quad (3.45)$$

where we use the notation $f^{(n)}(\boldsymbol{\psi})$ for the tensor with components $f_{a_1, \dots, a_n}^{(n)} = \partial_{\psi_{a_1}} \dots \partial_{\psi_{a_n}} f(\boldsymbol{\psi})$, and $\boldsymbol{\psi}^*$ is the global minimum of the function $f(\boldsymbol{\psi})$, which in particular satisfies

$$\frac{\partial}{\partial \psi_a} f(\boldsymbol{\psi}) = 0. \quad (3.46)$$

Besides providing us with a mean to calculate the free energy $F(\mathbf{g})$, the ensemble averages $\langle \phi \rangle$ and the susceptibilities $\hat{\chi}$, the notions defined above allow us to introduce the concept of *state*, which we will use to characterize the phenomenon of ergodicity breaking.

Definition 3.2. Consider a statistical model (ϕ, \mathbf{g}) which can be described by a set of order parameters $\boldsymbol{\psi}$, and such that at large N its partition function can be approximated by (3.45). Then we call a *state* any local minima of the saddle-point equations (3.46).

We will label any of those minima as $\boldsymbol{\psi}^{(\alpha)}$ with $\alpha = 1, \dots, P$, and use the superscript α to identify quantities associated with the state α , as for example

$$F^{(\alpha)}(\mathbf{g}) = -\log Z^{(\alpha)}(\mathbf{g}). \quad (3.47)$$

In principle just the state with smallest free energy $F^{(\alpha)}$ should be relevant for the computation of the partition function (3.45). Indeed all the other states have an interpretation according to the dynamics which governs the system. Such states are relevant in order to model the phenomenon of ergodicity breaking, which occurs whenever the configurations of a large system $s \in \Omega$ cannot be sampled according to the

probability distribution $p(s)$ in experiments of finite length T .⁴

In particular we informally remind that for large statistical models (ϕ, \mathbf{g}) endowed with a realistic dynamics (e.g., Metropolis-Hastings [80, 36]) leading in the limit of exponentially large T to the stationary distribution \mathbf{p} associated with (ϕ, \mathbf{g}) , states naturally emerge when observing a finite amount of configurations. In fact, the iteration of a dynamics for $T \ll 2^N$ time steps typically produces configurations belonging to the same state as the initial one, while in the opposite limit of large T the probability of observing a state belonging to a configuration α is proportional to $e^{Nf(\psi^{(\alpha)})}$. Hence, unless data obtained from an experiment are exponentially large in the size of the system (which isn't typically the case in real world applications of the inverse problem), one expects empirical averages to concentrate around averages which are in principle different from the ensemble ones, and that are associated with a specific state α . Accordingly, we define the notion of state average $\langle \phi^{(\alpha)} \rangle$, which is expected in the regime of $T \ll 2^N$ to model the averages obtained by experiments of finite length as follows:

Definition 3.3. Given a system (ϕ, \mathbf{g}) whose partition function can be approximated by the partition function (3.45), we define the *state averages*

$$\langle \phi_{\mu}^{(\alpha)} \rangle = -\frac{\partial F^{(\alpha)}}{\partial g_{\mu}} \quad (3.48)$$

and the state susceptibilities

$$\chi_{\mu, \nu}^{(\alpha)} = -\frac{\partial^2 F^{(\alpha)}}{\partial g_{\mu} \partial g_{\nu}}. \quad (3.49)$$

The correctness of above construction has been verified for several statistical models subject to different dynamics [57, 89], nevertheless to the best of our knowledge

⁴ We won't explicitly refer to the dynamics leading to the loss of ergodicity, even though this phenomenon is naturally associated with the stochastic process leading to the stationary distribution (2.1) and is more naturally discussed in the framework of a Markov chain [34].

no fully general, rigorous result concerning this phenomenon is available yet. In particular, in order to rigorously motivate the notion of state average, it would be necessary to show that for a generic, local dynamics a decomposition property of the form $p_s \xrightarrow{N \rightarrow \infty} \sum_{\alpha=1}^P q_\alpha p_s^{(\alpha)}$ where $\sum_{\alpha=1}^P q_\alpha = 1$ and $p^{(\alpha)} \in \mathcal{M}(\Omega)$ holds for the Gibbs measure, which again is known to be correct just for specific models.

In that case, the state averages and the susceptibilities can be explicitly computed by explicitly deriving the above free-energy, allowing to prove the following result.

Proposition 3.7. *The direct problem for a statistical model (ϕ, \mathbf{g}) which can be described with order parameters $\boldsymbol{\psi}$ and an order parameter free-energy $f(\boldsymbol{\psi})$ can be solved in saddle-point approximation in any state α , leading to*

$$F^{(\alpha)}(\mathbf{g}) = Nf(\boldsymbol{\psi}^{(\alpha)}) + \frac{1}{2} \log \det f^{(2)}(\boldsymbol{\psi}^{(\alpha)}) - \frac{1}{2} \log \frac{2\pi}{N} \quad (3.50)$$

$$\langle \phi_\mu^{(\alpha)} \rangle = \phi_\mu(\boldsymbol{\psi}^{(\alpha)}) + \frac{1}{N} \left[(f^{(2)})_{a,b}^{-1} \phi_{\mu;a,b}^{(2)} - (f^{(2)})_{a,b}^{-1} f_{b,a,d}^{(3)} (f^{(2)})_{d,e}^{-1} \phi_{\mu;e}^{(1)} \right] \quad (3.51)$$

$$\begin{aligned} \chi_{\mu,\nu}^{(\alpha)} &= \left[(f^{(2)})_{a,b}^{-1} \phi_{\mu;b,a}^{(2)} - [(f^{(2)})_{a,b}^{-1}]_{f}^{(1)} f_{b,a,d}^{(3)} (f^{(2)})_{d,e}^{-1} \phi_{\mu;e}^{(1)} \right. \\ &\quad - (f^{(2)})_{a,b}^{-1} f_{b,a,d;\nu}^{(3,1)} (f^{(2)})_{d,e}^{-1} \phi_{\mu;e}^{(1)} - (f^{(2)})_{a,b}^{-1} f_{b,a,d}^{(3)} [(f^{(2)})_{d,e}^{-1}]_{\nu}^{(1)} \phi_{\mu;e}^{(1)} \\ &\quad + \frac{1}{N} \left[[(f^{(2)})_{a,b}^{-1}]_{c}^{(1)} (f^{(2)})_{c,d}^{-1} \phi_{\nu;d}^{(1)} \phi_{\mu;b,a}^{(2)} + (f^{(2)})_{a,b}^{-1} \phi_{\mu;a,b,c}^{(3)} (f^{(2)})_{c,d}^{-1} \phi_{\nu;d}^{(1)} \right. \\ &\quad - [(f^{(2)})_{a,b}^{-1}]_{f}^{(1)} f_{b,a,d}^{(3)} (f^{(2)})_{d,e}^{-1} \phi_{\mu;e}^{(1)} f_{f,g}^{(2)} \phi_{\nu;g}^{(1)} - (f^{(2)})_{a,b}^{-1} f_{b,a,d,f}^{(4)} (f^{(2)})_{d,e}^{-1} \phi_{\mu;e}^{(1)} f_{f,g}^{(2)} \phi_{\nu;g}^{(1)} \\ &\quad \left. - (f^{(2)})_{a,b}^{-1} f_{b,a,d}^{(3)} (f^{(2)})_{d,e}^{-1} \phi_{\mu;e}^{(1)} f_{f,g}^{(2)} \phi_{\nu;g}^{(1)} - (f^{(2)})_{a,b}^{-1} f_{b,a,d}^{(3)} [(f^{(2)})_{d,e}^{-1}]_{f}^{(1)} \phi_{\mu;e,f}^{(2)} f_{f,g}^{(2)} \phi_{\nu;g}^{(1)} \right] \end{aligned} \quad (3.52)$$

where $\phi_{\mu;a_1,\dots,a_n}^{(n)}$ indicates the tensor $\frac{\partial}{\partial \psi_{a_1}} \dots \frac{\partial}{\partial \psi_{a_n}} \phi_\mu(\boldsymbol{\psi}^{(\alpha)})$,

$f_{a_1,\dots,a_m;\mu_1,\dots,\mu_N}^{(m,n)} = \frac{\partial}{\partial \psi_{a_1}} \dots \frac{\partial}{\partial \psi_{a_m}} \frac{\partial}{\partial g_{\mu_1}} \dots \frac{\partial}{\partial g_{\mu_n}} f(\boldsymbol{\psi}^{(\alpha)})$ and by convention repeated index are summed.

This result allows us to characterize the behavior of the inverse problem in the large N limit. In fact one can see that at leading order in N , the momentum matching

condition (2.27) becomes

$$\langle \phi_{\mu}^{(\alpha)} \rangle \xrightarrow{N \rightarrow \infty} \phi_{\mu}(\boldsymbol{\psi}^{(\alpha)}) = \bar{\phi}, \quad (3.53)$$

where we remark that the averages in the state α do not depend explicitly on \mathbf{g} , being their dependence contained in the order parameter $\boldsymbol{\psi}^{(\alpha)}$. This implies that, given two statistical models (ϕ, \mathbf{g}) and (ϕ, \mathbf{g}') with $\mathbf{g} \neq \mathbf{g}'$ such that there exist a couple of states (respectively α and α') solving the saddle point equations with $\boldsymbol{\psi}^{(\alpha)} = \boldsymbol{\psi}^{(\alpha')}$, in the large N limit those models cannot be discriminated.

Remark 3.4. *Consider an empirical dataset $\bar{\phi}^{(\alpha)}$ generated by a system in state α . Unless one doesn't consider a matching condition in which the state average contains the corrections of order $1/N$ indicated in the right term of formula (3.51), it is not generally guaranteed that it is possible to reconstruct the state α which generated the empirical averages.*

A rough criteria which can be used in order to check the expected number of solutions for the inverse problem is provided by the comparison of the number of solutions of the saddle-point equations P , the number of order parameters A and the number of couplings M . If in particular $M > A$, then the saddle-point equations are expected to have a continuous number of solutions \mathbf{g}^* specifying the same value of the order parameters for any of the P states $\boldsymbol{\psi}^{(\alpha)}$. If $M < A$ a unique set of couplings is expected to be associated with a value of an order parameters. Finally, if $M = A$, then a finite number of solutions for the couplings has to be expected.

3.4.1 Ergodicity breaking for a fully connected pairwise model

Consider the fully-connected pairwise model of section 3.3. In that case the construction above can be trivially applied by considering the only order parameter

$\psi(s) = \sum_{i=1}^N s_i$ (so that $A = 1$). The saddle-point equation for this model

$$m = \tanh(Jm + h) \tag{3.54}$$

can have either one solution m^* (thus, $P = 1$) or two stable solutions m_+^* and m_-^* ($P = 2$) according to the values of h and J . We consider as an illustrative example the case in which $J = J_- = 4$ and $h = h_- = 0.1$, hence $P = 2$ solutions are present. For this model the metastable state is characterized by $m_-^* \approx -0.9991754$, and it is easy to show that any pair (J, h) satisfying

$$m_-^* = \tanh(Jm_-^* + h) \tag{3.55}$$

has the same saddle point magnetization. In particular, it is possible to find $h < 0$ solutions corresponding to the stable $\alpha = +$ state characterized by the same value of the magnetization. For example, the stable state of the model $(J_+ \approx 3.39950, h_+ = -0.5)$ has magnetization $m_-(J_-, h_-) = m_+(J_+, h_+)$. In figure 3.1 we show how the models (J_-, h_-) and (J_+, h_+) lead to the same value of the state averages m and c in the thermodynamic limit $N \rightarrow \infty$: not even the state of a large fully connected ferromagnet can be reconstructed on the basis of a finite length experiment, unless the state averages are known with large precision. The difference of this result with

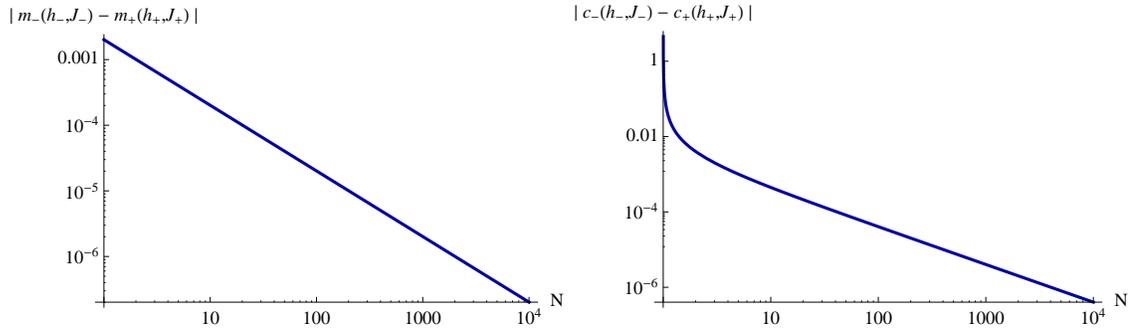


Figure 3.1: Absolute difference among the ensemble averages of models describing two different states of a fully connected model, as a function of the system size N .

respect to what found in section (3.3) lies in the fact that state averages can be matched by any solution of the form

$$h^* = \operatorname{arctanh} m - \delta J m \tag{3.56}$$

$$J^* = \delta J, \tag{3.57}$$

regardless of the sign of h^* (while in that case it had to be taken $\operatorname{sign}(h^*) = \operatorname{sign}(\bar{m})$). In both cases a continuous number of solutions for the inverse problem is present.

3.5 Disorder and heterogeneity: the regime of large N and large M

The results presented in section 3.3 for the Curie-Weiss model refer to a specific statistical model whose associated inverse problem shows interesting features in the limit of large N . Despite the fact that such properties generally hold for similar kind of models (section 3.4) one could wonder whether this behavior is retained in the more relevant case in which a large number of inferred parameters is present. Consider for example a general pairwise model (2.47), characterized by a set of N external fields and $\frac{N(N-1)}{2}$ pairwise couplings. In this case one may have several problems in studying the features introduced in section 3.2 as we did above. In particular:

- The averages \mathbf{m} , $\hat{\mathbf{c}}$ and the generalized susceptibility $\hat{\chi}$ are hard to compute for a generic value of $\hat{\mathbf{J}}$ and \mathbf{h} . Therefore, it is not possible to understand which points of marginal polytope $\mathcal{G}(\phi)$ are associated with zero modes in $\hat{\chi}$. Moreover, the limit $N \rightarrow \infty$ is ambiguously defined if no prescription is provided for how should the empirical averages scale with N .

- For the same reason, it is not possible to find in which points one expects model concentration to occur, as this would require knowing which eigenvalues of $\hat{\chi}$ are divergent in the thermodynamic limit $N \rightarrow \infty$ for generic points $(\mathbf{m}, \hat{\mathbf{c}}) \in \mathcal{G}(\phi)$.
- No saddle-point approach is justified for generic empirical averages $\mathbf{m}, \hat{\mathbf{c}}$. Thus, an approach analogous to the one in 3.4 cannot be considered, and the notion of state cannot be described in such terms.

These difficulties could be overcome by resorting to the notion of *disorder*, which is commonly used in the field of statistical mechanics of heterogeneous systems. In particular we want to show, as a possible outlook of this work, an approach to the analysis of the large M limit borrowed from that field [56] which could be applied to this problem.

3.5.1 Self-averaging properties and inverse problem

Given an operator set ϕ , consider a set of statistical models $\mathcal{M}(\phi)$ and a prior $P_0(\mathbf{g})$ on this space. Then, suppose that a statistical model (ϕ, \mathbf{g}) is sampled according to $P_0(\mathbf{g})$, and successively a set of empirical data of length T is drawn by such distribution. Several functions of the estimator $\mathbf{g}^*(\bar{\phi})$ can be built in order to analyze the properties of an instance of the inverse problem, such as the quantities

$$\Delta(\bar{\phi}, \mathbf{g}) = \sqrt{\sum_{\mu=1}^M \frac{1}{M} (g_{\mu}^*(\bar{\phi}) - g_{\mu})^2}, \quad (3.58)$$

which quantifies the average error in the inferred coupling and

$$\frac{1}{M} \log \det \hat{\chi}(\mathbf{g}^*(\bar{\phi})) = \frac{1}{M} \text{tr} \log \hat{\chi}(\mathbf{g}^*(\bar{\phi})), \quad (3.59)$$

whose divergence signals critical properties of the generalized susceptibility matrix $\hat{\chi}$. If these quantities are self-averaging for large N and T (i.e., they concentrate around

an average value determined by $P_0(\mathbf{g})$, then one expects that specific instances of inverse problems drawn by the same prior $P_0(\mathbf{g})$ to share the same collective features. As an example, if one considers a Gaussian prior for the ferromagnetic model of the type $P_0(\mathbf{h}, \hat{\mathbf{J}}) \propto \exp\left(-N \sum_{i < j} \frac{(J_{ij} - J_0/N)^2}{2\delta J^2}\right) \exp\left(-\sum_i \frac{(h_i - h_0)^2}{2\delta h^2}\right)$ with $J_0 \neq 0$, then it is known that the macroscopic behavior of the model approaches in the large N limit the one of a fully-connected ferromagnet (3.23) defined by the only parameters (h_0, J_0) [56]. In section 5.3 we will support this claim through a specific example, showing a case in which the properties of a homogeneous model allow to describe very accurately the collective features of the inverse problem for an heterogeneous one. Nevertheless, it would be interesting to repeat the calculations shown in the previous sections in this more general scenario in which disorder is present, and prove through the so-called *replica* formalism [56] the correctness of these expectations.

Remark 3.5. *The idea of disorder in the context of the inverse problem is obviously linked to the existence of a prior $P_0(\mathbf{g})$ on the space $\mathcal{M}(\phi)$, so that in principle the case of a flat prior cannot be treated with these techniques. Nevertheless, fixing implicitly a specific class of models through $P_0(\mathbf{g})$ is the price to pay to answer to very interesting questions, which wouldn't otherwise be well-posed namely: (i) can a specific model be learnt with high probability according to a given inference prescription? (ii) Are the global properties of an heterogeneous system equivalent the the ones of an homogeneous one? (iii) Is it possible to understand the generic properties of $\hat{\chi}$?*

Chapter 4

Complete representations

In this chapter we will introduce the notion of complete family of operators, which can be used to gain some insight about the inverse problem. Although in general this approach may introduce a high degree of over fitting, dealing with complete families allows to discuss very transparently some features of inference which are related to algorithmic complexity (section 4.2.1). Moreover, completeness allows for an explicit reparametrization of the probability distribution (2.1) in terms of state probabilities, allowing for a complete understanding of properties of the inverse problem which are less clear by using the Gibbs form for the probability density. More interestingly, in this language we will be able to differentiate local features of the direct and of the inverse problem, which in turn rely on the locality of the marginals. In this chapter the inverse problem for some models will be exactly and explicitly solved, while some ideas will be presented in order to generalize this methods to more relevant problems (sections 4.2.3 and 4.2.4). In section 4.3 we will present some specific examples illustrating these ideas.

4.1 Orthogonality and completeness

We define in this section the notion of *orthogonality* and *completeness* for families of operators. While the orthogonality condition is related to the one of minimality, the one of completeness will allow to formally invert the relation among ensemble averages $\langle \phi \rangle$ and couplings \mathbf{g} .

Definition 4.1. Given a family of operators ϕ , we call it *orthogonal* if it satisfies

$$\frac{1}{|\Omega|} \sum_s \phi_{\mu,s} \phi_{\nu,s} = \delta_{\mu,\nu} , \quad (4.1)$$

while it will be called *complete* if it holds

$$\frac{1}{|\Omega|} \sum_{\mu} \phi_{\mu,s} \phi_{\mu,s'} = \delta_{s,s'} . \quad (4.2)$$

Property (4.1) can be seen as expressing the fact that in an orthogonal family any pair of operators decorrelate when averaged with respect to a uniform probability density (at infinite temperature in the language of statistical mechanics). Additionally, if $\phi_0 \in \phi$, one can see that in an orthogonal family, for $\mu \neq 0$

$$\frac{1}{|\Omega|} \sum_s \phi_{\mu,s} = 0 , \quad (4.3)$$

i.e., ϕ_{μ} has zero mean at infinite temperature for any $\mu \neq 0$. Finally, if ϕ is an orthogonal family, then it is easy to see that $\phi \setminus \{\phi_0\}$ is minimal. The main result that derives instead from equation (4.2) is the explicit one-to-one mapping between couplings \mathbf{g} , state probabilities \mathbf{p} and averages $\langle \phi \rangle$, as clarified by the next proposition.

Proposition 4.1. *Given a family ϕ satisfying (4.1) and (4.2), the statistical model $(\phi \setminus \{\phi_0\}, \mathbf{g})$ associated with the probability density \mathbf{p} satisfies*

$$\langle \phi_\mu \rangle = \sum_s \phi_{\mu,s} \exp \left(\sum_\nu g_\nu \phi_{\nu,s} \right) \quad (4.4)$$

$$g_\mu = \frac{1}{|\Omega|} \sum_s \phi_{\mu,s} \log \left(\frac{1}{|\Omega|} \sum_\nu \langle \phi_\nu \rangle \phi_{\nu,s} \right). \quad (4.5)$$

Additionally, state probabilities can be expressed as

$$p_s = \frac{1}{|\Omega|} \sum_\mu \langle \phi_\mu \rangle \phi_{\mu,s}. \quad (4.6)$$

Proof. These relations are a direct consequence of the axioms (4.1) and (4.2) and can be checked by direct substitution. □

Monomials

Throughout most of the following discussion, we will focus on families of operators ϕ formed by *monomials*, for which axiom (4.1) trivially applies. More precisely, given a cluster of spins Γ , we define the monomial $\phi_\Gamma(s)$ as

$$\phi_\Gamma(s) = \prod_{i \in \Gamma} s_i, \quad (4.7)$$

while the identity is associated with the empty cluster $\phi_0(s) = \phi_\emptyset(s) = 1$. It is easy to show the following:

Proposition 4.2. *Given a collections of clusters $(\Gamma_0, \dots, \Gamma_M)$ with $\Gamma_i \neq \Gamma_j \forall (i, j)$ it holds for the family $\phi = \{\phi_{\Gamma_0}, \dots, \phi_{\Gamma_M}\}$ that*

- ϕ is an orthogonal family;
- ϕ is complete if and only if it contains all possible monomials, whose number is $|\Omega| = 2^N$.

Moreover, monomials satisfy a very important relation which will be used extensively in the following.

Proposition 4.3. *Consider a complete family of monomials ϕ . Then the marginals of the probability density \mathbf{p} associated with the model $(\phi \setminus \{\phi_0\}, \mathbf{g})$ can be expressed as*

$$p^\Gamma(s^\Gamma) = \frac{1}{2^{|\Gamma|}} \sum_{\Gamma' \subseteq \Gamma} \langle \phi_{\Gamma'} \rangle \phi_{\Gamma', s} \quad (4.8)$$

Proof. This can be checked by using equation (4.6) and showing that for each monomial it holds

$$\frac{1}{2} \sum_{s_i} \phi_{\Gamma, s} = \delta_{i \notin \Gamma} \phi_{\Gamma, s} . \quad (4.9)$$

□

This property expresses the locality of marginals once they are expressed in terms of ensemble averages. This should be compared with the expression of a marginal written as a function of the couplings (2.4), in whose form the locality properties are hidden by the interaction structure.

4.2 Inference on complete models

4.2.1 The complete inverse problem

The techniques shown in the above section can be used to write a formal solution of the inverse problem in full generality. The main drawback of this procedure is the overfitting issue which has to be associated with the presence of an exponential number of couplings, which in practical cases makes this approach unfeasible unless the system has small size (typically $N \sim 10^1$). Indeed, as the solution of the complete inverse problem illustrates with simplicity some very general features of many inverse problem, we choose to present its solution.

Definition 4.2. The *complete inverse problem* is the inverse problem associated with the statistical model defined by the complete family of monomials $\{\phi_\Gamma(s)\}_{\Gamma \subseteq V}$. Its probability density can be written as

$$p(s) = \exp \left(\sum_{\Gamma \subseteq V} g_\Gamma \phi_\Gamma(s) \right). \quad (4.10)$$

It is easy to write the formal solution for the entropy by using the relation (4.2), while its differentiation (or the direct use of the relation (4.5)) leads to an exact expression for the couplings and the susceptibility matrix.

Proposition 4.4. *The expression for the entropy of the complete inverse problem reads*

$$S(\bar{\phi}) = -\frac{1}{|\Omega|} \sum_s \left(\sum_{\Gamma \subseteq V} \bar{\phi}_\Gamma \phi_{\Gamma,s} \right) \log \left(\frac{1}{|\Omega|} \sum_{\Gamma' \subseteq V} \bar{\phi}_{\Gamma'} \phi_{\Gamma',s} \right), \quad (4.11)$$

while the inferred couplings \mathbf{g}^* and the inverse susceptibility matrix $\hat{\chi}^{-1}$ result

$$g_\Gamma^* = \frac{1}{|\Omega|} \sum_s \phi_{\Gamma,s} \log \left(\frac{1}{|\Omega|} \sum_{\Gamma' \subseteq V} \bar{\phi}_{\Gamma'} \phi_{\Gamma',s} \right) \quad (4.12)$$

$$\chi_{\Gamma,\Gamma'}^{-1} = \frac{1}{|\Omega|^2} \sum_s \frac{\phi_{\Gamma,s} \phi_{\Gamma',s}}{\frac{1}{|\Omega|} \sum_{\Gamma'' \subseteq V} \bar{\phi}_{\Gamma''} \phi_{\Gamma'',s}}. \quad (4.13)$$

This solution has a simple interpretation in terms of empirical frequencies, once one rewrites above expression using the relation $\bar{p}_s = |\Omega|^{-1} \sum_{\Gamma' \subseteq V} \bar{\phi}_{\Gamma'} \phi_{\Gamma',s}$ as

$$g_\Gamma^* = \frac{1}{|\Omega|} \sum_s \phi_{\Gamma,s} \log \bar{p}_s \quad (4.14)$$

$$\chi_{\Gamma,\Gamma'}^{-1} = \frac{1}{|\Omega|^2} \sum_s \frac{\phi_{\Gamma,s} \phi_{\Gamma',s}}{\bar{p}_s}. \quad (4.15)$$

In this form it is possible to appreciate that the solution simply corresponds to a matching of state probabilities with empirical probabilities.

Remark 4.1. *This last observation can be made more precise by exploiting the identity*

$$\log P_T(\hat{\mathbf{s}}|\mathbf{g}) = T \sum_{\Gamma} g_{\Gamma} \bar{\phi}_{\Gamma} = T \sum_s \log p_s \bar{p}_s , \quad (4.16)$$

which can be used to express the log-likelihood function as a function of the probabilities \mathbf{p} instead of the coupling vector \mathbf{g} . Its maximization can be seen equivalently as performed over the state probabilities \mathbf{p} . In this case, the obvious solution is $\mathbf{p}^ = \bar{\mathbf{p}}$, so that the expression (4.14) is describing an approach in which the state probabilities are matched with the empirical ones one-by-one. In particular, if a configuration is not observed, the inferred probability for that configuration is strictly zero.*

Divergencies

The formal solution (4.14) shows that the inferred couplings can be infinite if there are states which are never sampled in the data $\hat{\mathbf{s}}$. In particular if data are generated by an actual probability distribution \mathbf{p} assigning zero weight to some configuration, the Ω space splits into an accessible and a non-accessible sector, and divergencies can be seen as are required to implement an hard constraint on the set of accessible configurations. Couplings obtained by using this scheme are finite either when all states are measured or when divergencies cancel out for a given region of the coupling space. Indeed, the presence of an unaccessible sector has to be considered a spurious result unless $p_s \approx \bar{p}_s$, which is expected to hold just in the large T limit. In particular for $T < |\Omega|$, $\bar{p}_s = 0$ for at least $|\Omega| - T > 0$ configurations, regardless of the presence or absence of a forbidden sector. Therefore it is not possible to distinguish if divergencies are due to the presence of an unaccessible sector or to poor sampling. In this case, regularization schemes such as the use of Laplacian smoothing or an L-2 norm can be used to obtain finite results. This basically corresponds to lift the probability for non-measured configurations from zero to some finite value. For example, Laplacian

smoothing procedure [70] corresponds to the choice:

$$p_s^* = \frac{\bar{p}_s + \lambda}{1 + |\Omega|\lambda} \quad (4.17)$$

Finally, we remark that the same type of divergence arises in all the cases that will be analyzed (see sections 4.2.3 and 4.2.4), and is a very general characteristic of inverse problems, which typically relates to under sampling. This is the simplest setting in which this problem can be analyzed in full generality.

Observed sector

The expression for the inferred couplings (4.14) involves a summation over all the configuration space Ω , so that a summation over $|\Omega| = 2^N$ terms seems to be required to calculate any of them. Indeed, those expressions may be rewritten exploiting the orthogonality relation (4.1), which implies that

$$\frac{1}{|\Omega|} \sum_{s \in \bar{\mathcal{I}}} \phi_{\Gamma,s} = \delta_{\Gamma,0} - \frac{1}{|\Omega|} \sum_{s \notin \bar{\mathcal{I}}} \phi_{\Gamma,s} \quad (4.18)$$

where $\bar{\mathcal{I}} = \{s \in \Omega \mid \bar{p}_s > 0\}$ is the set of observed configuration. Then, one can rewrite (4.14) as

$$g_{\Gamma}^* = \frac{1}{|\Omega|} \sum_{s \in \bar{\mathcal{I}}} \phi_{\Gamma,s} \log \bar{p}_s + \log \bar{p}_0 \left(\delta_{\Gamma,0} - \frac{1}{|\Omega|} \sum_{s \in \bar{\mathcal{I}}} \phi_{\Gamma,s} \right), \quad (4.19)$$

where the term proportional to $\bar{p}_0 = 0$ account for the divergencies, and the sum over states runs over a number $|\bar{\mathcal{I}}| \leq T$ terms. In the case of the regularized complete inverse problem (section 4.2.2), we will see that it will be possible to write an analogous expression for the couplings, in which the weight assigned to non-observed configuration will be finite.

Rate of convergence

Given an underlying statistical model \mathbf{p} for the complete inverse problem, large deviation theory (as described in section 2.2.4) states that for large T the variance of the inferred couplings \mathbf{g} with respect to the measure given by $P_T(\mathbf{p}|\bar{\mathbf{p}}) \propto P_T(\bar{\mathbf{p}}|\mathbf{p})$ is

$$\text{Var}(g_\Gamma^*) = \frac{\chi_{\Gamma,\Gamma}^{-1}}{T} = \frac{1}{T} \left(\frac{1}{|\Omega|^2} \sum_s \frac{1}{p_s} \right). \quad (4.20)$$

Incidentally, the same quantity can also be obtained by averaging with respect to the $\langle \dots \rangle_T$ measure, a result which allows to express the rate of convergence for the complete inverse problem (appendix D.1). While the $1/T$ pre factor expresses the expected scaling for the error on the inferred coupling, the $\chi_{\Gamma,\Gamma}^{-1}$ term is non trivial. In particular we observe that:

1. The fluctuations of the inferred couplings are identical for all the operators.
2. The value of the fluctuations is bound by the inequality:

$$\frac{1}{T} \leq \text{Var}(g_\Gamma^*) \leq \frac{1}{T|\Omega|p_{min}} \quad (4.21)$$

where $p_{min} = \min_s p_s$.

3. The speed of convergence is limited by the presence of rare configurations. In particular if $p_{min} = 0$, the variance diverges.

The generalization to the case in which the sector of observable states $\mathcal{I} = \{s \in \Omega \mid p_s > 0\}$ is smaller than the entire phase space $\mathcal{I} \subset \Omega$ is straightforward (appendix D.1). Indeed, it is necessary to define a set of *regular* operators ϕ^{reg} such that $\phi^{reg} = \{\phi_\Gamma \in \phi \mid \sum_{s \in \mathcal{I}} \phi_{\Gamma,s} = 0\}$. For couplings associated with regular operators it

holds the asymptotic property

$$\text{Var}(g_{\Gamma}^{\star reg}) = \frac{1}{T|\Omega|^2} \sum_{s \in \mathcal{I}} \frac{1}{p_s} . \quad (4.22)$$

If the sector of observable states has cardinality $|\mathcal{I}| = \alpha|\Omega|$, then the fluctuations on the regular couplings satisfy the bound

$$\frac{\alpha^2}{T} \leq \text{Var}(g_{\Gamma}^{\star reg}) \leq \frac{\alpha}{T|\Omega|p_{min}} \quad (4.23)$$

where $p_{min} = \min_{s \in \mathcal{I}} p_s$. Even in the cases analyzed in section 4.2.3 and 4.2.4 the presence of rare configurations will limit the speed of convergence of the inferred couplings to their actual value.

4.2.2 Regularization of the complete inverse problem

The generality of the complete inverse problem renders its regularization relevant for a strong theoretical reason. In fact, the complete inverse problem is totally non-parametric in the sense that the probability distribution (4.10) contains all possible statistical models describing a set of N binary variables. Then one could think of selecting the most appropriate statistical model to describe a dataset of binary data simply by applying a suitable regularizer to this general problem, and let the regularization term itself perform the task of model selection (an approach successfully adopted in [63, 86] in a less general scenario). We present in the following the results obtained by using different regularization terms, and comment about the interpretation of the solutions of the regularized inverse problem. Finally, we will characterize a symmetry property of regularizers which can be used to study their suitability in the field of high-dimensional inference (i.e., for large values of N).

L-2 regularization

The simplest regularized version of the complete inverse problem is the one defined by the function

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu} + \frac{\beta}{2} \sum_{\mu=1}^M g_{\mu}^2 \quad (4.24)$$

which implements the Gaussian prior over the L-2 norm of the coupling vector described in section 2.3.2. In terms of state probabilities, equation (4.24) can be written as

$$H(\mathbf{p}|\hat{\mathbf{s}}) = -T \sum_s \log p_s \bar{p}_s + \frac{\beta}{2} \left[\left(\frac{1}{|\Omega|} \sum_s \log^2 p_s \right) - \left(\frac{1}{|\Omega|} \sum_s \log p_s \right)^2 \right] \quad (4.25)$$

and its minimization with respect to p_s (constrained to $\sum_s p_s = 1$) leads to the set of implicit equations

$$p_s^* = \bar{p}_s - \frac{\beta}{T|\Omega|} \left(\log p_s^* - \frac{1}{|\Omega|} \sum_{s'} \log p_{s'}^* \right). \quad (4.26)$$

Its solution determines the value of the couplings g through the relation

$$g_{\mu}^* = \frac{1}{|\Omega|} \sum_s \phi_{\mu,s} \log p_s^*. \quad (4.27)$$

We observe that:

1. The summation over the configuration space requires considering in principle an exponential number of terms, but this issue can be avoided as explained in section 4.2.1.
2. The expression for g_{μ}^* is always finite, as the presence of infinite couplings is suppressed by the cost associated with the L-2 norm.

3. The parameter β controls the total value of the L-2 norm of the coupling vector \mathbf{g}^* and the entropy of the inferred distribution. In particular the total L-2 norm can be expressed as

$$\sum_{\mu} g_{\mu}^{*2} = \frac{1}{|\Omega|} \sum_s \log^2 p_s^*, \quad (4.28)$$

where the statistical weights \mathbf{p}^* are fixed by equation (4.26).

4. The additional problem of solving the system of equations for p_s requires in principle the numerical solution of $|\Omega| = 2^N$ equations. Indeed, all equations linked with unobserved configurations are equal, and defining as above the probability p_0 for non-measured configurations, the number of independent equations that have to be solved is $|\bar{\mathcal{I}}| + 1 \leq |\mathcal{I}| + 1 \leq T + 1$.

This considered, the expression for the couplings obtained using this regularization scheme is

$$g_{\mu}^* = \frac{1}{|\Omega|} \sum_{s \in \mathcal{I}} \phi_{\mu,s} \log \left(\frac{p_s^*}{p_0^*} \right) + \delta_{\mu 0} \log p_0^*, \quad (4.29)$$

where the p_s and the p_0 satisfy the set of implicit equations:

$$\begin{cases} p_s^* &= -\frac{\beta}{|\Omega|T} \left(\log p_s^* - \frac{1}{|\Omega|} \sum_{s' \in \mathcal{I}} \log p_{s'}^* - \frac{|\Omega| - |\mathcal{I}|}{|\Omega|} \log p_0^* \right) + \bar{p}_s \\ p_0^* &= -\frac{\beta}{|\Omega|T} \left(\log p_0^* - \frac{1}{|\Omega|} \sum_{s' \in \mathcal{I}} \log p_{s'}^* - \frac{|\Omega| - |\mathcal{I}|}{|\Omega|} \log p_0^* \right). \end{cases} \quad (4.30)$$

We remark that the calculation of the regularized couplings can be performed in polynomial time in T .

Entropy regularization

Another choice for the regularization is motivated by the following argument. If a dataset $\hat{\mathbf{s}}$ of length T is associated with an entropy $S(\bar{\mathbf{p}}) \sim \log T$, with $\log T \ll N$, it is likely for the model to be in the under sampled regime, as the entropy per variable

is expected to be finite (i.e., $S(\mathbf{p}) \sim N$) for well-behaved models. Then, it is possible to consider a regularizing term which penalizes low entropy distribution, so that

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu} - \beta S(p) , \quad (4.31)$$

where as usual \mathbf{p} is the density associated with the statistical model (ϕ, \mathbf{g}) , so that $S(\mathbf{p}) = -\sum_{\mu=0}^M g_{\mu} \langle \phi_{\mu} \rangle$. The minimization of above expression with respect to g_{μ} leads to

$$\bar{\phi}_{\mu} = \langle \phi_{\mu} \rangle + \frac{\beta}{T} \sum_{\nu=1}^M g_{\nu} \frac{\partial \langle \phi_{\nu} \rangle}{\partial g_{\mu}} . \quad (4.32)$$

After some manipulation and after using the completeness relation (4.2) one finds that

$$\bar{p}_s = p_s + \frac{\beta}{T} p_s \log p_s - \frac{\beta}{T} p_s \left(\sum_{s'} p_{s'} \log p_{s'} \right) . \quad (4.33)$$

Finally, by writing $s_s = -p_s \log p_s$, one is led to a set of implicit equations

$$p_s = \frac{\bar{p}_s + \frac{\beta}{T} s_s}{1 + \frac{\beta}{T} \sum_{s'} s_{s'}} , \quad (4.34)$$

which is analogous to the one described in the L-2 case. Also in this case the system has to be solved numerically, by exploiting the fact that the probabilities p_s depend on the s index through the empirical frequency \bar{p}_s (i.e., states visited the same number of times are associated with the same inferred probability). Equation (4.2) can finally be used to extract the inferred couplings from the probability density of p_s .

Susceptibility regularization

The inverse generalized susceptibility of a model $\hat{\chi}^{-1}$ provides an indication of the generalizability of an inference procedure through equation (2.32), which implies that

the response of the inferred couplings \mathbf{g}^* to a shift of the empirical averages $\bar{\phi}$ is

$$\chi_{\mu,\nu}^{-1} = \frac{\partial g_{\mu}^*}{\partial \bar{\phi}_{\nu}} . \quad (4.35)$$

Then one could think to favor generalizability in an inference procedure by introducing a regularization term of the form

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu} + \beta \text{tr}(\hat{\chi}^{-1}) . \quad (4.36)$$

By employing equation (4.13) it is easy to see that the inverse susceptibility matrix can be written as a function of the coupling vector g as

$$\chi_{\mu,\nu}^{-1} = \frac{1}{|\Omega|^2} \sum_s \phi_{\mu,s} \phi_{\nu,s} \exp\left(-\sum_{\rho=0}^M g_{\rho} \phi_{\rho,s}\right) , \quad (4.37)$$

so that the total energy can be written as

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu} + \beta \left(\frac{|\Omega|-1}{|\Omega|^2}\right) \sum_s p_s^{-1} . \quad (4.38)$$

Its minimization leads to

$$\bar{\phi}_{\mu} = \langle \phi_{\mu} \rangle + \frac{\beta}{T} \left(\frac{|\Omega|-1}{|\Omega|^2}\right) \sum_s [\langle \phi_{\mu} \rangle - \bar{\phi}_{\mu}] p_s^{-1} \quad (4.39)$$

whose solution requires solving a set of implicit equations analogous to (4.30) and (4.34) of the form

$$p_s = \frac{\bar{p}_s + \frac{\beta}{T} \left(\frac{|\Omega|-1}{|\Omega|^2}\right) p_s^{-1}}{1 + \frac{\beta}{T} \left(\frac{|\Omega|-1}{|\Omega|^2}\right) \sum_{s'} p_{s'}^{-1}} . \quad (4.40)$$

By using equation (4.6) the solution \mathbf{p}^* can be used to explicitly express \mathbf{g}^* .

Remark 4.2. Notice that this regularization scheme artificially pushes the inferred couplings \mathbf{g}^* towards regions of the space $\mathcal{M}(\phi)$ in which fluctuations are high. This is a very general feature of inference procedures which favor the stability of the inferred model: requiring a model to be stable forces the generalizes susceptibility to be large, or equivalently, ensemble averages to have strong fluctuations.

L-1 regularization

We will write the L-1 regularized problem for the complete inverse problem described as in section 2.3.2, with the idea that its solution it would be equivalent to a complete, non-parametric solution of the problem of binary inference. In the more optimistic scenario, the problem of model selection would be implicitly solved by the L-1 norm, without the need of explicitly breaking the symmetry among the operators by choosing (*a priori*) the more relevant ones, as it is usually done by means of the maximum entropy principle (appendix A.1). Relevant operators should arise as conjugated to non-zero couplings in a regularized problem of the form

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu} + \beta \sum_{\mu=1}^M |g_{\mu}| . \quad (4.41)$$

The minimization of above expression with respect to \mathbf{g} leads to

$$p_s \in \bar{p}_s - \frac{\beta}{T|\Omega|} \sum_{\mu=1}^M \text{sgn}(g_{\mu}) \phi_{\mu,s} , \quad (4.42)$$

where we define the set valued function $\text{sgn}(x)$ as in appendix C. We remark several issues concerning this regularizer:

1. Unlike the L-2 case, the completeness relation does not allow to switch from a summation on operators to a summation over configurations, hence algebraic properties cannot be fully exploited to manipulate the above equation.

2. The minimization condition is a system of $|\Omega|$ implicit equations in which the inferred values of p_s on non-observed configurations are generally different. This is due to the term $\sum_{\mu}(\text{sgn } g_{\mu})\phi_{\mu,s}$, which is different even for $s \notin \mathcal{I}$ (see example 4.3.2).
3. L-1 norm is associated with a compact description of the probability distribution (it is used to enforce sparsity in the number of non-zero couplings), while in the case $T \ll 2^N$ one deals with few observations of the system (sparsity in the number of observed configurations). As the change of parametrization (4.6) from \mathbf{p} to \mathbf{g} is strongly non-local (i.e., what is sparse in a parametrization is not sparse in the other one), the problem becomes hard to solve due to *frustration*, alias the simultaneous request of incompatible conditions in a constraint satisfaction problem.
4. Even if a fast (i.e., polynomial in N) algorithm to find a solution for a single coupling g_{μ} was available, a preliminary selection of the couplings to focus on would nevertheless be needed. In fact, even in the scenario in which the calculation of a single g_{μ} can be achieved in polynomial time, a constrained optimization problem should be formulated in order to select *which* subset of couplings is non-zero given a specific value of β .

Explicit selection of couplings

An interesting case is the one in which a specific inverse problem – such as the inverse pairwise model – is seen as a regularized version of the complete inverse problem. This implicitly implies that unlike with the previous regularizers, in this particular example we are not interested in the problem of model selection, but we mean to offer a different perspective on a problem which is known to be hard, in order to characterize it from a different point of view. In particular, we consider the regularized minus-log-

likelihood

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu} + \frac{\beta}{2} \sum_{\mu=1}^M \theta_{\mu} g_{\mu}^2 \quad (4.43)$$

in which $\theta \in \{0, 1\}^M$ determines the couplings that are penalized by the L-2 norm, and we consider the limit of large, positive β , so that $g_{\mu}^* \approx 0$ if $\theta = 1$. The minimization of (4.43) leads to the set of equations

$$\bar{\phi}_{\mu} = \sum_s p_s^* \phi_{\mu,s} + \frac{\beta}{T} \theta_{\mu} g_{\mu}^*, \quad (4.44)$$

which in the parametrization of states becomes

$$\bar{p}_s = p_s^* + \frac{\beta}{T|\Omega|} \sum_{\mu} \phi_{\mu,s} \theta_{\mu} g_{\mu}^*. \quad (4.45)$$

The last term in (4.45) is finite in the limit of large β , and encodes the constraint specified by θ . Within this formulation the intrinsic difficulty of an inverse, non-complete problem emerges as the fact that the probabilities p_s^* can be different for states visited with the same frequency. This is associated with the dependence of the second term of equation (4.45) upon the index s associated with the operators $\phi_{\mu,s}$, and is analogous to the case of the L-1 norm described above.

Remark 4.3. *A formal solution for this problem can be written by studying the limit $\beta \rightarrow \infty$, which is associated with couplings $g_{\mu}^* = 0$ for $\theta_{\mu} \neq 0$. The equation $g_{\mu}^* = 0$ can be expressed in term of operator averages by using equation (4.5) as follows:*

$$1 = \prod_s \left(\frac{1}{|\Omega|} \sum_{\nu} \langle \phi_{\nu}^* \rangle \phi_{\nu,s} \right)^{\phi_{\mu,s}}, \quad (4.46)$$

where $\langle \phi_{\nu}^* \rangle$ indicates the ensemble average of the operator ϕ_{ν} under the distribution p_s^* . This result expresses a relation among observables which must hold whenever couplings

are zero, which is typically used to express higher order correlations in terms of low order ones, which can be expressed as roots of polynomial equations. Then equation (4.5) can be used to write the remaining couplings, and the roots of equation (4.46) can in principle be used to obtain an expression for the non-zero components of \mathbf{g}^* .

Symmetry properties of the regularizers

The limit of large N of the regularized complete inverse problem provides an insight on the structure of the regularizers which have been examined in the previous sections. In particular, we can consider the regime in which N is large, while T scales polynomially in N ($T \sim N^\alpha$) so that $T \ll |\Omega| = 2^N$, and provide an argument about the behavior of the regularized inverse problem. Indeed, we will first need to define the notion of symmetric regularizer.

Definition 4.3. Consider the complete inverse problem defined by the model $(\{\phi_\Gamma\}_{\Gamma \subseteq V} \setminus \phi_\emptyset, \mathbf{g})$ and a regularizer $H_0(\mathbf{g})$. Then, we call $H_0(\mathbf{g})$ a *symmetric* regularizer if for any pair of states s and s' it holds

$$\bar{p}_s = \bar{p}_{s'} \Rightarrow p_s^* = p_{s'}^* \quad (4.47)$$

For example, the L-2 regularizer, the entropy regularizer and the susceptibility regularizer analyzed above are symmetric regularizers. Obviously, the non-regularized problem $H_0(\mathbf{g}) = 0$ is also symmetric. The following proposition holds for symmetric regularizers.

Proposition 4.5. Consider the complete inverse problem defined by the model $(\{\phi_\Gamma\}_{\Gamma \subseteq V} \setminus \phi_\emptyset, \mathbf{g})$ and a symmetric regularizer $H_0(\mathbf{g})$. Suppose additionally that the empirical probability vector $\bar{\mathbf{p}}$ has elements only in $\bar{\mathbf{p}} \in \{0, 1/T\}^{|\Omega|}$. Then the solution

of the regularized inverse problem is given by

$$g_\mu^* \propto \bar{\phi}_\mu. \quad (4.48)$$

This result intuitively indicates that symmetric regularizers are unable to distinguish correlations and interactions unless states are sampled more than once. As in the large N regime described above one expects (for well-behaved probability distributions) single states to appear either one or zero times, then this indicates that non-parametric inference procedures should be performed with non-symmetric regularizers in order to extract informative results about interactions. From another perspective, this shows that in the extremely under sampled limit $T \ll |\Omega|$, the more biased couplings are the ones associated with biased empirical averages. Notice that while in the case of the explicit coupling selection the regularizer is expected not to be symmetric by construction (states are biased according to their overlaps with the explicitly selected operators), it is interesting to see that the L-1 norm breaks the state symmetry without the need of biasing specific operators (example 4.3.2).

Proof. To prove the above proposition it is sufficient to notice that by symmetry the coupling vector \mathbf{g}^* depends on the two values p_0^* and $p_{1/T}^*$ associated with the states sampled zero ($\bar{p}_s = 0$) and once ($\bar{p}_s = 1/T$). Then equation (4.6) implies that

$$\begin{aligned} g_\mu^* &= \frac{1}{|\Omega|} \log \left(\frac{p_{1/T}^*}{p_0^*} \right) \sum_{s \in \bar{\mathcal{I}}} \phi_{\mu,s} + \delta_{\mu 0} \log p_0^* \\ &= \frac{T}{|\Omega|} \log \left(\frac{p_{1/T}^*}{p_0^*} \right) \bar{\phi}_\mu + \delta_{\mu 0} \log p_0^*. \end{aligned} \quad (4.49)$$

□

Remark 4.4. *The symmetry broken by the L-1 regularizer and by the explicit coupling selection is associated with the following consideration: in principle, unless there is an explicit information that allows to distinguish between states s and s' that are*

observed the same number of times, then inference should assign the same weight to those states. In the first case such symmetry is spontaneously broken (the information injected by the prior doesn't specifically favor any state), while in the second it is explicitly broken.

4.2.3 Pairwise model on trees

One of the simplest cases in which the pairwise model defined by equation (2.47) can be explicitly solved is when the topology of the interaction matrix \mathbf{J} is the one of a tree. In that case it is well known that *message passing* algorithms [54] can find the solution to the direct problem in a time linear in N . Indeed, there are several reasons which make the inverse problem worth studying. The first one is the observation that the factorization property (4.51) allows to write an explicit, closed form solution of the inverse problem. The second one is the exceptional stability of the inverse problem with respect to the direct one. Finally, the a full analogy with the complete case can be discussed, and a general scheme for the structure of solutions for inverse problems can be sketched speculating on this simple example.

Definition 4.4. Consider the pairwise model described in section 2.2.5, defined by the probability density

$$p(s) = \frac{1}{Z(\mathbf{h}, \hat{\mathbf{J}})} \exp \left(\sum_{i \in V} h_i s_i + \sum_{(i,j) \in E} J_{ij} s_i s_j \right), \quad (4.50)$$

in the case in which the set of edges E does not contain any cycle. Then this model is called a *tree* (see appendix D.2 for a more precise definition).

For such models the inverse problem is easy to solve due to the factorization property shown in appendix D.2, which allows to write the probability density as

$$p(s) = \prod_{(i,j) \in E} p^{\{i,j\}}(s_i, s_j) \prod_{i \in V} [p^{\{i\}}(s_i)]^{1-|\partial i|}, \quad (4.51)$$

where $\partial i = \{\phi_{\{i,j\}} \in \Phi \mid (i,j) \in E\}$. Hence, the entropy can be written as

$$S(\mathbf{m}, \hat{\mathbf{c}}) = \sum_{(i,j) \in E} S^{\{i,j\}}(m_i, m_j, c_{ij}) + \sum_{i \in V} (1 - |\partial i|) S^{\{i\}}(m_i) \quad (4.52)$$

and the inverse problem can be solved, as shown in the next proposition.

Proposition 4.6. *For a the pairwise model of the form (2.47) with a tree topology, the entropy $S(\mathbf{m}, \hat{\mathbf{c}})$ can be written as*

$$\begin{aligned} S(\mathbf{m}, \hat{\mathbf{c}}) &= \sum_{(i,j) \in E} \sum_{s_i, s_j} \left[\frac{1}{4} (1 + m_i s_i + m_j s_j + c_{ij} s_i s_j) \right] \log \left[\frac{1}{4} (1 + m_i s_i + m_j s_j + c_{ij} s_i s_j) \right] \\ &+ \sum_{i \in V} (1 - |\partial i|) \sum_{s_i} \left[\frac{1}{2} (1 + m_i s_i) \right] \log \left[\frac{1}{2} (1 + m_i s_i) \right], \end{aligned} \quad (4.53)$$

while the fields \mathbf{h}^* and the couplings \mathbf{J}^* result

$$\begin{aligned} h_i^* &= \frac{1}{4} \sum_{j \in \partial i} \sum_{s_i, s_j} s_i \log \left[\frac{1}{4} (1 + m_i s_i + m_j s_j + c_{ij} s_i s_j) \right] \\ &+ \frac{1}{2} (1 - |\partial i|) \sum_{s_i} s_i \log \left[\frac{1}{2} (1 + m_i s_i) \right] \\ J_{ij}^* &= \frac{1}{4} \sum_{s_i, s_j} s_i s_j \log \left[\frac{1}{4} (1 + m_i s_i + m_j s_j + c_{ij} s_i s_j) \right], \end{aligned} \quad (4.54)$$

and the inverse susceptibility matrix $\hat{\chi}^{-1}$ is given by

$$\begin{aligned}
 \chi_{\{i,j\},\{k,l\}}^{-1} &= \frac{1}{16} \sum_{s_i, s_j} \frac{\delta_{i,k} \delta_{j,l} + \delta_{i,l} \delta_{j,k}}{\bar{p}^{\{i,j\}}(s_i, s_j)} \\
 \chi_{\{i,j\},\{k\}}^{-1} &= \frac{1}{16} \sum_{s_i, s_j} \frac{\delta_{i,k} s_j + \delta_{j,k} s_i}{\bar{p}^{\{i,j\}}(s_i, s_j)} \\
 \chi_{\{i\},\{j\}}^{-1} &= \frac{1}{16} \sum_{k \in \partial i} \sum_{s_i, s_k} \frac{\delta_{i,j} + s_i s_k \delta_{k,j}}{\bar{p}^{\{i,k\}}(s_i, s_k)} + \frac{1}{4} (1 - |\partial i|) \sum_{s_i} \frac{\delta_{i,j}}{\bar{p}^{\{i\}}(s_i)}
 \end{aligned} \tag{4.55}$$

The structure of this solution is reminiscent of the one shown in the case of the complete inverse problem described in section 4.2.1, and can intuitively be understood as follows. To solve an inverse problem it is necessary to find the clusters which allow to express the entropy (in that case all clusters had to be included, while in this case single spin and two spins clusters alone are sufficient to write the full entropy). Couplings are obtained as sums over cluster contributions, in which each of them contributes with a value proportional to the average of the conjugated operator, weighted by the log-probability of each cluster configuration. Inverse generalized susceptibilities quantify the amount of cluster fluctuations, and are large if local fluctuations are rare.

The presence of a large number of delta functions is due to the fact that the entropy is built by a small number of cluster contributions, so that the response of the couplings to a shift in the value of the conjugated average is strongly localized: either the perturbation is applied to a neighbor, in whose case the response is finite, or it is zero. This has to be compared with the direct problem, in which a perturbation in the couplings changes the average of a finite number of operators in general. In that case, one roughly expects that

$$\chi_{\{i\},\{j\}} \propto e^{|i-j|/\xi}, \tag{4.56}$$

where ξ is the correlation length of the system. This was first noted in [25, 26], where it is shown that for a large number of statistical models that the structure of $\hat{\chi}$ is dense, while the one of $\hat{\chi}^{-1}$ tends to be sparse.

4.2.4 One-dimensional periodic chain with arbitrary range couplings

An interesting application of the inference scheme presented in this chapter concerns the solution of the inverse problem for one-dimensional chains. Despite the fact that an exact solution of this problem has been first presented in [35], we will be interested in providing a rigorous proof relying on completeness properties. Also in this case a complete analogy with the previous example can be drawn. Consider a set of binary spins $s \in \Omega$ and a family of operators of range R (i.e. acting on the first R spins) $\phi(s_1, \dots, s_R) = (\phi_1(s_1, \dots, s_R), \dots, \phi_M(s_1, \dots, s_R))$ subject to the periodic boundary conditions $s_i = s_{i+N}$. Then the notion of one-dimensional chain can be introduced through the action of *translation operators* $\mathbf{T} = \{T_n\}_{n=0}^{N/\rho-1}$, defined through their action on the ϕ

$$T_n \phi_\mu(s_1, \dots, s_R) = \phi_\mu(s_{1+n\rho}, \dots, s_{R+n\rho}), \quad (4.57)$$

which corresponds to a shift of the argument of ϕ on the next set of $n\rho$ spins, so that $\rho < R$ is characterized as the periodicity of the chain.

Definition 4.5. A one-dimensional chain is defined as the probability distribution on the space $s \in \Omega$

$$p(s) = \frac{1}{Z(\mathbf{g})} \exp \left(\sum_{\mu=1}^M g_\mu \sum_{n=0}^{N/\rho-1} T_n \phi_\mu(s) \right), \quad (4.58)$$

where \mathbf{T} is a set of translation operators characterized by a periodicity parameter ρ and ϕ is a set of M operators of range R .

We are interested in solving the inverse problem for this type of system, which means to calculate the entropy $S(\mathbf{T}\bar{\phi})$ as a function of the empirical averages of the operators $\mathbf{T}\phi = \sum_{n=0}^{N/\rho} T_n\phi$. In order for the entropy to be well-behaved, and in order to exploit the property of completeness (4.2), we need to require a specific choice for the set ϕ .

Definition 4.6. A one-dimensional chain defined by a family of operators ϕ and translation operators \mathbf{T} is *orthogonal* and *complete* if

- For any $m, n \in (0, \dots, N/\rho - 1)$, $\sum_s T_n\phi_{\mu,s}T_m\phi_{\nu,s} = \delta_{m,n}\delta_{\mu,\nu}$
- For any generic operator $\phi \neq 1$ of range R , and any $m \in (0, \dots, N/\rho - 1)$, there exist n and μ such that $T_m\phi = T_n\phi_\mu$.

A possible choice for a family ϕ satisfying those requirements is provided by a suitable choice of monomials. More precisely, one can define a set $\Gamma_0 = \{1, \dots, R\}$ and a set $\gamma_0 = \{\rho + 1, \dots, R\}$, so that the family of operators $\phi = \{\phi_\Gamma\}_{\Gamma \subseteq \Gamma_0} \setminus \{\phi_\gamma\}_{\gamma \subseteq \gamma_0}$ describes the $|\phi| = 2^R(1 - 2^{-\rho})$ monomials belonging to Γ_0 which are not contained in γ_0 (appendix D.3). Intuitively, this corresponds to define the problem through all operators located inside the unit cell, so that any other operator of range R can be generated in a unique way by using the translation operators \mathbf{T} .

For a one-dimensional chain, it is possible to prove (appendix D.3) that the probability density \mathbf{p} can be factorized as

$$p(s) = \prod_{n=0}^{N/\rho-1} \frac{p^{\Gamma_n}(s^{\Gamma_n})}{p^{\gamma_n}(s^{\gamma_n})}, \quad (4.59)$$

where $\Gamma_n = T_n\Gamma_0 = \{1 + n\rho, \dots, R + n\rho\}$ while $\gamma_n = T_n\gamma_0 = \{1 + (n+1)\rho, \dots, R + n\rho\}$. Consequently the entropy can be written as

$$S(\mathbf{T}\bar{\phi}) = \sum_{n=0}^{N/\rho-1} [S^{\Gamma_n}(\mathbf{p}^{\Gamma_n}) - S^{\gamma_n}(\mathbf{p}^{\gamma_n})]. \quad (4.60)$$

This relation, together with equation (4.9) which expresses the locality of marginals, allows to explicitly find the expression of the entropy of a one-dimensional chain.

Proposition 4.7. *The inverse problem for an orthogonal, complete one-dimensional chain of monomials has the following solution. The entropy can be expressed as¹*

$$\begin{aligned}
 S(\mathbf{T}\bar{\phi}) &= \frac{N}{\rho} \left\{ \sum_{s^{\Gamma_0}} \left[\frac{1}{2^R} \sum_{\Gamma \in \Gamma_0} \sum_{\mu \in \phi} c_{\mu, \Gamma} \bar{\phi}_\mu \phi_{\Gamma, s^{\Gamma_0}} \right] \log \left[\frac{1}{2^R} \sum_{\Gamma \in \Gamma_0} \sum_{\mu \in \phi} c_{\mu, \Gamma} \bar{\phi}_\mu \phi_{\Gamma, s^{\Gamma_0}} \right] \right. \\
 &\quad \left. - \sum_{s^{\gamma_0}} \left[\frac{1}{2^{R-\rho}} \sum_{\gamma \in \gamma_0} \sum_{\mu \in \phi} c_{\mu, \gamma} \bar{\phi}_\mu \phi_{\gamma, s^{\gamma_0}} \right] \log \left[\frac{1}{2^{R-\rho}} \sum_{\gamma \in \gamma_0} \sum_{\mu \in \phi} c_{\mu, \gamma} \bar{\phi}_\mu \phi_{\gamma, s^{\gamma_0}} \right] \right\}, \tag{4.61}
 \end{aligned}$$

where $c_{\mu, \Gamma} = 1$ if $\exists n$ such that $T_n \phi_\mu = \phi_\Gamma$ and $c_{\mu, \Gamma} = 0$ otherwise. The couplings result

$$\begin{aligned}
 g_\mu^* &= \sum_{s^{\Gamma_0}} \left[\frac{1}{2^R} \sum_{\Gamma \in \Gamma_0} c_{\mu, \Gamma} \phi_{\Gamma, s^{\Gamma_0}} \right] \log \left[\frac{1}{2^R} \sum_{\Gamma \in \Gamma_0} \sum_{\nu \in \phi} c_{\nu, \Gamma} \bar{\phi}_\nu \phi_{\Gamma, s^{\Gamma_0}} \right] \\
 &\quad - \sum_{s^{\gamma_0}} \left[\frac{1}{2^{R-\rho}} \sum_{\gamma \in \gamma_0} c_{\mu, \gamma} \phi_{\gamma, s^{\gamma_0}} \right] \log \left[\frac{1}{2^{R-\rho}} \sum_{\gamma \in \gamma_0} \sum_{\nu \in \phi} c_{\nu, \gamma} \bar{\phi}_\nu \phi_{\gamma, s^{\gamma_0}} \right] \tag{4.62}
 \end{aligned}$$

while the inverse susceptibilities are given by

$$\begin{aligned}
 \chi_{\mu, \nu}^{-1} &= \frac{\rho}{N} \left\{ \sum_{s^{\Gamma_0}} \frac{\left[\frac{1}{2^R} \sum_{\Gamma \in \Gamma_0} c_{\mu, \Gamma} \phi_{\Gamma, s^{\Gamma_0}} \right] \left[\frac{1}{2^R} \sum_{\Gamma \in \Gamma_0} c_{\nu, \Gamma} \phi_{\Gamma, s^{\Gamma_0}} \right]}{\bar{p}^{\Gamma_0}(\mathbf{T}\bar{\phi})} \right. \\
 &\quad \left. - \sum_{s^{\gamma_0}} \frac{\left[\frac{1}{2^{R-\rho}} \sum_{\gamma \in \gamma_0} c_{\mu, \gamma} \phi_{\gamma, s^{\gamma_0}} \right] \left[\frac{1}{2^{R-\rho}} \sum_{\gamma \in \gamma_0} c_{\nu, \gamma} \phi_{\gamma, s^{\gamma_0}} \right]}{\bar{p}^{\gamma_0}(\mathbf{T}\bar{\phi})} \right\}. \tag{4.63}
 \end{aligned}$$

Also in this case the structure of the solution is analogous to the one found in section 4.2.1 for the complete inverse problem and in section 4.2.3 for the inverse pairwise tree. The expression of the entropy is a sum of cluster contributions associated with unit cells. Such contributions are all equal due to periodicity, so that

¹ Notice that with abuse of notation we are writing $\bar{\phi}_\mu$ instead of $\frac{\rho}{N} \sum_n T_n \bar{\phi}_\mu$.

two clusters only (Γ^0 and γ^0) are sufficient to write the exact expression for the full entropy. The stability of the problem is instead determined by the fluctuations inside Γ^0 and γ^0 , and divergencies occur whenever any state in Γ^0 is not observed.

Remark 4.5. *In the case of a one-dimensional chain, the role which in the previous examples was played by of the number of observations T is played by the quantity TN/ρ , which measures the number of sampled unit cells. For this type of system even in the case of one single observation ($T = 1$) the noise on the inferred couplings can be small if the system is large enough.*

Remark 4.6. *In order to apply these ideas to empirical data, the information about the one-dimensional nature of the problem should be a priori known. Indeed, the exact nature of the interactions needs not to be known, provided that the R parameter is larger than the actual range of the interactions.*

4.3 Applications

4.3.1 Complete inverse problem

The techniques shown in section 4.2.1 have been tested on synthetic datasets in order to check their performance. As expected, they are suitable for systems in the small N regime, due to the slow convergence in T of the inferred coupling vector \mathbf{g}^* to the true coupling vector \mathbf{g} , which can be seen as a consequence of the over fitting problem associated with the presence of an exponential number of couplings. We have considered for simplicity a system of $N = 8$ spins, with couplings corresponding to several models, namely:

1. **Pure Noise:** A model with $g_\Gamma = 0$ describing the flat distribution $p_s = 1/|\Omega|$.
2. **Pairwise model:** A model with two body interactions (i.e., $g_\Gamma = 0$ if $|\phi_\Gamma| > 2$), and couplings equal to $g_\Gamma = 1/N$.

3. **Arbitrary couplings and hidden sector:** A model with infinite couplings associated to four random operators, in order to test the behavior of the algorithm in presence of divergent couplings.

In all those cases, we were able to compute by enumeration the partition function of the model, and to sample from the exact probability distribution a set of $T \in \{100, \dots, 50000\}$ states which have been used to construct the vectors of empirical frequencies \bar{p} and empirical averages $\bar{\phi}$. Formulas derived in the above sections have been used to solve the inverse problem for those sets of sampled states. For the case 1. of a flat probability distribution, we were able to check formula (4.20) describing the concentration of the couplings towards their expected value $g_\mu = 0$, as shown in figure 4.1. Beyond the naive inference scheme described in section 4.2.1, we have employed

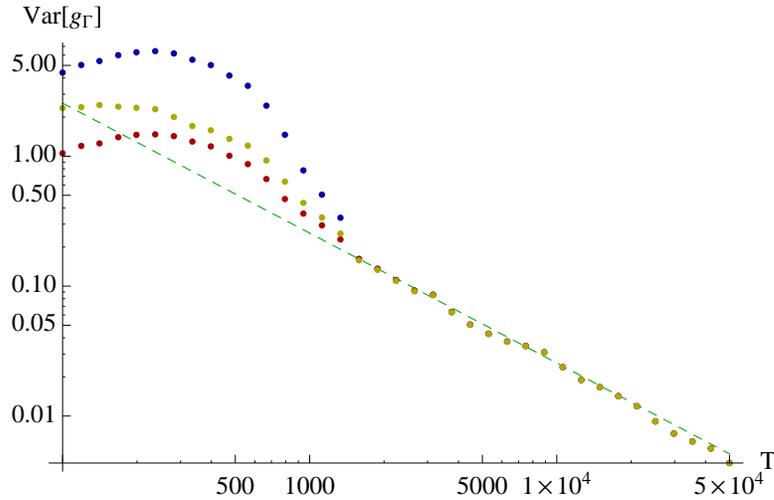


Figure 4.1: Variance of the inferred coupling vector as a function of the number of samples T for a flat probability distribution. Un-regularized inference procedure (4.14) corresponds to the blue line, the yellow one indicates an L-2 regularization scheme with $\beta = 10$ while the red one is obtained by using a cutoff in the divergencies of the form $p_0 \propto \log \epsilon = -\frac{1}{2} \log T$. The green line corresponds to the expected scaling for the error (4.20) in the case of a flat distribution.

an L-2 regularization scheme (yellow line) and a simple cutoff for divergencies of the form $\log \epsilon = -\frac{1}{2} \log T$ (red line). This last prescription is motivated by the simple consideration that for a multinomial distribution the variance on the empirical

probabilities scales as T^{-1} , so that the error on the sampled probability p_0 is expected to be of the order of $T^{-1/2}$. In figure 4.2, we plot an histogram of the couplings obtained for various values of T in order to show the shape of the posterior $P_T(\mathbf{g}|\hat{\mathbf{s}})$. Finally, we show in figure 4.3 that there is no cluster size $|\Gamma|$ which dominates the coupling vector for any value of T , implying that no model is favored by this inference scheme. For the case 2. of a pairwise model (section 2.2.5), we have considered a

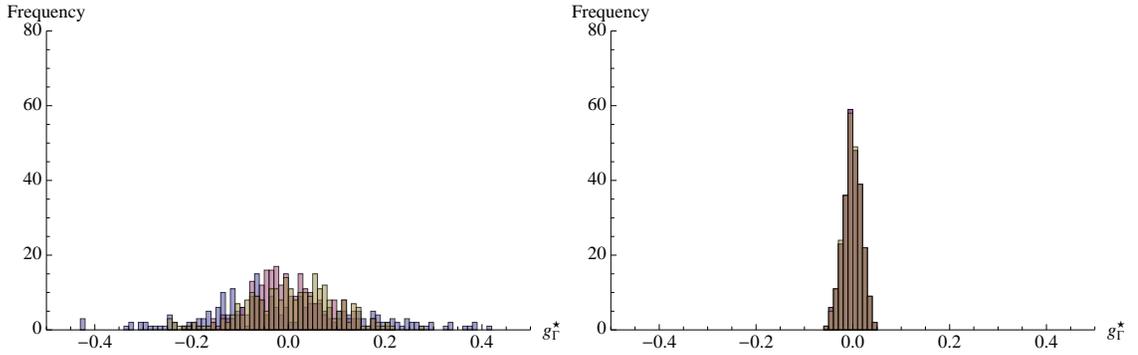


Figure 4.2: We plot the histogram of the inferred couplings for a complete model with $N = 8$ and $g_\Gamma = 0$, hence describing the posterior probability $P_T(\mathbf{g}|\hat{\mathbf{s}})$ for $T = 237$ (left panel) and $T = 2657$ (right panel). We employed the same type of regularizers as in figure 4.1.

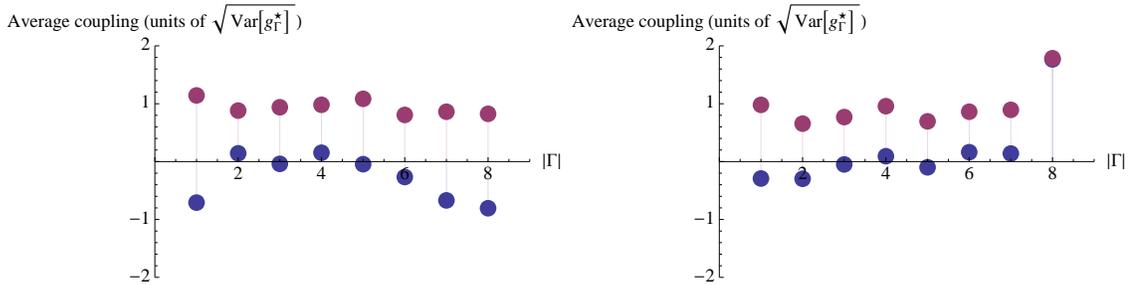


Figure 4.3: We plot the average mean inferred coupling $\binom{N}{k}^{-1} \sum_{|\Gamma|=k} g_\Gamma^*$ (blue line) and the average mean absolute coupling (red dots) $\binom{N}{k}^{-1} \sum_{|\Gamma|=k} |g_\Gamma^*|$ for $T = 237$ (left panel) and $T = 2657$ (right dots) in units of the error $\sqrt{\text{Var}(g_\Gamma^*)}$. The figure indicates that no specific size for the cluster Γ is preferred, as predicted by the expression for the error (4.20).

model with $h_i = 0 \forall i$ and $J_{ij} = 1/N \forall i < j$. We performed the same analysis and collected the same statistics as in the previous case. In figure 4.4 we plot the variance

of the inferred coupling distribution against the number of samples T , finding that as indicated by the inequality (4.21), the pre factor $\frac{1}{|\Omega|^2} \sum_s \frac{1}{p_s}$ controlling the convergence to zero of the errors is higher than for a flat probability distribution. Also in this case

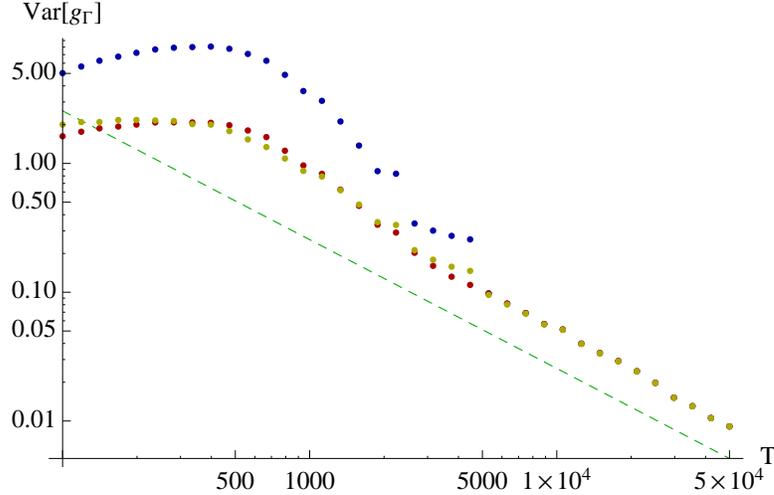


Figure 4.4: Variance of the inferred coupling vector as a function of the length of the number of samples T , for a pairwise model with $N = 8$, $h_i = 0$ and $J_{ij} = 1/N$. See figure 4.1 for the color convention and the type of regularizers adopted. The green line shows the expected scaling of the variance for a flat distribution, indicating that the reconstruction of a pairwise model is affected by a higher error than the one of a flat distribution.

we plot the histogram of the inferred coupling for various values of T , comparing the unimodal distribution of couplings in the noise-dominated regime ($T \lesssim 10^3$) with the bimodal distribution emerging for large sample size ($T \gtrsim 10^3$), in which the shrinking noise peak leaves room for the genuine signal concentrated in $g_\Gamma \approx 1$. The plot of the mean value and the mean absolute value of the couplings with fixed cluster size shows that even in this case no particular cluster size is biased except for $|\Gamma| = 2$.

Finally, we show how this procedure might be employed in the case in which one or more couplings are infinite. We consider complete models in which all couplings g_Γ are put to zero, but a random set which are set to $g_\Gamma = \infty$. As an illustrative example, we consider the case $g_{\{1\}} = g_{\{7\}} = \{3, 6\} = \{1, 4, 5, 7\} = \infty$, which lead to a set of observable states \mathcal{I} with $|\mathcal{I}| = 2^4$, and a set of regular (i.e., non divergent)

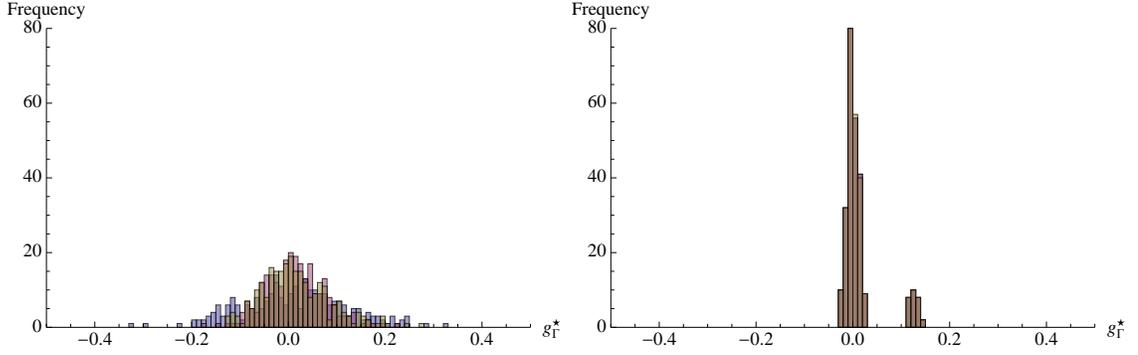


Figure 4.5: Histogram of the inferred couplings for the pairwise model described in figure 4.4 for $T = 1121$ (left panel) and $T = 14934$ (right panel), where the color convention is also described. Notice the transition from a unimodal distribution in the noise-dominated regime to the bimodal distribution obtained for large T .

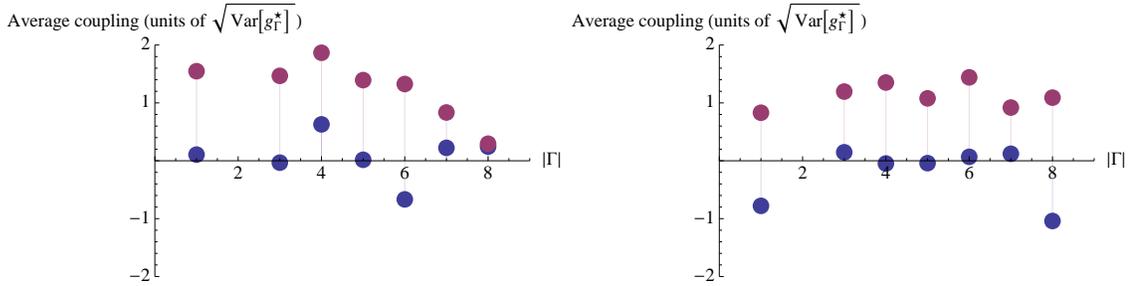


Figure 4.6: Average mean inferred coupling (blue points) and average mean absolute coupling (red points) for $T = 1121$ (left panel) and $T = 14934$ (right panel) in units of the error $\sqrt{\text{Var}(g_\Gamma^*)}$. Just clusters with $|\Gamma| = 2$ are favored (and hence out of scale in this plot).

couplings \mathbf{g}^{reg} of size $|\mathbf{g}^{reg}| = 240$. We plot in figure 4.7 the variance of the regular inferred couplings to their exact value against the length of the dataset T , while in figure 4.8 we show how non-regular couplings approach infinity.

4.3.2 L-1 norm vs L-2 norm: emergence of state symmetry breaking

In section 4.2.2 we have defined a notion of symmetry for the regularizers of the complete inverse problem, by saying that a regularizer is *symmetric* if it holds for any pair of states s, s' that $\bar{p}_s = \bar{p}_{s'} \Rightarrow p_s^* = p_{s'}^*$. We want to show through a very

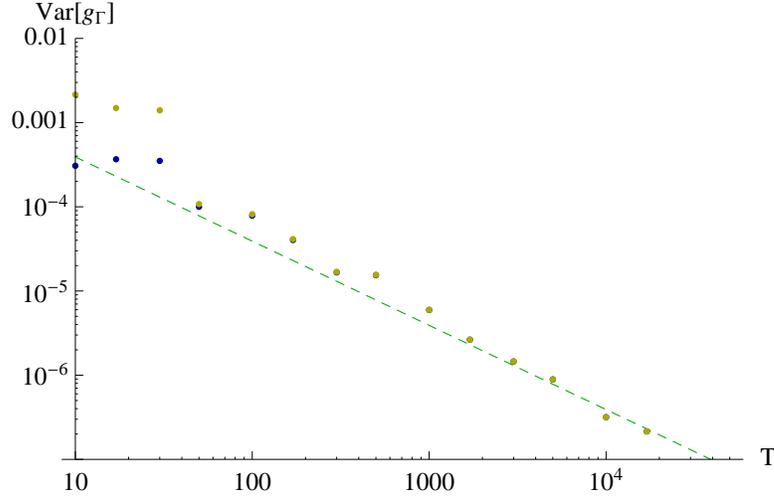


Figure 4.7: Variance of the regular (i.e., non-divergent) couplings as a function of the length of the number of samples T , for a model with $N = 8$ a set of $|\mathcal{I}| = 16$ observable states. Blue and yellow line respectively denote the non-regularized and the L-2 regularized value of the couplings (with $\beta = 5$). The green line shows the expected scaling of the variance for a flat distribution over the set of observable states.

simple example that the L-1 norm is non-symmetric and hence, according to the argument presented in section 4.2.2, it is not expected to have a trivial limit in the high-dimensional inference regime $T \sim N^\alpha \ll |\Omega|$. To show this, we consider a system of $N = 3$ spins, described by a complete model consisting of $|\phi| = 7$ operators and compare the inferred probability \mathbf{p}^* obtained by using an L-1 regularization with the one obtained by using an L-2 regularization. To do this, we numerically minimized (see appendix C for the details) the function

$$H(\mathbf{g}|\hat{\mathbf{s}}) = -T \left(F(\mathbf{g}) + \sum_{\Gamma \subseteq V, \Gamma \neq \emptyset} g_\Gamma \bar{\phi}_\Gamma \right) + H_0(\mathbf{g}) \quad (4.64)$$

with either $H_0(\mathbf{g}) = \beta \sum_{\Gamma \subseteq V, \Gamma \neq \emptyset} |g_\Gamma|$ or $H_0 = \frac{\beta}{2} \sum_{\Gamma \subseteq V, \Gamma \neq \emptyset} g_\Gamma^2$ for respectively the L-1 and the L-2 norm. We assumed the sampled configuration vector to be $\bar{\mathbf{p}} = \frac{1}{3}(\delta_{s,----} + \delta_{s,+--+} + \delta_{s,++++})$, in order to deal with only two different values for the empirical probability vector $\bar{\mathbf{p}}$. The results obtained in the case of the L-2 norm for the inferred probabilities are shown in figure 4.9, where it is possible to appreciate the uniform

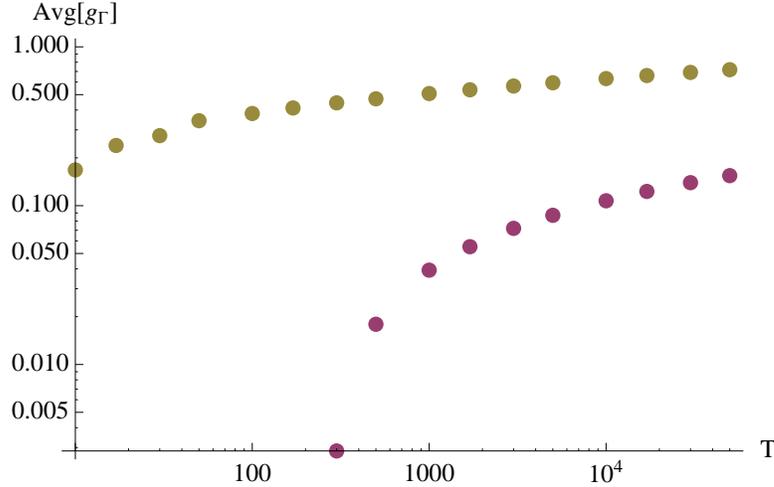


Figure 4.8: Divergence with T of the non-regular couplings, for the model described in previous plot. Red and yellow line respectively denote the values obtained putting $\log \epsilon = -\frac{1}{2} \log T$ and using an L-2 regularization ($\beta = 5$). Notice that the divergence is very slow, as it is expected to be logarithmic in T .

lifting of non-observed configurations, while probabilities associated with observed states are uniformly decreased as predicted by equation (4.30). In the case of the

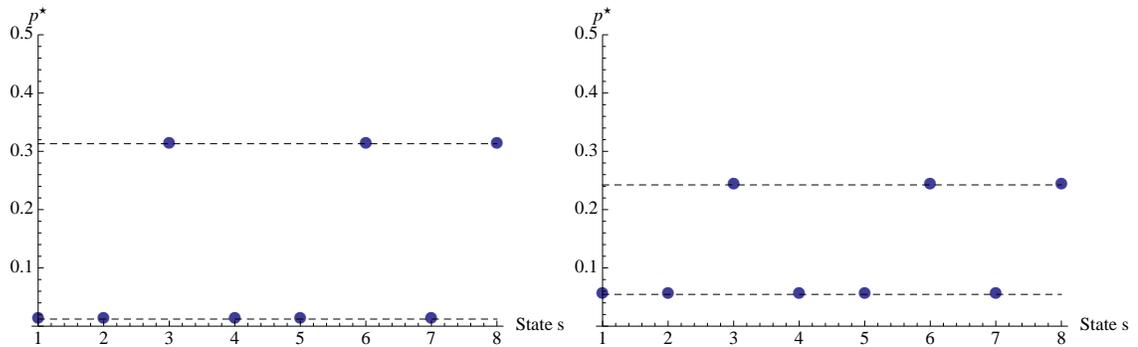


Figure 4.9: Inferred probability \mathbf{p}^* for the L-2 regularized complete inverse problem, in the case $\beta = 0.1$ (left panel) and $\beta = 0.8$ (right panel) in the highly under sampled limit $\bar{p}_s \in \{0, 1/T\}$. Equal empirical frequencies \bar{p}_s are mapped to equal inferred probabilities p_s^* .

L-1 norm (figure 4.10) we found that the vector of inferred probabilities can assign three different weights to the inferred state probability vector \mathbf{p}^* . In particular the configuration corresponding to the non-observed state $(-1, -1, -1)$ is lifted to a non-trivial value which breaks the state symmetry.

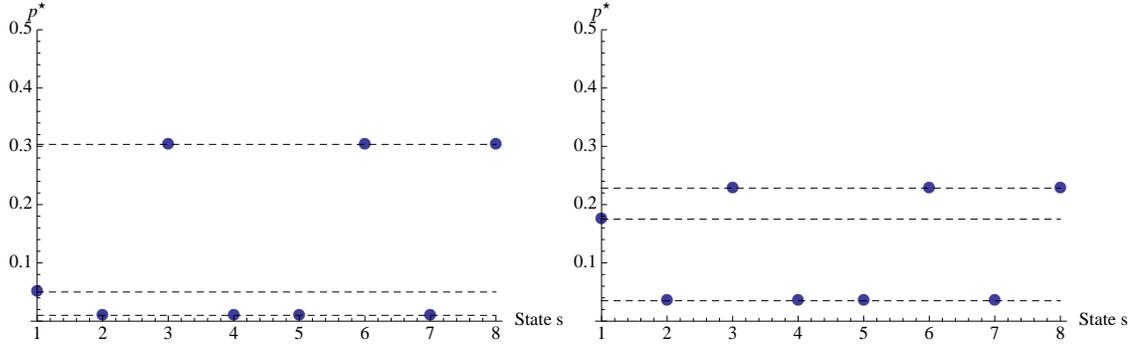


Figure 4.10: Inferred probability \mathbf{p}^* for the L-1 regularized complete inverse problem with $N = 3$, in the case $\beta = 0.1$ (left panel) and $\beta = 0.3$ (right panel) in the highly under sampled limit $\bar{p}_s \in \{0, 1/T\}$. The state symmetry which associates the same weight to configurations sampled the same number of times is spontaneously broken.

4.3.3 Pairwise model on a tree

We tested the results shown in section 4.2.3 providing a solution for the inverse problem for pairwise models with tree-like structure. We considered trees of size $N = 50$, and studied the behavior of the solution of the inverse problem for samples of length T up to 10^6 . The model which we considered was defined by the couplings J and h randomly and uniformly drawn in the interval $[0, 1]$. Datasets that we used did not consist of i.i.d. configurations sampled from the exact probability distribution, rather we sampled the states by using a Monte-Carlo simulation of T sweeps with a Metropolis-Hastings algorithm [50, 47]. We selected an initial condition of the form $\{1, \dots, 1\}$ in order to enforce a solution of positive \mathbf{m} in case of ergodicity breaking. Figure 4.11 shows the variance of the inferred couplings as a function of the length of the time series T , comparing it against a reference scaling $1/T$ for a random instance of a problem (i.e., a specific choice of \mathbf{h} and \mathbf{J}). We find that formula (4.54) correctly predicts the inferred couplings and their scaling to the actual ones. We remark that in this case, errors arise not only due to the finite number of samples, but is also introduced from an imperfect sampling of the empirical averages m and c . Indeed, as long as $\langle \phi \rangle - \bar{\phi} \sim T^{-1/2}$, the results obtained display the correct scaling of the variance. We also considered the case in which we produce a random instance of the

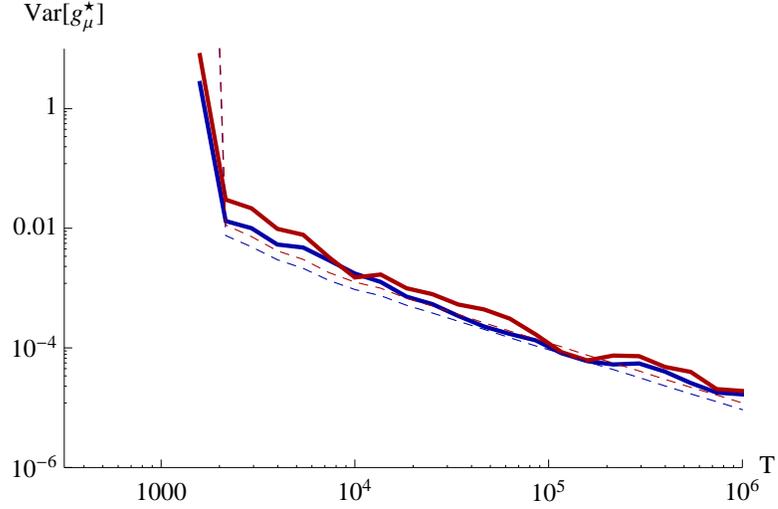


Figure 4.11: Variance of the inferred couplings h (red line) and J (blue line) against the number of samples T for a pairwise tree. The dashed lines plotted for reference indicates the error predicted by equation (4.55).

problem, and consider all the models obtained by multiplying the couplings with an *inverse temperature* β controlling the width of the fluctuations, in order to model the cases in which the noise is enhanced (β large) and the one in which it is suppressed ($\beta \rightarrow 0$). In particular, we considered a random instance of the model defined by couplings g randomly extracted in $[0, 1]$, and multiplied by a parameter $\beta \in [1/2N, 1]$, from which we extracted via MonteCarlo a set of $T = 10^5$ samples. In figure 4.12 we plotted the variance of the inferred coupling against the inverse temperature β . This plot shows that it is not possible to discriminate an overall strength of a couplings from a temperature parameter modulating the fluctuations. This implies that the maximum accuracy in inferring the products $\beta \mathbf{h}$ and $\beta \mathbf{J}$ is obtained when fluctuations are maximum ($h_i = J_{ij} = 0$), while the maximum accuracy for the inferred vector (h, J) is achieved by finding a compromise between maximum signal (favoring high couplings) and minimum noise (favoring high temperature, or equivalently low β). We also studied how the quality of the reconstruction of the couplings degrades by raising the β parameter. We find that within this inference scheme it is possible to reconstruct accurately the couplings as long as local fluctuations are sampled. More

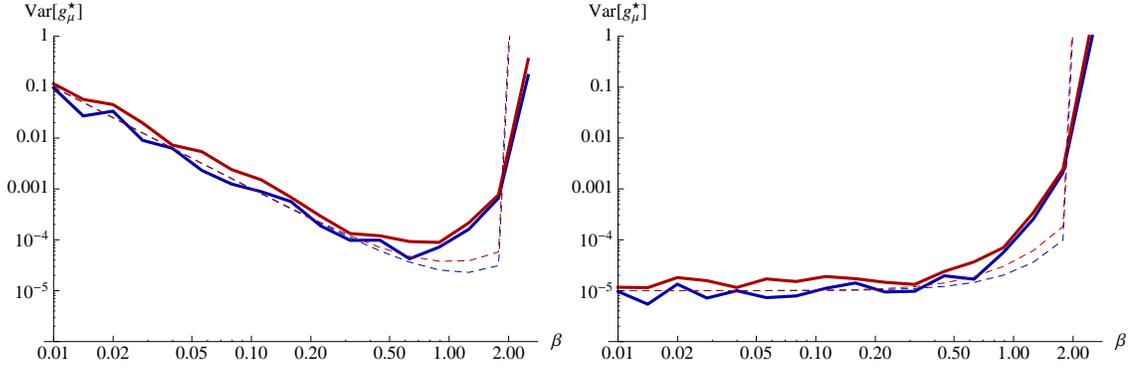


Figure 4.12: Variance of the inferred couplings \mathbf{h}^* (red line) and \mathbf{J}^* (blue) against the inverse temperature β for a pairwise tree, obtained by using $T = 10^5$ MonteCarlo samples. We plot both the variance of \mathbf{h}^* and \mathbf{J}^* (left panel) and the one of the products $\mathbf{h}^*\beta$ and $\mathbf{J}^*\beta$ (right panel), in order to show that this inference procedure cannot discriminate an overall interaction strength from an inverse temperature. The dashed lines indicate the value of the error estimated through equation (4.55).

precisely, expression (4.55) states that couplings can be accurately reconstructed as long as all the four possible states belonging to clusters of interacting spins (i, j) are well-sampled. This indicates that pushing β to large values, the configuration $(s_i, s_j) = (1, 1)$ gets more biased, eventually leading to the absence of other states if T is finite. Then, error can be large or divergent as shown in section 4.2.1 for the case of the complete inverse problem.

Remark 4.7. Notice that an accurate reconstruction of the couplings is obtained when local fluctuations (i.e., fluctuations relative to clusters of two spins) are sampled. It is not necessary to probe global fluctuations, which indicates that even in a phase in which ergodicity is broken, it is possible to accurately reconstruct the couplings, although no global fluctuations of the empirical average $m = \frac{1}{N} \sum_i m_i$ are observed. This indicates that it is not crossing the critical point what degrades the quality of the inference procedure, rather it is the lack of local fluctuations in the empirical samples.

4.3.4 One-dimensional periodic chain

We studied the performance of the inference procedure described in section 4.2.4 in inferring the couplings of a one-dimensional periodic chain with arbitrary range interactions. The analysis confirms the validity of the expression (4.62) for the couplings and (4.63) for the inverse susceptibilities. As an illustrative example, we consider the case of a periodic complete chain of size $N = 50$ with interactions of range $R = 4$ and periodicity parameter $\rho = 2$. We sampled via MonteCarlo a set of up to 10^6 configurations for a model in which the couplings g_Γ have been randomly and uniformly extracted from the interval $[0, 1/2N]$ (see above section for the details about the sampling procedure). The results for the variance of the inferred couplings (a set of $|\mathbf{g}^*| = 2^R(1 - 2^{-\rho})$ values) are represented in figure 4.13, where we study their dependence on the number of sampled unit cells NT/ρ . As we did above, we studied

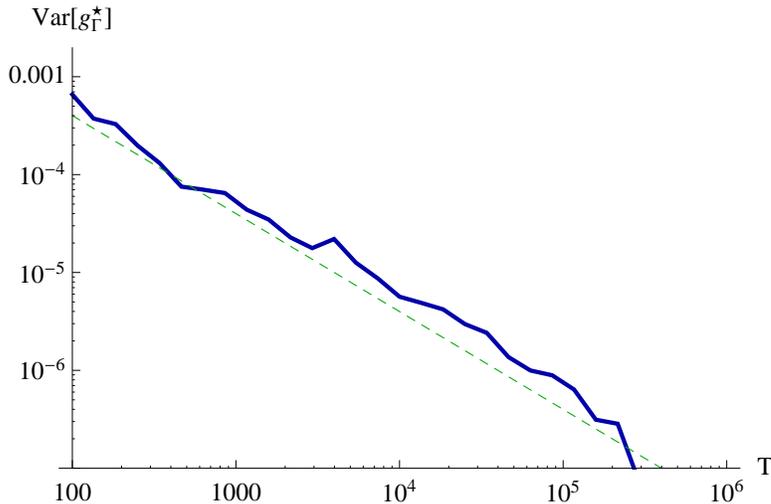


Figure 4.13: Variance of the inferred coupling vector \mathbf{g}^* (blue line) plotted against the number of sampled unit cells NT/ρ , obtained by MonteCarlo sampling of a model describing a complete one-dimensional periodic chain of size $N = 50$, range $R = 4$ and periodicity $\rho = 2$. The green dashed line shows the error predicted by equation (4.63).

the behavior of this inference procedure after modulating the interaction strength with an overall inverse temperature parameter β controlling the intensity of the fluc-

tuations for a random instance of the model . The results are shown in figure 4.14, where we show both the variance for the parameters \mathbf{g}^* and the one for the product $\beta\mathbf{g}^*$. Also in this case it is apparent that for a flat distribution ($\beta = 0$) the error

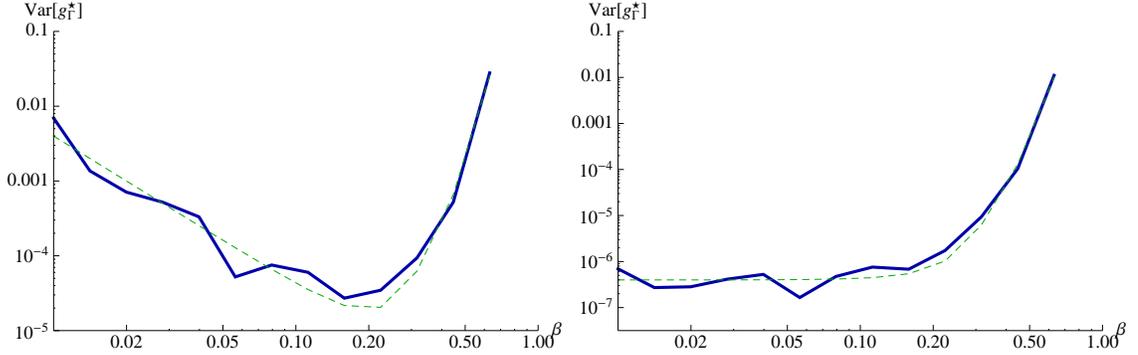


Figure 4.14: Variance of the inferred couplings \mathbf{g}^* (blue line) against the inverse temperature β for a one dimensional periodic chain. We have sampled 10^5 configurations via MonteCarlo to construct the empirical averages $\bar{\phi}$. The left panel shows the results for the inferred couplings \mathbf{g}^* , while the right one displays the results for the product $\beta\mathbf{g}^*$. The dashed lines indicate the estimation of the error obtained through equation (4.63).

on $\beta\mathbf{g}^*$ is minimum, while for the parameters \mathbf{g}^* the reconstruction error is minimal for a finite value of β which optimize the signal-to-noise ratio. We remark that also in this case the quality of the reconstruction of the couplings is determined by the sampling of the configurations belonging to clusters of R spins. If local fluctuations are not sampled well-enough, the error on the inferred couplings is large as predicted by equation (4.63). As observed above, it is not necessary to probe global fluctuations of the system in order to accurately reconstruct the couplings.

Chapter 5

Information geometry and criticality

In this chapter we will be interested in studying the natural structure of Riemannian manifold which characterizes the space of probability distributions [9]. This structure provides a mean to rigorously define a distance between statistical models, which can be used to characterize the consistency of the solution of the inverse problem through the notion of distinguishable distribution [61]. The metric structure of the coupling space becomes especially interesting in the case of models displaying a critical behavior at large N , as it allows for a characterization of (second-order) criticality from the point of view of information theory. In this scenario critical points can be seen as regions of the space of statistical models which are infinitely descriptive, in the sense that any finite region of the coupling space around a critical point can encode an anomalously high number of distinguishable statistical models. We call this phenomenon *model condensation*. An illustrative example is presented by discussing the thermodynamic limit of a fully connected ferromagnet. Finally, we will introduce a model of a stochastic point-process known as Hawkes process which we will use as a toy model to study the features of the inverse problem when applied to a realistic

dataset, and compare the results to the ones obtained by studying real data describing financial transactions in a stock market. This will allow to distinguish among spurious and the genuine collective features which emerge from the analysis of empirical data similar to the one considered in [72, 76, 24] in the context of neurobiology.

5.1 Metric structure of the probability space

5.1.1 Fisher information as a metric

Any statistical model (ϕ, \mathbf{g}) of the form (2.1) defines a probability density $p(s)$ on the configuration space Ω which is parametrically specified by a coupling vector \mathbf{g} . As such, one can see the space $\mathcal{M}(\phi)$ of all the probability densities obtained by varying the coupling vector $\mathbf{g} \in \mathbb{R}^M$ as an M -dimensional, smooth manifold, in which the role of the coordinates is played by the coupling vector \mathbf{g} . The advantage gained by taking this point of view is that the space $\mathcal{M}(\phi)$ is no longer associated with any particular parametrization of the probability space, rather it is characterized in term of the densities \mathbf{p} independently of their functional form. This is the point of view taken in the field of *information geometry*, in which the geometric properties of the space of probability distributions are inquired by using methods of differential geometry (see [9, 10] and [11] for a pedagogical review), which we will briefly present in the following sections. We will be interested in using these methods to answer several questions, namely: (i) is it possible to define a meaningful measure of distance in the space $\mathcal{M}(\phi)$? (ii) Is it possible to define a notion of *volume* in such $\mathcal{M}(\phi)$? (iii) Can a measure of *complexity* be defined? We will see that a positive answer to those points can be given by means of the Fisher information matrix.

Definition 5.1. Consider a minimal family ϕ and its corresponding manifold $\mathcal{M}(\phi)$. Then its tangent space $\mathcal{T}(\phi)$ is equipped by a canonical basis

$(\partial_1, \dots, \partial_M) = (\frac{\partial}{\partial g_1}, \dots, \frac{\partial}{\partial g_M})$, and given two tangent vectors¹ $X = \sum_{\mu=1}^M X_\mu \partial_\mu$ and $Y = \sum_{\mu=1}^M Y_\mu \partial_\mu$ and a point $\mathbf{p} \in \mathcal{T}(\phi)$ one can define the scalar product $\langle \cdot, \cdot \rangle_{\mathbf{p}} : \mathcal{T}(\phi) \times \mathcal{T}(\phi) \rightarrow \mathbb{R}$ as:

$$\langle X, Y \rangle_{\mathbf{p}} = \sum_{\mu, \nu} \chi_{\mu, \nu} X_\mu Y_\nu. \quad (5.1)$$

It can be shown (appendix A.2) that for any $X, Y \in \mathcal{T}(\phi)$ one has

$$\langle X, Y \rangle_{\mathbf{p}} > 0 \quad (5.2)$$

$$\langle X, Y \rangle_{\mathbf{p}} = \langle Y, X \rangle_{\mathbf{p}} \quad (5.3)$$

Hence, $\langle \cdot, \cdot \rangle_{\mathbf{p}}$ is a metrics which we define the *Fisher metrics* associated with $\mathcal{M}(\phi)$.

Notice that the scalar product $\langle X, Y \rangle_{\mathbf{p}}$ is independent of the parametrization used to describe the distribution \mathbf{p} due to the transformation law of $\chi_{\mu, \nu} = \langle \partial_\mu \log p(s) \partial_\nu \log p(s) \rangle$. This fact, and the choice of this metric itself, will be intuitively justified in the next section where the notion of distinguishable distribution will be introduced. The Fisher metrics allows to define the length of a curve in the space $\mathcal{M}(\phi)$.

Definition 5.2. Given a curve γ , i.e., a one-to-one function $\gamma : [a, b] \subset \mathbb{R} \rightarrow \mathcal{M}(\phi)$ with components $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)$, we define its *length* as

$$\ell(\gamma) = \int_a^b dt \sqrt{\sum_{\mu, \nu} \frac{d\gamma_\mu}{dt} \frac{d\gamma_\nu}{dt} \chi_{\mu, \nu}} \quad (5.4)$$

It is easy to show that the length of a curve (i) is independent of the parametrization of γ , (ii) is independent of the parametrization of $\mathcal{M}(\phi)$ (iii) is additive, i.e.,

¹ It is customary in literature to use superscripts for contravariant tensors and subscripts for covariant ones. We will disregard for simplicity this distinction and use lower indices for any tensor or vector field, as their use will be unambiguous.

given $a < b < c$, $\gamma_1 : [a, b] \rightarrow \mathcal{M}(\phi)$, $\gamma_2 : [b, c] \rightarrow \mathcal{M}(\phi)$ and $\gamma : [a, c] \rightarrow \mathcal{M}(\phi)$ such that $\gamma(t) = \gamma_1$ if $t < b$ and $= \gamma_2$ if $t \geq b$, one has $\ell(\gamma) = \ell(\gamma_1) + \ell(\gamma_2)$. Finally, a notion of distance $d(\cdot, \cdot) : \mathcal{M}(\phi) \times \mathcal{M}(\phi) \rightarrow \mathbb{R}$ between points in $\mathcal{M}(\phi)$ can be defined through

$$d(\mathbf{p}, \mathbf{q}) = \min_{\gamma \in \gamma(\mathbf{p}, \mathbf{q})} \ell(\gamma) \quad (5.5)$$

where $\gamma(\mathbf{p}, \mathbf{q})$ denotes the set of curves in $\mathcal{M}(\phi)$ starting in \mathbf{p} and ending in \mathbf{q} .

Definition 5.3. The curve

$$\gamma^* = \arg \min_{\gamma \in \gamma(\mathbf{p}, \mathbf{q})} \ell(\gamma) \quad (5.6)$$

is called a *geodesics*, and its coordinates $\gamma^* = (\gamma_1^*, \dots, \gamma_M^*)$ satisfy the linear differential equation

$$\frac{\partial^2 \gamma_\mu}{\partial t^2} + \sum_{\nu, \rho} \Gamma_{\nu, \rho}^\mu \frac{\partial \gamma_\nu}{\partial t} \frac{\partial \gamma_\rho}{\partial t} = 0, \quad (5.7)$$

where the *Christoffel symbols* $\Gamma_{\nu, \rho}^\mu$ are given by

$$\Gamma_{\nu, \rho}^\mu = \frac{1}{2} \chi_{\mu, \sigma}^{-1} \left(\frac{\partial \chi_{\sigma, \nu}}{\partial g_\rho} + \frac{\partial \chi_{\sigma, \rho}}{\partial g_\nu} - \frac{\partial \chi_{\nu, \rho}}{\partial g_\sigma} \right) \quad (5.8)$$

In appendix (E.1) we prove this well-known result by explicitly varying the length functional $\ell(\gamma)$.

Proposition 5.1. *The function $d(\cdot, \cdot) : \mathcal{M}(\phi) \times \mathcal{M}(\phi) \rightarrow \mathbb{R}$ satisfies for any $\mathbf{p}, \mathbf{p}', \mathbf{p}'' \in \mathcal{M}(\phi)$ the following relations: (i) $d(\mathbf{p}, \mathbf{p}') \geq 0$, (ii) $d(\mathbf{p}, \mathbf{p}') = 0$ if and only if $\mathbf{p} = \mathbf{p}'$, (iii) $d(\mathbf{p}, \mathbf{p}') = d(\mathbf{p}', \mathbf{p})$, (iv) $d(\mathbf{p}, \mathbf{p}') \leq d(\mathbf{p}, \mathbf{p}'') + d(\mathbf{p}'', \mathbf{p}')$. Hence, it is a proper measure of distance.*

We will show in the next section that this distance relates to the inverse problem by intuitively counting how many error bars away are two distributions away one from the other, given a fixed experiment length T . A related concept is the one of volume,

which can be used to quantify the number distributions that cannot be distinguished one from the other on the basis of an experiment of finite length.

Definition 5.4. Given a sub-manifold $\mathcal{M} \subseteq \mathcal{M}(\phi)$, we define the *volume* of \mathcal{M} as the value

$$\mathcal{N}(\mathcal{M}) = \int_{\mathcal{M}} d\mathbf{g} \sqrt{\det \hat{\chi}}, \quad (5.9)$$

which can trivially be shown to be invariant under reparametrization of \mathbf{p} .

Finally, we define along the lines of [61] the *complexity* of a manifold $\mathcal{M}(\phi)$ as the integral

$$\mathcal{N}(\mathcal{M}(\phi)) = \int_{\mathcal{M}(\phi)} d\mathbf{g} \sqrt{\det \hat{\chi}}. \quad (5.10)$$

The relevance of this measure will be elucidated in section 5.1.3.

5.1.2 Sanov theorem and distinguishable distributions

The metric introduced in section 5.1.1 can be justified by providing an intuitive interpretation in terms of distinguishable distribution, a concept which we will present starting from a simple consistency requirement. Suppose to be given a dataset $\hat{\mathbf{s}}$ of length T generated by an underlying (unknown) distribution. Then, given an operator set ϕ it is possible to construct the empirical averages $\bar{\phi}$ and to infer the maximum likelihood estimate of the couplings $\mathbf{g} = \mathbf{g}^*(\bar{\phi})$ describing the data, and to use them to generate a *different* dataset $\hat{\mathbf{s}}'$ of the same length as $\hat{\mathbf{s}}$. The maximum likelihood estimator $\mathbf{g}' = \mathbf{g}^*(\bar{\phi}')$ of $\hat{\mathbf{s}}'$ will, in general, be different from \mathbf{g} . Thus, distributions labeled by \mathbf{g} and \mathbf{g}' cannot be distinguished on the basis of a dataset of length T , as sketched in figure 5.1. What one expects is that by increasing T , the model \mathbf{g}' gets closer and closer to \mathbf{g} . This idea can be rigorously formulated by means of Sanov theorem (presented in section 2.2.4), which allows to prove the following corollary.

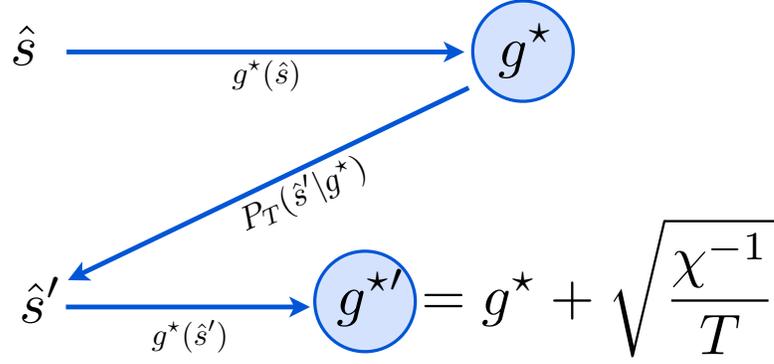


Figure 5.1: Cartoon illustrating the notion of indistinguishable distributions.

Corollary 5.1. *Consider a statistical model (ϕ, \mathbf{g}) associated with a probability density \mathbf{p} . Then, given a set of empirical averages $\bar{\phi}$ generated by \mathbf{p} and a maximal likelihood estimator \mathbf{g}^* , the probability that the maximum likelihood estimator \mathbf{g}^* takes a value close to \mathbf{g}' on the dataset associated with $\bar{\phi}$ is given by*

$$\lim_{\delta \rightarrow \mathbf{0}} \lim_{T \rightarrow \infty} -\frac{1}{T} \log \text{Prob}(\mathbf{g}^*(\bar{\phi}) - \mathbf{g}' \in \delta \mathbf{g}) = D_{KL}(\mathbf{p}' || \mathbf{p}), \quad (5.11)$$

where \mathbf{p}' is defined by the statistical model (ϕ, \mathbf{g}') and $\delta \mathbf{g} = [-\delta, \delta]^M$.

The proof of this corollary is presented in appendix (E.3). What it implies is that the Kullback-Leibler divergence controls the probability that after the resampling procedure explained above one ends in a model very different from the starting one. As expected, such probability is exponentially small in T . We will informally rewrite above corollary in the form

$$-\frac{1}{T} \log \text{Prob}(\mathbf{g}^*(\bar{\phi}) = \mathbf{g}') \xrightarrow{T \rightarrow \infty} D_{KL}(\mathbf{p}' || \mathbf{p}), \quad (5.12)$$

implying a choice of δ enforcing $\delta \mathbf{g}$ very close to $\mathbf{0}$. This will allow us to characterize the concept of indistinguishable distribution.

Definition 5.5. Consider two models \mathbf{g} and \mathbf{g}' within the same family of operators ϕ . Then, given a dataset of length T and empirical averages $\bar{\phi}$ sampled by the

model (ϕ, \mathbf{g}) and an *accuracy* $\epsilon > 0$, we say that \mathbf{g} and \mathbf{g}' are *indistinguishable* if the maximum likelihood estimator \mathbf{g}^* satisfies

$$-\log \text{Prob}[\mathbf{g}^*(\bar{\phi}) = \mathbf{g}'] \leq \epsilon \quad (5.13)$$

Given corollary (5.11), it is easy to prove (appendix E.3) that for large T the distinguishability of two distributions is determined by the generalized susceptibility, as stated in the next proposition.

Proposition 5.2. *Given two models (ϕ, \mathbf{g}) and (ϕ, \mathbf{g}') , in the limit of large T they are indistinguishable if*

$$\frac{1}{2} [(\mathbf{g}' - \mathbf{g})^T \hat{\chi} (\mathbf{g}' - \mathbf{g})] \leq \frac{\epsilon}{T}. \quad (5.14)$$

Remark 5.1. *Although the notion of indistinguishability inherits asymmetry in \mathbf{g} and \mathbf{g}' from the Kullback-Leibler divergence $D_{KL}(\mathbf{p}||\mathbf{p}')$, above proposition shows that for large T the definition symmetrizes.*

Remark 5.2. *This proposition clarifies the role of the Fisher metric (5.1): it shows that the distance among two close-by distributions is proportional to the log-probability that the maximum likelihood estimator of a statistical model (ϕ, \mathbf{g}) takes value $\mathbf{g}^* = \mathbf{g}'$. From this perspective, it is non-trivial to notice that this result is invariant after reparametrization of the probability densities.*

This last property identifies an approximatively elliptical region of indistinguishability in the space $\mathcal{M}(\phi)$ around each statistical model (ϕ, \mathbf{g}) , whose volume $\mathcal{V}_{T,\epsilon}(\mathbf{g})$ can be easily calculated in the large T limit, and is given by

$$\mathcal{V}_{T,\epsilon}(\mathbf{g}) = \frac{1}{\sqrt{\det \hat{\chi}}} \left[\frac{1}{\Gamma(\frac{M}{2} + 1)} \left(\frac{2\pi\epsilon}{T} \right)^{\frac{M}{2}} \right] \quad (5.15)$$

as shown in appendix E.4. Besides displaying the scaling of the volume with T expected by dimensional analysis, equation (5.15) shows that the Fisher information controls how wide is each region of indistinguishability inside the space $\mathcal{M}(\phi)$. In particular, the more the fluctuations are relevant in a given region $\mathcal{M} \subseteq \mathcal{M}(\phi)$, the better models in \mathcal{M} can be discriminated on the basis of a finite length experiment. Finally, the volume $\mathcal{V}_{T,\epsilon}(\mathbf{g})$ allows to define the concept of density of models, and to link it to the metrics described in section 5.1.1.

Definition 5.6. Consider the statistical model (ϕ, \mathbf{g}) and the space of models $\mathcal{M}(\phi)$. Then for any fixed T and $\epsilon > 0$ we define the *density of states* $\rho_{T,\epsilon}(\mathbf{g})$ as

$$\rho_{T,\epsilon}(\mathbf{g}) = \frac{1}{\mathcal{V}_{T,\epsilon}(\mathbf{g})} \propto \sqrt{\det \hat{\chi}} . \quad (5.16)$$

For large enough values of T , the density of models can be used to count the number of distinguishable models $\mathcal{N}_{T,\epsilon}(\mathcal{M}) = \int_{\mathcal{M}} dg \rho_{T,\epsilon}(\mathbf{g}) \propto \mathcal{N}(\mathcal{M})$ in a region of the space $\mathcal{M}(\phi)$. Then the Fisher metrics (5.1) has a natural interpretation through the notion of indistinguishable distributions, and the integration measure $\sqrt{\det \hat{\chi}}$ induced by the metric $\hat{\chi}$ is proportional to the density of models $\rho_{T,\epsilon}(\mathbf{g})$. The notion of distance defined in the previous section also has a simple interpretation in this setting. Consider in fact the discretization of the manifold $\mathcal{M}(\phi)$ induced by a sample size T and an accuracy ϵ , in which a curve $\gamma : [a, b] \in \mathbb{R} \rightarrow \mathcal{M}(\phi)$ is given. Suppose that one is interested in counting the number of ellipsoids (i.e., regions of indistinguishability) crossed by γ . Then one can see using equation (5.14) that the number of such regions $\ell_{T,\epsilon}(\gamma)$ tends in the large T limit to

$$\ell_{T,\epsilon}(\gamma) \left(\frac{2\epsilon}{T} \right)^{1/2} \xrightarrow{T \rightarrow \infty} \int_a^b dt \sqrt{\chi_{\mu,\nu} \dot{\gamma}_\mu(t) \dot{\gamma}_\nu(t)} = \ell(\gamma) . \quad (5.17)$$

A geodesic is interpreted in this setting as measuring the minimum number of models which have to be crossed to link two probability densities \mathbf{p} and \mathbf{q} with a curve γ , and

the corresponding distance $d(\mathbf{p}, \mathbf{q})$ is proportional to such number through the trivial pre factor $(T/2\epsilon)^{1/2}$. Summarizing, the link among the notions of length and volume defined in 5.1.1 and the corresponding notions in the field of statistical learning is provided by the relations

$$\ell_{T,\epsilon}(\gamma) = \ell(\gamma) \left(\frac{T}{2\epsilon}\right)^{1/2} \quad (5.18)$$

$$d_{T,\epsilon}(\mathbf{p}, \mathbf{q}) = d(\mathbf{p}, \mathbf{q}) \left(\frac{T}{2\epsilon}\right)^{1/2} \quad (5.19)$$

$$\mathcal{N}_{T,\epsilon}(\mathcal{M}) = \mathcal{N}(\mathcal{M}) \left[\Gamma\left(\frac{M}{2} + 1\right) \left(\frac{T}{2\pi\epsilon}\right)^{\frac{M}{2}} \right] \quad (5.20)$$

5.1.3 Complexity measures and criticality

One of the most relevant problems in the field of statistical learning is the one of choosing the most appropriate model in order to fit an empirical dataset $\hat{\mathbf{s}}$ generated by an unknown distribution. In particular it is well-known that models containing a large number of parameters typically lead to large values for the likelihood function $P_T(\hat{\mathbf{s}}|\mathbf{g})$, while parsimonious models tend to produce worst in-sample values. Conversely, parsimonious models tend to generalize better, while complex models tend to fit noisy components of data leading to a poor out-of-sample performance. Using a prior function $P_0(\boldsymbol{\phi}, \mathbf{g})$ which keeps into account the complexity of the model itself is a practical strategy which can be used to find an optimal compromise between faithfulness to the data and generalizability of the model. Popular priors used to achieve those goals are:

- **Akaike information criterion:** The Akaike information criterion (AIC) can be associated with the choice of a prior which penalizes the number of inferred parameters M through [7]

$$P_0(\boldsymbol{\phi}, \mathbf{g}) = e^{-M}, \quad (5.21)$$

which leads to the score

$$AIC = 2H(\boldsymbol{\phi}, \mathbf{g}|\hat{\mathbf{s}}) = 2M + 2H_0(\boldsymbol{\phi}, \mathbf{g}|\hat{\mathbf{s}}) . \quad (5.22)$$

- **Bayesian information criterion:** The Bayesian information criterion (BIC) considers a prior of the type [73]

$$P_0(\boldsymbol{\phi}, \mathbf{g}) = e^{-\frac{M}{2} \log T} , \quad (5.23)$$

leading to a score of the form

$$BIC = 2H(\boldsymbol{\phi}, \mathbf{g}|\hat{\mathbf{s}}) = M \log T + 2H_0(\boldsymbol{\phi}, \mathbf{g}|\hat{\mathbf{s}}) , \quad (5.24)$$

in which both the number of parameters and the sample size are taken into account. The BIC is closely related to the so-called Minimal Description Length criterion (MDL), in the sense that the score function $H(\mathbf{g}|\hat{\mathbf{s}})$ is proportional to the one obtained in [66, 65] by favoring models which lead to compressible data descriptions. In this sense, the notion of *simplicity* for a statistical model is related to the one compressibility and algorithmic complexity.

We will show in the following that the above results of information geometry allow to construct a measure of complexity which generalizes the BIC stated above, retaining the main feature of being completely invariant under reparametrization of the model [61, 13]. In order to do this, we consider the prior:

$$P_0(\boldsymbol{\phi}, \mathbf{g}) = \frac{\sqrt{\det \tilde{\boldsymbol{\chi}}}}{\mathcal{N}(\boldsymbol{\phi})} , \quad (5.25)$$

where the term $\mathcal{N}(\boldsymbol{\phi})$ is the volume of $\mathcal{M}(\boldsymbol{\phi})$ defined in (5.10).

Proposition 5.3. *Consider the probability for an unknown dataset $\hat{\mathbf{s}}$ of length T to belong to a given class of statistical models ϕ . Under the prior (5.25) this is given by*

$$P(\phi|\hat{\mathbf{s}}) \propto \int_{\mathcal{M}(\phi)} d\mathbf{g} P_T(\hat{\mathbf{s}}|\mathbf{g}) \left(\frac{\sqrt{\det \hat{\chi}}}{\mathcal{N}(\phi)} \right). \quad (5.26)$$

In the limit $T \rightarrow \infty$, this quantity concentrates according to:

$$P(\phi|\hat{\mathbf{s}}) \xrightarrow{T \rightarrow \infty} \left(\frac{P_T(\hat{\mathbf{s}}|\mathbf{g}^*)}{\mathcal{N}(\phi)} \right) \left(\frac{2\pi}{T} \right)^{M/2}, \quad (5.27)$$

where \mathbf{g}^ is the maximum likelihood estimator of \mathbf{g} .*

The proof of this result is completely analogous to the one shown in appendix A.6, and is obtained through a saddle-point expansion of the likelihood function $P_T(\hat{\mathbf{s}}|\mathbf{g})$. This result implies that the score assigned to the model ϕ converges to (up to an irrelevant constant in ϕ)

$$-\log P(\phi|\hat{\mathbf{s}}) \xrightarrow{T \rightarrow \infty} -\log P_T(\hat{\mathbf{s}}|\mathbf{g}^*) + \frac{M}{2} \log T + \log \mathcal{N}(\phi). \quad (5.28)$$

Remark 5.3. *The first two terms of the score (5.28) match the ones obtained by considering the BIC. The extra term $\log \mathcal{N}(\phi)$ quantifies a geometric contribution to the complexity of the model, which takes into account not only the number of parameters M , but also the detailed shape of the manifold $\mathcal{M}(\phi)$.*

Our interest lies in the fact that, assuming that on the basis of dimensional analysis the complexity measure $\log \mathcal{N}(\phi)$ scales like

$$\log \mathcal{N}(\phi) \sim M \log \ell, \quad (5.29)$$

where ℓ is a characteristic length scale, when high-dimensional models are considered, the scaling of the complexity might be *anomalous*, in the sense that ℓ can scale in the

limit $N \rightarrow \infty$ as a power of N . This argument additionally suggests that models ϕ containing critical points should be penalized by the prior (5.25), which assigns low scores to complex models. Intuitively, it has to become very costly to describe critical points even if the number of parameters of the model M is not large.

More specifically, if one assumes the scaling $\ell \sim N^\alpha$, then it is

$$H(\phi|\hat{\mathbf{s}}) = -\log P(\phi|\hat{\mathbf{s}}) \sim H_0(\phi, \mathbf{g}^*|\hat{\mathbf{s}}) + M \left(\frac{1}{2} \log T + \alpha \log N \right), \quad (5.30)$$

where one has the scaling $H_0(\phi, \mathbf{g}^*|\hat{\mathbf{s}}) \sim T$. Then one can intuitively expect that a fixed scaling of T, N and M is required in order for the inverse problem to be meaningful (i.e., the left term side of (5.30) to dominate the score). Hence, when dealing with high-dimensional inference, avoiding overfitting requires not only to study how M scales with N , but also to consider that the geometric properties of the model themselves can play a role through the logarithmic correction in the last term of (5.30).

5.1.4 Examples

The independent spin case

Consider the independent spin model described in section 2.2.5. By using equation (2.45) it is possible to find that

$$\det \hat{\chi} = \prod_{i \in V} \cosh^{-2} h_i. \quad (5.31)$$

Hence, the number of distinguishable independent spin models which can be described in an experiment of final length T with accuracy ϵ is

$$\mathcal{N}_{T,\epsilon} = \int d\mathbf{h} \rho_{T,\epsilon}(\mathbf{h}) = \left[\Gamma \left(\frac{N}{2} + 1 \right) \left(\frac{\pi T}{-2 \log \epsilon} \right)^{\frac{N}{2}} \right], \quad (5.32)$$

so that for example, just $\mathcal{N}_{T,\epsilon} \approx 5$ distinguishable models can be described by means of $T = 100$ observations of $N = 1$ spin with an accuracy of $e^{-\epsilon} = 1\%$, while for $T = 1000$ and $e^{-\epsilon} = 10\%$ one gets $\mathcal{N}_{T,\epsilon} \approx 23$. The finiteness of $\mathcal{N}_{T,\epsilon}$ also implies that infinite regions of the $\mathbf{h} \in \mathbb{R}^N$ space belong to the same distinguishable distribution. This can easily be checked, and one can see for example that for $N = 1$ the condition

$$1 = \int_{-\infty}^{h_{min}} dh \rho_{T,\epsilon}(h) \quad (5.33)$$

implies that for $T = 100$ and $e^{-\epsilon} = 1\%$ all models with h smaller than $h_{min} \approx -1.16$ (or h larger than $h_{max} = -h_{min}$) belong to the same region of indistinguishability.

Fully connected ferromagnet

Let's consider the fully connected ferromagnet described in section 3.3. In that case the calculation of $\det \hat{\chi}$ is non-trivial, and requires an analysis of the finite N corrections to the saddle-point solution of the model presented in appendix B.1, where it is shown that to leading order in N one has

$$\sqrt{\det \hat{\chi}} = \sqrt{\frac{N}{2}} \left(\chi_{s.p.}^{3/2} + \delta(h) \theta(J-1) \sqrt{2\pi^2 m_{s.p.}^2 \chi_{s.p.}} \right) \quad (5.34)$$

Also in this case it is possible to count the number of distinguishable models in a given region of space by explicitly integrating this measure. For example, we can calculate $\mathcal{N}_{T,\epsilon}$ in the semiplane $J \geq J_{max} \gg 1$ stripped of the $h = 0$ line. In that case it results that

$$\det \hat{\chi} \approx \sqrt{N} \left(4\sqrt{2} e^{-3(J+|h|)} \right), \quad (5.35)$$

which implies that in such region $\mathcal{N}_{T,\epsilon} \approx T\sqrt{N} \left(\frac{4\sqrt{2}}{-9\pi \log \epsilon} \right) e^{-3J_{max}}$. This indicates that no $J \gtrsim J_{max} \sim \frac{1}{3} \log T + \frac{1}{6} \log N$ can be discriminated by J_{max} unless $h \approx 0$. Interestingly, the number of models contained in the critical line $h \approx 1/N$ dominates $\mathcal{N}_{T,\epsilon}$ in the semiplane $J > J_{max}$. In fact the term of (5.34) proportional to $\delta(h)$

contributes with

$$\int_{-\infty}^{+\infty} dh \det \hat{\chi} \approx \sqrt{N} (2\pi e^{-J}) , \quad (5.36)$$

so that keeping into account the transition line one gets $\mathcal{N}_{T,\epsilon} \approx T\sqrt{N} \left(\frac{1}{-\log \epsilon}\right) e^{-J_{max}}$, and values of J which cannot be discriminated by J_{max} are the ones for which $J \gtrsim \log T + \frac{1}{2} \log N$. Finally, one can notice that $\chi_{s.p.}$ is divergent for $(h, J) = (0, 1)$. In particular the analysis of $\sqrt{\det \hat{\chi}}$ shows that along the line $J = 1$, the divergence is of the type $\sqrt{\det \hat{\chi}} \propto |h|^{-1}$, while for $h = 0$ and $J < 1$, one has $\sqrt{\det \hat{\chi}} \propto |1 - J|^{-3/2}$. Both divergencies are non-integrable, implying that the number of distinguishable models contained in a finite region around the point $(0, 1)$ dominates the total volume of the coupling space. This singularity is smeared out by finite-size effects when $N < \infty$, indeed those characteristics emerge by studying the scaling for finite N of the volume \mathcal{N} , as shown in figure 5.2. We plot in figure 5.3 the density of distinguishable

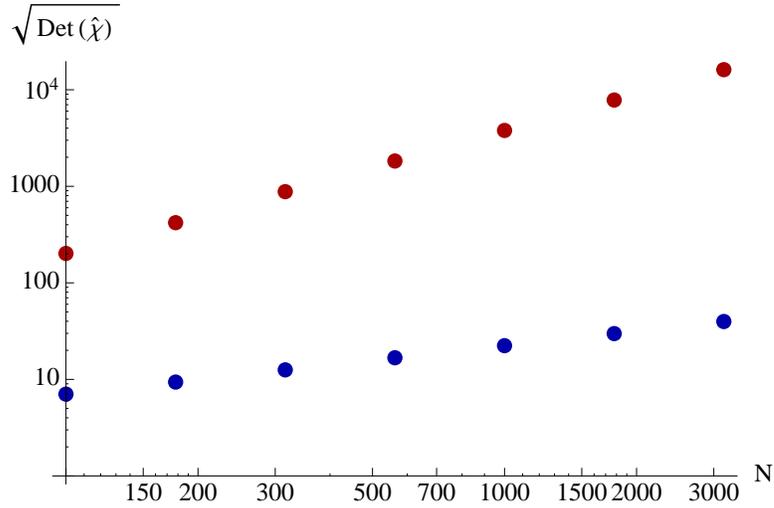


Figure 5.2: Finite size scaling of the measure $\sqrt{\det \hat{\chi}}$ for a fully connected ferromagnet computed via exact enumeration. The value obtained for the models $(h, J) = (0, 1)$ (red points) and $(h, J) = (0, 0)$ (blue points) are plotted.

models for this model in the case $N = 100$, computed both by exact enumeration and via saddle point approximation. The geodesics for this model can also be numerically computed by solving the differential equation (5.7) explicitly. As an example, we

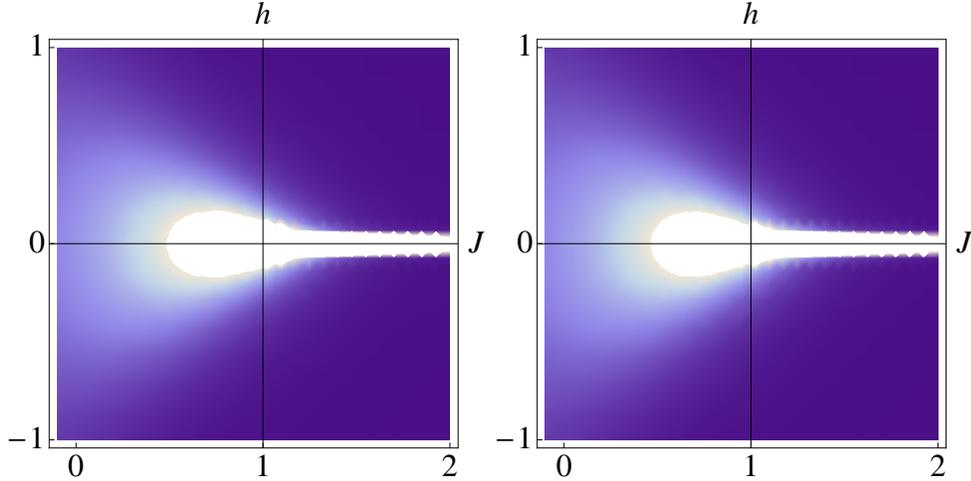


Figure 5.3: Density of models $\rho(h, J) \propto \det \hat{\chi}$ for the fully connected ferromagnet. The left panel shows the exact value calculated for $N = 100$, while the right panel displays the saddle-point approximation described in appendix B.1.

plot in figure 5.4 a set of geodesics of length $\ell(\gamma) = 1$ calculated for a system of size $N = 50$.

5.2 Inference of a non-equilibrium model

Many recent works in the field of neurobiology focus on neuronal ensembles which are described by means of strings of binary variables encoding the activity pattern of a set of $N \sim 10^1$ or $N \sim 10^2$ neurons [72, 76, 24]. Such compact description of the fundamental units of those system has been argued to be meaningful, triggering the expectation that techniques such as the ones described in chapter 2 might be applied on empirical data in order to extract relevant information about the interaction patterns of networks of real neurons. As a result of those expectations, striking features of neural ensembles started to emerge from the solution of the inverse problems applied to experimental data [82, 58, 78]. These findings posed a challenging question, whose answer has yet to be fully clarified in order to assess their validity, namely: *how much of those emerging features depends on the inference procedure which has been applied, and how much is intrinsically associated with structural properties of the*

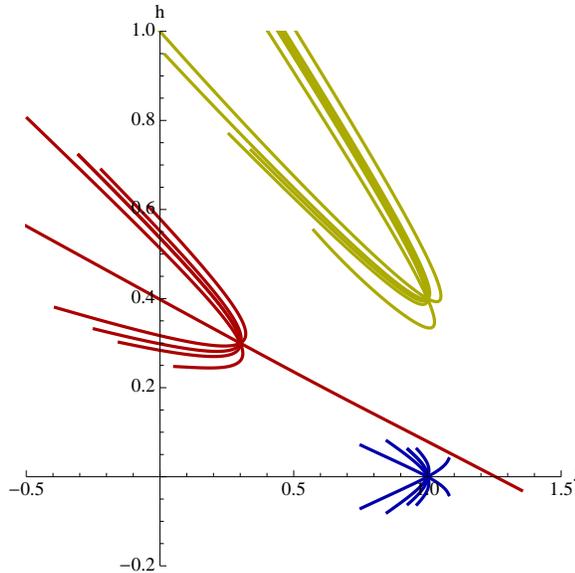


Figure 5.4: Three sets of geodesics of length $\ell(\gamma) = 1$ plotted for a system of size $N = 50$. Blue (respectively, red, yellow) lines describe curves passing through the points $(h, J) = (0, 1), (0.3, 0.3), (0.4, 1)$. It is possible to notice how volume shrinks around the critical point $(0, 1)$ and the presence of a quasi-null mode of $\hat{\chi}$ along the direction $(-m, 1)$.

system? The implications of the answer go well beyond the field of neurobiology, and apply more generally to the field of statistical learning. In this section we want to provide a partial answer to this point, and show that procedures similar to the ones used to study such neural networks may generate spurious features in the inferred models, as well as genuine ones. We address in particular the issue of *criticality*, which we identify from the point of view of statistical mechanics with the presence of long-range correlations in a system as a result of strong collective interactions among its constituents. We apply those ideas to two datasets whose nature is similar to the one considered in [72, 76, 24]: a set of simulated realizations of a Hawkes point-process [38, 37] and a dataset describing transactions in a financial market.

5.2.1 The Hawkes process

We will introduce the Hawkes point-process as a null-model to describe a system consisting of N interacting units which are able produce *events* in time and cross-influence each other in *absence* of remarkable collective behaviors (i.e., the emergence of long-range correlations in time or space). The study of the discretized version of this model will allow an analysis of the genuine and the spurious features of the inferred model under the procedure described in chapter 2.

Definition and basic properties

We will briefly remind the notion of point process, which we will use to construct the Hawkes process, addressing the reader to [14, 19] for a more detailed description.

Definition 5.7. We consider an N -variate point-process described by a non-decreasing, right-continuous *counting function* $\mathbf{X} = (X_1, \dots, X_N) : [0, \infty) \subset \mathbb{R} \rightarrow \mathbb{N}^N$, such that

$$\frac{\text{Prob}(dX_i(\tau) = 1 | \mathbf{X}(\tau')(\tau' < \tau))}{d\tau} \xrightarrow{d\tau \rightarrow 0} \lambda_i(\tau | \mathbf{X}(\tau')(\tau' < \tau)) \quad (5.37)$$

$$\frac{\text{Prob}(dX_i(\tau) > 1 | \mathbf{X}(\tau')(\tau' < \tau))}{d\tau} \xrightarrow{d\tau \rightarrow 0} 0, \quad (5.38)$$

where $dX_i(\tau) = X_i(\tau + d\tau) - X_i(\tau)$, and the (possibly stochastic) value $\boldsymbol{\lambda}(\tau) = (\lambda_1(\tau), \dots, \lambda_N(\tau))$ is referred as the *conditional intensity* (or more simply, *intensity*) function.

Intuitively, $X_i(\tau)$ counts the number of *events* of type i falling in the interval $[0, \tau]$, being the probability of the occurrence of an event in $(\tau, \tau + d\tau]$ equal to $\lambda_i(\tau)d\tau$, and being the one associated with the outcome of two events of order smaller than $d\tau$. A well-known example is provided by the Poisson process, which is a point-processes specified by a constant, deterministic value for the intensity $\lambda_i(\tau) = \mu_i$.

Finally, we will say that a point-process is (asymptotically, weak-sense) *stationary* if the mean $\mathbb{E}[d\mathbf{X}(\tau)] \xrightarrow[\tau \rightarrow 0]{\tau \rightarrow \infty} \boldsymbol{\lambda}(\tau) d\tau$ is independent of τ and the covariance $\text{Cov}(dX_i(\tau), dX_j(\tau')) \xrightarrow[\tau \rightarrow 0]{\tau \rightarrow \infty} \sigma_{ij}(\tau, \tau') d\tau^2$ depends just upon the difference $\tau - \tau'$.

Definition 5.8. We will call a *Hawkes point-process* the stationary, N -variate point-process $\mathbf{X}(\tau) = (X_1(\tau), \dots, X_N(\tau))$ defined by a stochastic intensity vector $\boldsymbol{\lambda}(\tau) = (\lambda_1(\tau), \dots, \lambda_N(\tau))$ of the form

$$\lambda_i(\tau) = \mu_i + \sum_{j=1}^N \int_{-\infty}^{\tau} dX_j(\tau') K_{ij}(\tau - \tau'), \quad (5.39)$$

such that $\hat{\mathbf{K}}(\tau)$ is a positive matrix kernel satisfying

$$K_{ij}(\tau) \geq 0 \quad \text{if } \tau \geq 0 \quad (5.40)$$

$$K_{ij}(\tau) = 0 \quad \text{if } \tau < 0 \quad (5.41)$$

$$\max_n |K_n| < 1, \quad (5.42)$$

where $\{K_n\}_{n=1}^N$ are the eigenvalues of the Fourier transform $\hat{\mathbf{K}}(\omega) = \int d\omega e^{i\omega\tau} \hat{\mathbf{K}}(\tau)$ calculated in the point $\omega = 0$, so that condition (5.42) ensures the stationarity of the process (5.39).

This model describes a self-excitatory process (i.e., $\text{Cov}(dX_i(\tau), dX_j(\tau')) \geq 0$) due to the positive, linear coupling of the stochastic intensities $\lambda_i(\tau)$ with the process itself. The interest in this model resides in the fact that it can describe *clustering* of events: just as non-interacting (i.e., Poisson) point-processes describe events which occur at times uniformly drawn from the time axis, Hawkes point-processes model events which tend to take place in close-by regions in time due to an attractive interaction modeled by the kernel $\hat{\mathbf{K}}(\tau)$.

We focus on the properties of this model in the stationary regime, which is guaranteed to exist for the choice of the spectral radius of the kernel $\hat{\mathbf{K}}(\tau)$ that we specified

through (5.42). Despite the fact that both averages and two point correlations of $\mathbf{X}(\tau)$ can be analytically computed for a large class of functions $\hat{\mathbf{K}}(\tau)$ [38, 37], in the following discussion we will just require the knowledge of the average intensity $\boldsymbol{\lambda} = \mathbb{E}[\boldsymbol{\lambda}(\tau)]$.

Proposition 5.4. *Given a stationary Hawkes point-process, the average intensity vector $\boldsymbol{\lambda}$ is given by*

$$\boldsymbol{\lambda} = \left(\hat{\boldsymbol{\delta}} - \hat{\mathbf{K}}(\omega = 0) \right)^{-1} \boldsymbol{\mu}, \quad (5.43)$$

as one can easily see by taking the expectation value of equation (5.39) and imposing the stationarity condition $\boldsymbol{\lambda}(\tau) = \boldsymbol{\lambda}$.

We employ the notation $\mathbb{E}[\dots]$ to indicate an average taken in the stationary state of the model, and $\hat{\boldsymbol{\delta}}$ denotes the identity matrix in dimension N .

We want to highlight some of the features of the Hawkes process which differentiate it from statistical models such as the ones described in section 2.1.

- **Dynamics:** The Hawkes process describes a stochastic process characterized by the dynamics (5.39), while a statistical model $(\boldsymbol{\phi}, \boldsymbol{g})$ of the form (2.1) describes a stationary probability density. This implies that any information concerning the directionality in time (e.g., causality) of the interactions is lost when passing to a description in terms of i.i.d. binary strings.²
- **Non-stationarity:** For any non-stationary generalization of the Hawkes process in which the kernel changes in time (i.e., it is of the form $\hat{\mathbf{K}}(\tau, \tau')$), or the exogenous intensity is a function $\boldsymbol{\mu}(\tau)$, it is likely that inferring a stationary model may lead to errors in the interpretation of the results. In particular, what is described as an interacting, stationary system in the language of the inferred

²This can be understood by noting that any dataset $\pi[\hat{\mathbf{s}}] = \{s^{(\pi_t)}\}_{t=1}^T$ obtained by applying any permutation π_t to a raw dataset $\hat{\mathbf{s}} = \{s^{(t)}\}_{t=1}^T$ leads to the same inverse problem.

model $p(s) \propto \exp\left(J \sum_{i<j} s_i s_j + h \sum_i s_i\right)$ may correspond to a non-interacting, non-stationary real system [84].

- **Criticality:** The divergence of the mean intensity $\mathbb{E}[\boldsymbol{\lambda}(\tau)]$ doesn't indicate criticality of the statistical model describing the stationary state of the Hawkes process. In particular, the divergence of $\boldsymbol{\lambda}(\tau)$ is not linked to collective effects, as it is present even for finite N , while a proper phase transition in the statistical mechanics sense can arise just in the large N limit.

These considerations also apply when considering the binary encoding of the stochastic process describing spiking neurons, or more generally when considering any point-process which is binned and discretized in order to perform an inference procedure such as the one described in chapter 2.

The fully-connected Hawkes process

We introduce here the notion of fully-connected Hawkes process, which we will relate to the fully-connected pairwise model in the following part of the discussion.

Definition 5.9. Consider an N -dimensional Hawkes point-process, whose intensity vector $\boldsymbol{\lambda}(\tau)$ is defined by

$$\lambda_i(\tau) = \mu_i + \sum_j \int_{-\infty}^{\tau} dX_j(\tau') \alpha_{ij} e^{-\beta(\tau-\tau')}, \quad (5.44)$$

which corresponds to the choice of an exponentially decaying influence kernel $K_{ij}(\tau - \tau') = \alpha_{ij} e^{-\beta(\tau-\tau')} \theta(\tau - \tau')$. Let $\hat{\boldsymbol{\alpha}}$ be a matrix of the form

$$\alpha_{ij} = \frac{\alpha}{N-1} (1 - \delta_{ij}) \quad (5.45)$$

and the vector $\boldsymbol{\mu}$ to be equal to $\mu_i = \mu$ for each i . Then such process will be called a *fully-connected Hawkes process*.

For a fully-connected Hawkes process, it is easy to see by employing formula (5.43) that

$$\mathbb{E}[\lambda_i(\tau)] = \mu \left(1 - \frac{\alpha}{\beta}\right)^{-1}, \quad (5.46)$$

while the stationarity condition (5.42) reduces to $\alpha < \beta$.

Binning and discretization

In order to establish a connection between a spin system and a Hawkes process, we consider a discretization in time and a binarization of the signal dealt according to the following procedure.

Definition 5.10. Given a realization of an N -dimensional Hawkes process described by a counting function $\mathbf{X}(\tau)$ with $\tau \in [0, \tau_{max}]$ and a *bin size* $\delta\tau$, we define for any $i \in \{1, \dots, N\}$ and $t \in \{1, \dots, \tau_{max}/\delta\tau = T\}$ the *binning functions*

$$b_i^{(t)}(\mathbf{X}, \delta\tau) = \min\{1, X_i(t\delta\tau) - X_i(t\delta\tau - \delta\tau)\} \quad (5.47)$$

which is 1 if an event of type i occurred in the interval $\tau \in \delta\tau [t - 1, t]$ and zero otherwise. We analogously define the functions

$$s_i^{(t)}(\mathbf{X}, \delta\tau) = 2b_i^{(t)}(\mathbf{X}, \delta\tau) - 1, \quad (5.48)$$

which evaluate to 1 if an event of type i occurred in the interval $\tau \in \delta\tau [t - 1, t]$ and to -1 otherwise.

In order to shorten the notation, we will often write $b_i^{(t)} = b_i^{(t)}(\mathbf{X}, \delta\tau)$ and $s_i^{(t)}(\mathbf{X}, \delta\tau) = s_i^{(t)}$. Those functions provide a mean to map an Hawkes process to an empirical dataset $\hat{\mathbf{s}}$ through $(\mathbf{X}, \delta\tau) \rightarrow \hat{\mathbf{s}} = \{s_i^{(t)}(\mathbf{X}, \delta\tau)\}_{t=1}^T$. Notice that an empirical dataset $\hat{\mathbf{s}}$ constructed according to this procedure does not consist in general of i.i.d. observations.

5.2.2 Trades in a financial market

Financial markets are complex systems in which a large number of individuals interact by buying and selling contracts at variable prices according to an unknown, dynamically varying set of criteria (e.g., their specific needs, their past experience, their future expectations). In this sense, markets can be seen as intermediary entities implicitly defined by a set of trading rules which mediate the interactions of individuals. Those rules should be such that efficient allocation of resources is achieved, so that price of traded goods reflects correct information about their fundamental value [33]. Evidence that this is not always the case has dramatically emerged in recent times [16, 46, 83]. Part of the responsibility has been attributed to the instability of the microscopic mechanism by means of which financial markets process information, producing prices and providing liquidity for investors [17]. Hence, it makes sense to characterize empirically how such mechanism operates, and to identify its weaknesses, its sources of inefficiencies and potential causes of its instability. With this ideas in mind, we want to characterize from the empirical point of view a part of the complex process leading to price formation.

Types of market data

In modern financial markets the action of the participants is constantly recorded, and most of the events taking place during its activity are electronically stored. In some cases, part of this data is available for investigation. In particular some main categories of datasets describing market activity which can be identified and classified on the basis of the timescale they are associated with. The most detailed level of description (timescales ranging from tens of milliseconds to the second) is achieved when informations about single market events triggered by individual agents are available [48, 59, 29]. A more coarse-grained description of the market is obtained by focusing on the price process and its variations. More precisely, it is possible to define an

instantaneous price for any contract, and to keep track of all its variations (tick-by-tick data) throughout the duration of the market activity. Data describing all events changing the price (called either *trade* or *quote* events) are necessary to achieve this level of description (being the typical time resolution required the one of the second) [28]. Finally, data corresponding to market behavior at lower frequencies are often publicly available, and involve, beyond the daily opening and closing price, the volume traded and the highest and lowest daily price for all traded goods (e.g., they can be found in [5]). In this discussion, we focus on data describing *trade* events, which belong to the intermediate regime in which the price process is monitored with the resolution of around one second. Any of those trade events corresponds to the transfer of a contract from a seller to a buyer at a given price, for a given quantity (volume) of a good.

Cross-correlation of trade events

It has been observed in empirical data across several markets that trade events of single securities are not independent one from another, rather they influence each other leading to interesting clustering phenomena. Moreover, by considering multiple securities traded in the same market venue, it is possible to check that even event times associated with the to trade of different instruments are strongly correlated among each other. Then, one can be interested in answering the following question: *do correlations in trading times arise from correlated exogenous phenomena driving market activity, or do they form due to an endogenous contagion process spreading across the market?* While the former scenario would correspond to a picture in which market activity reflects fundamental exogenous information, the latter would be associated with the scenario of a (potentially unstable) market which self-interacts without necessarily assimilating external information. Those scenarios can in principle coexist, although it is not easy to construct a quantitative, empirically measurable notion dis-

tinguishing the two regimes [44]. It should also be added that part of the explanation for long-range correlation in trading times has been identified in the mechanism of *order-splitting*: the finite amount of liquidity available in the market forces traders to split large orders (*meta-orders*, *care-orders* or *hidden-orders*) in smaller lots which are traded incrementally, leading to long-range correlation of trading times (from hours to days, sometimes even up to weeks). Indeed a relevant role could also be played by collective interactions across different securities, which could lead to correlated order flow. This possibility is empirically inquired in section (5.3.2), where we apply the techniques described in chapter 2 to this type of financial system, and try to understand the results on the basis of what we presented in the early part of this chapter.

5.3 Applications

With these ideas in mind, we consider two sets of realizations of a point-processes $\mathbf{X}(\tau)$.

- **Hawkes processes:** We considered simulated data corresponding to several realizations of a multivariate ($N = 100$) fully-connected Hawkes processes with parameters in a variable range.
- **Financial data:** We studied trade events corresponding to one year of activity (2003) in a specific stock market, the New York Stock Exchange (NYSE), for the $N = 100$ most traded assets.

The counting functions $\mathbf{X}(\tau)$ have been discretized in both cases by using a sliding window of size δt in order to build the datasets $\hat{\mathbf{s}}(\mathbf{X}, \delta t)$ by using the binning function (5.48). Datasets $\hat{\mathbf{s}}$ have been used to construct the empirical magnetizations $\mathbf{m} = (m_i)_{i \in V}$ and the correlation matrix $\hat{\mathbf{c}} = \{c_{ij}\}_{i < j \in V}$, together with the average

magnetization $m = \sum_{i=1}^N m_i$ and the average correlation $c = \frac{2}{N(N-1)} \sum_{i<j \in V} c_{ij}$. Then we solved the inverse problem for this sets of data by considering two type of models:

- **Fully-connected ferromagnet:** We considered the operator set $\phi = \{\sum_i s_i, \frac{1}{N} \sum_{i<j} s_i s_j\}$ defining the model (3.23) and extracted the conjugated parameters $\mathbf{g}^* = (h^*, J^*)$ given the empirical averages $\bar{\phi} = (Nm, \frac{N-1}{2}c)$ as shown in section 3.3.
- **Disordered fully-connected ferromagnet:** We considered the operator set $\phi = \{s_i\}_{i \in V} \cup \{\frac{1}{N} s_i s_j\}_{i<j \in V}$ defining the model (2.47) and extracted the conjugated parameters $\mathbf{g}^* = \{\mathbf{h}^*, \hat{\mathbf{J}}^*\}$ given the empirical averages $\bar{\phi} = (\mathbf{m}, \hat{\mathbf{c}})$ by using the algorithms described in section 3.1.2

5.3.1 Pairwise fully-connected model for Hawkes processes

In the case of the fully-connected Hawkes process, we considered N -variate models with $N = 100$ for various set of parameters (μ, α, β) . We fixed without loss of generality $\mu = 0.011 \text{ s}^{-1}$ (as a common factor in the choice of the parameters can be reabsorbed into a suitable definition of the time coordinate τ) and simulated datasets consisting of 5×10^3 events with α in the range $[0, \beta]$. We first studied the behavior of the average magnetization and correlations, finding the results summarized in figure 5.5 for the generic case $\mu = 0.011 \text{ s}^{-1}, \alpha = 0.015 \text{ s}^{-1}, \beta = 0.03 \text{ s}^{-1}$ and described in the following.

Relations among bin size and empirical observables

- The average magnetization ranges from -1 to 1 depending on $\delta\tau$, being the crossover determined from the value of $\mathbb{E}[\lambda(\tau)]$. We plot for reference the curve $1 - 2e^{-\delta\tau\mu/(1-\alpha/\beta)}$ corresponding to the average value of the magnetization in the stationary state.

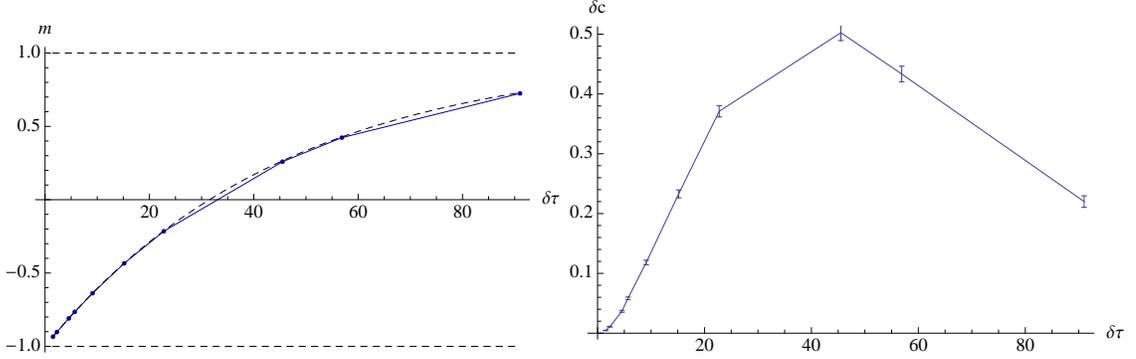


Figure 5.5: Average magnetization (left panel) and average correlation (right panel) as functions of the bin size $\delta\tau$ (in units of seconds) for simulated data corresponding to a fully-connected Hawkes process defined by parameters $\mu = 0.011 \text{ s}^{-1}$, $\alpha = 0.015 \text{ s}^{-1}$, $\beta = 0.03 \text{ s}^{-1}$. δc indicates the normalized connected correlation $\delta c = N(c - m^2)$ defined in section 3.3. The dashed line in the left panel indicates the reference value $m = 1 - 2e^{-\mu\delta t / (1-\alpha/\beta)}$.

- Correlations drop to zero as the window $\delta\tau$ is made smaller (a phenomenon known in the field of finance as Epps effect [32]). In particular if the $\delta\tau$ is smaller than the natural scale for the dynamics of the system β^{-1} , one expects correlations not to be fully developed. Conversely, when the bin size includes on average multiple events ($\delta\tau \sim \mathbb{E}[\lambda(\tau)]^{-1}$) correlations start to drop due to the binarization of the data.

This leads to a general consideration involving the optimal bin size required to perform inference: while a large $\delta\tau$ implies less statistics (due to $T = \tau_{\max}/\delta\tau$) and leads to multiple events thus decreasing correlations, it generates less correlated samples (as the auto-correlation decays exponentially in $\beta\delta\tau$). Conversely, small values of $\delta\tau$ imply more statistics, at the price of decreasing the independence of the samples. Eventually, for $\delta\tau$ very small no dynamics is observed due to Epps effect. All those features can be qualitatively motivated in a simple approximation which allows to compute the averages $\mathbb{E}[s_i^{(t)}]$ and $\mathbb{E}[s_i^{(t)}s_j^{(t)}]$ (appendix E.5).

Features of the inferred models

Extracting the couplings of a fully connected model from the values of magnetization and correlation described above leads to the results depicted in figure 5.6, where we consider both the disordered case $\mathbf{g} = (h, \hat{J})$ and the two-parameter model $\mathbf{g} = (h, J)$. We stress in the following some of the main features.

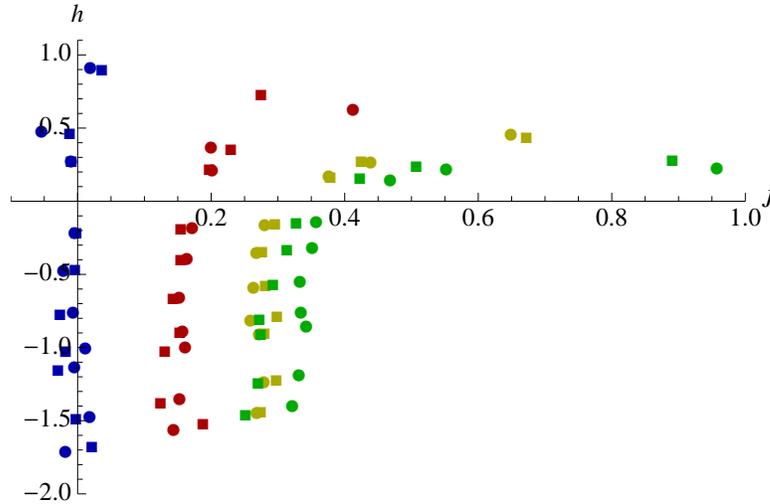


Figure 5.6: Inferred couplings obtained for several choices of Hawkes processes, for various choices of the bin size $\delta\tau$. We considered models with $\mu = 0.01 \text{ s}^{-1}$, $\alpha = 0, 0.0075, 0.015, 0.0225 \text{ s}^{-1}$, $\beta = 0.03 \text{ s}^{-1}$ (respectively, blue, red, yellow, green line), and bin sizes ranging from 20 to 80 s. Circles correspond to average couplings inferred from a heterogeneous model, while squares indicate couplings obtained by fitting a homogeneous model.

- The Poisson point-process is mapped on the line $J = 0$, while models with increasing interaction parameter α for a fixed β are mapped on monotonically increasing values of J . In this sense, interactions in the original model are genuinely mapped in couplings J within the inferred model. Moreover, increasing interaction parameters lead to curves which are closer to the critical point.
- The inferred fields do not increase monotonically in $\delta\tau$, and the asymptotic behavior when $\delta\tau \rightarrow +\infty$ may be either $h \rightarrow +\infty$ (for $\alpha > \beta/2$) or $h \rightarrow -\infty$ ($\alpha < \beta/2$). This indicates that the inference procedure that we use can generate

metastable states (see section 3.4) as legitimate solutions of the inverse problem. The metastability can be understood as a spurious result of the inference procedure, as it doesn't correspond to any instability of the underlying Hawkes point-process.

- Adopting a criterium of maximum information efficiency in order to select $\delta\tau$ would lead to a choice of an inferred model which is maximally close to the critical point, where the stability of the model is infinite (section 5.1.4). Equivalently, adding to the criteria required to choose $\delta\tau$ listed above also the stability would poise the inferred model artificially close to the critical point, where statistical models generalize better.
- Interestingly, the inferred model doesn't lie on the line $h = 0$ where most models concentrate (section 3.3). This is because the scaling $\sim 1/N$ of the kernel $\hat{\mathbf{K}}(\tau)$ leads to correlations proportional to $1/N$ (see appendix E.5 for a qualitative understanding of this behavior).

These results have been obtained both for the disordered (using naive mean-field and TAP equations, which lead to similar results) and the non-disordered model (using formulae (3.35) and (3.36)) in order to check the artificial degree of heterogeneity which the inference procedure would have induced if the permutational symmetry among the N spins wouldn't have been known in advance. In figure 5.7 we plot an histogram of the off-diagonal elements of the connected correlation matrix $c_{ij} - m_i m_j$ and of the inferred couplings J_{ij} for a specific case. In figure 5.8 we plot the histogram of the eigenvalues obtained in the same case.

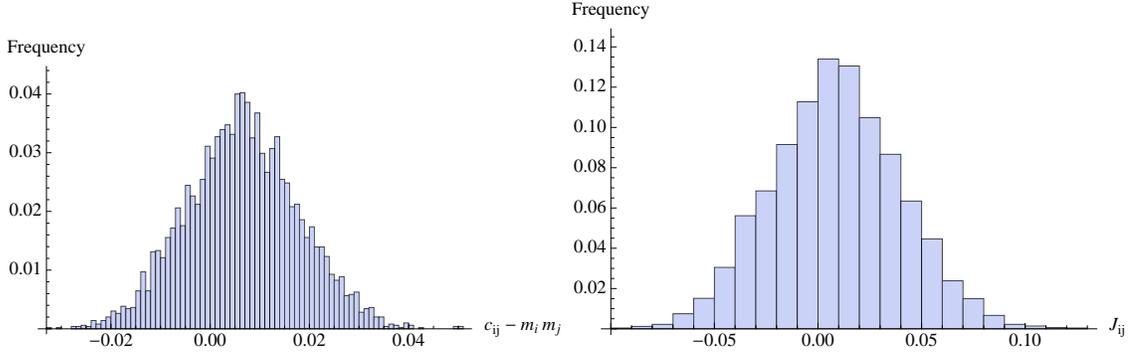


Figure 5.7: Histogram of the off-diagonal values of the correlation matrix $\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$ (left panel) and the inferred interaction matrix $\hat{\mathbf{J}}^*$ (right panel) for an Hawkes process defined by $\mu = 0.01 \text{ s}^{-1}$, $\alpha = 0.025 \text{ s}^{-1}$, $\beta = 0.03 \text{ s}^{-1}$, binned with a resolution of $\delta\tau \approx 30 \text{ s}$. Data corresponds to 5000 events, corresponding to approximately $T = \tau_{max}/\delta\tau \approx 4167$.

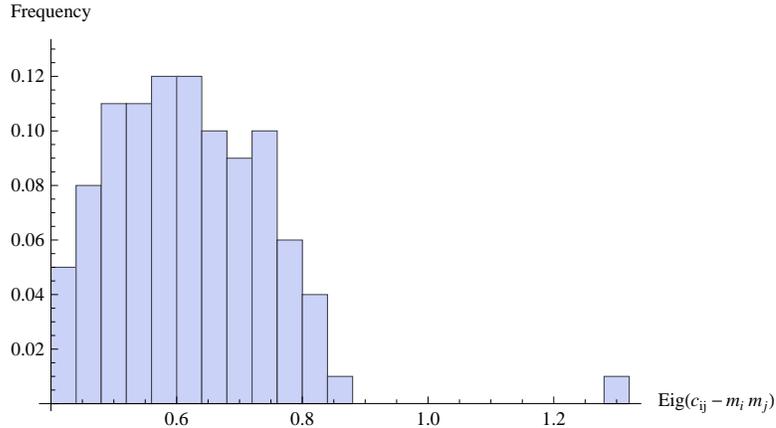


Figure 5.8: Histogram of the eigenvalues of the connected correlation matrix $\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$ for the Hawkes process described in figure 5.7. Notice that due to symmetry, one would expect for large T to have $N - 1$ degenerate eigenvalues of size $1 - m^2 - \delta c/N \approx 0.62$ and a larger eigenvalue of size $1 - m^2 + \delta c(N - 1)/N \approx 1.26$, whose associated eigenvector is of the form $(1, \dots, 1)/\sqrt{N}$.

5.3.2 Pairwise fully-connected model for NYSE trade events

We now focus on a dataset describing 100 days of trading activity (from 02.01.2003 to 05.30.2003) in the NYSE for the 100 most traded stocks. We consider only on the central part of each trading day ($\tau_{max} = 10^4 \text{ s}$), in order to avoid non-stationary effects linked with the opening and the closing hours of the market [19]. Any financial

transaction in this period has been defined as an event, independently on the buy or sell direction of the trade. The total data available allowed us to study 10^6 s of market activity corresponding to $\sim 10^5$ trade events, which have been binned by using sliding windows of size $\delta t \in \{2, \dots, 100\}$ s. The results obtained for the average magnetization and the average correlations as functions δt are reported in figure (5.9), in which it is possible to appreciate at which scale the magnetization changes from -1 to 1 (around 10 s), the one at which correlations form (~ 10 s) and decrease due to the presence of multiple events (~ 30 s).

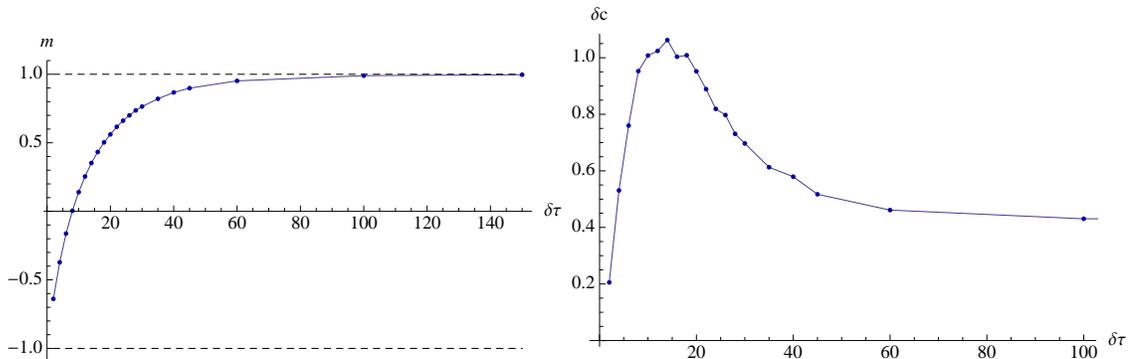


Figure 5.9: Average magnetization (left panel) and average correlation (right panel) for data corresponding to 100 days of financial transactions in the NYSE. $\delta\tau$ indicates the bin size in seconds, δc is the normalized correlation coefficient. The plot refers to a representative stock of the ensemble, specifically it is associated with the asset Analog Devices Inc. (ADI).

Features of the inferred model

By considering a fully-connected ferromagnet, such $\bar{\phi} = (m, c)$ data has been inverted in order to obtain the interactions $\mathbf{g} = (h, J)$, as shown in figure 5.10, where we also plotted the quantities $(\frac{1}{N} \sum_{i=1}^N h_i, \frac{2}{N(N-1)} \sum_{i<j} J_{ij})$ obtained by considering a disordered fully-connected ferromagnet. While for the non-disordered model we used formula (3.36) to invert the averages for the couplings, in the case of the disordered model we used mean-field equations – both naive Mean-Field (equations (3.10) and

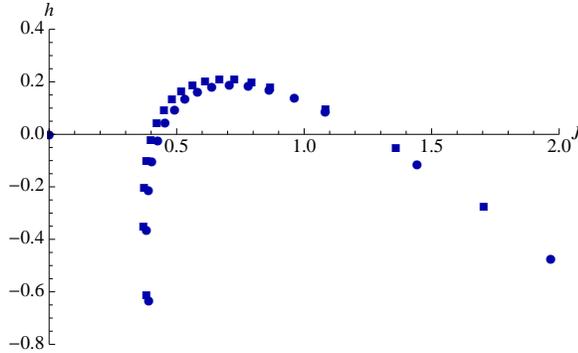


Figure 5.10: Inferred couplings h^* and J^* obtained with financial data, for various choices of the bin size $\delta\tau$. Squares indicate the result of inferring a homogeneous model, while circles indicate the averages of the vector \mathbf{h} and of the matrix $\hat{\mathbf{J}}$ obtained by inferring a disordered model.

(3.11)) and TAP equations (equations (3.14) and (3.15)) – which produced consistent results. We stress some features of the results we obtained:

- The ratio h/J changes according to δt , so that it is not possible to interpret h as measuring exogenous driving factors and J as a genuine interaction. Moreover, as explained in section 3.3), this inference procedure may mix interactions with external fields due to the approximate symmetry $\mathbf{g}^* \rightarrow \mathbf{g}^* + \delta J(-m, 1)$. What is possible to say is that the Hawkes process which best describes this (h, J) curve is defined by parameters $\mu \approx 0.011 \text{ s}^{-1}$, $\alpha \approx 0.022 \text{ s}^{-1}$, $\beta \approx 0.03 \text{ s}^{-1}$, so that the exogenous intensity μ corresponds to approximatively one fourth of the average intensity $\mathbb{E}[\lambda(\tau)]$.
- Results describing the Hawkes process allow to understand the proximity of the inferred parameters (h, J) to the critical point as related to the divergence of the average intensity $\mathbb{E}[\lambda(\tau)]$, rather than arising from a collective effect.
- As in the previous case, the inferred model doesn't lie on the critical line $h = 0$. This is due to the fact (section 3.3) that correlations are of the order of $1/N$, so that a description in terms of fully-connected ferromagnet leads to a non-degenerate description of the data.

Remark 5.4. *A procedure which has been proposed to estimate the distance of an inferred model from the critical point consists in rescaling of all the couplings by a common factor β (i.e., performing a shift $\mathbf{g}^* \rightarrow \beta\mathbf{g}^*$) which is interpreted as a fictitious inverse temperature. Studying how the elements of the susceptibility matrix $\hat{\chi}$ vary with respect to β should allow to identify criticality in the inferred model by the presence of peaks close to $\beta = 1$ in specific components of the matrix. We perform this procedure with our data and plot the results in figure 5.11, finding that:*

- *This procedure is not isotropic, in the sense that the shift $\mathbf{g} \rightarrow \beta\mathbf{g}$ implicitly indicates that the direction $(1, \dots, 1)$ should be the preferred one in order to evaluate distances in the coupling space.*
- *This type of measure does not describe the distance of the inferred model from the critical point in term of distinguishable models (equivalently, this measure of distance is not invariant under reparametrization of the statistical model (ϕ, \mathbf{g})).*

These points can lead to problems when model condensation is present: very different models may be described by slightly shifting β . Moreover an inferred model may lie close to the critical point due not only due to model condensation, but also due to the choice of a stable inference procedure, so that it is likely that $\hat{\chi}$ attains large value in the point \mathbf{g}^ , and that by moving from that point one can expect fluctuations to strongly decrease. A better measure of distance would be provided by considering geodesics in the coupling space under the Fisher metrics $\hat{\chi}$, as shown in section 5.1.2. We remind that properties (2.36) and (2.37) allows to informally identify this measure as counting how many error bars is one away from the critical point. This approach doesn't specify any privileged direction in the coupling space (as the geodesic distance is associated with whatever path in the coupling space is minimizing the number of such error bars), nor varies according to the reparametrization (as in that case also error bars are reparametrized accordingly). As an example, one finds that the distance $d_{T,\epsilon}(\mathbf{p}_{crit}, \mathbf{p}^*)$*

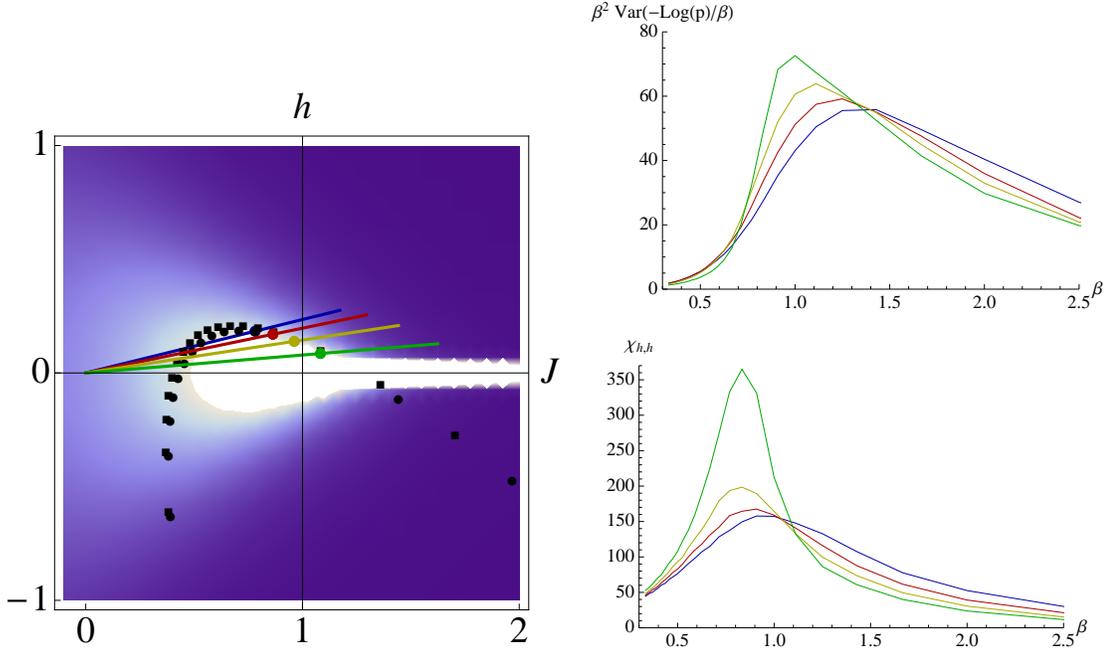


Figure 5.11: In the left panel we plot the regions of the phase space which are probed by shifting the inferred couplings for financial data \mathbf{g}^* by a fictitious inverse temperature β for various bin sizes (blue, red yellow and green correspond, respectively, to $\delta\tau = 24, 26, 28, 30$ s), while the background shows the model density $\rho(\mathbf{g}) \propto \det \hat{\chi}$. In the right plots we show the *specific heat* $\beta^2 \text{Var}[-\log p/\beta]$ and the *susceptibility* $\chi_{h,h}$ as a function of the inverse temperature for the same bin sizes as in the left plot.

defined by equation 5.19 between the critical point and the inferred parameters for $\delta\tau = 28$ s ($h^* \approx 0.14, J^* \approx 0.96$) is $d_{T,\epsilon}(\mathbf{p}_{crit}, \mathbf{p}^*) \gtrsim 10^2$ for $\epsilon = -\log 1\%$ and $T \sim 10^6$ s/28 s.

We also performed an analysis of the empirical connected correlation matrix $\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$ and of the inferred interaction matrix $\hat{\mathbf{J}}^*$ in order to check the compatibility of data with an homogeneous model. The corresponding histograms have been plotted in figure 5.12, showing that data is qualitatively similar to the one which would have been obtained with an homogeneous model. The principal component analysis of the matrices $\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$ and $\hat{\mathbf{J}}$ indicates in both cases the presence of a large eigenvalue, whose associated eigenvector is roughly of the form $\frac{1}{\sqrt{N}}(1, \dots, 1)$ as shown in the histogram (5.13). This findings can be interpreted as indicating that a significant part

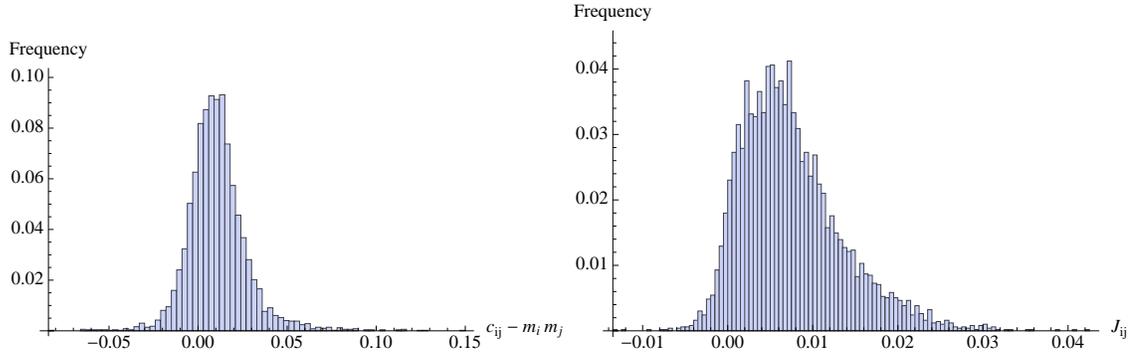


Figure 5.12: Histogram of the off-diagonal values of the correlation matrix $\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$ (left panel) and the inferred interaction matrix $\hat{\mathbf{J}}^*$ (right panel) for financial data, binned with a resolution of 30 s.

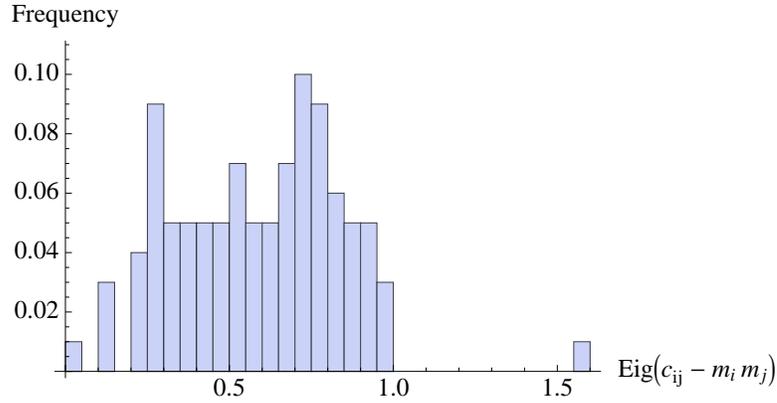


Figure 5.13: Histogram of the eigenvalues of the correlation matrix $\hat{\mathbf{c}} - \mathbf{m}\mathbf{m}^T$ for financial data, binned with a resolution of 30 s.

of the structure of the cross-excitatory network can be captured by an homogeneous model.³

³This is somewhat similar to what one finds for the statistics of stock price variations [51, 18]. In that case the correlation matrix has a large eigenvalue of size proportional to N (also called *market mode*), together with a small number of isolated eigenvalues, whose associated eigenvectors usually identify *financial sectors*. Interestingly, the inspection of the eigenvalues of $\hat{\mathbf{c}}$ and $\hat{\mathbf{J}}$ beyond the largest one evidences different sectors with respect to the ones found by studying stock price variations.

Chapter 6

Conclusion

In this work we have presented a general approach to the field of statistical learning, in which the problem of estimating parameters describing a complex system is seen as an inverse problem in the field of statistical mechanics. This perspective has been proven to be especially relevant in order to study extended systems, which in this language are associated to physical systems in the thermodynamic limit. This regime is well-known in physics, and several techniques (mean-field approximations) are known to solve the inverse problem in this framework with high accuracy. Interesting collective features emerge in this regime: models can *condensate* leading to regions of the space of parameters which are able to describe anomalously well very diverse datasets, and *null-modes* can develop leading to degenerate representations of a dataset. All of these factors have to be kept into account when studying inverse problem for empirical datasets, in order to disentangle the genuine features of a system from the spurious ones depending on the inference procedure which is applied. We have also shown that *complete representations* of the inverse problem lead to the exact solution of several systems (complete systems, one-dimensional systems, tree-like interaction networks), and allow a general understanding of the locality and stability features of the inverse problems, which are *easier* and more *resilient* to noise than the direct

ones. The notion of regularizer has also been discussed and its use has been clarified by specific, solvable examples, in order to show general features of non-parametric inference. We find that a symmetry property characterizes the regularizers, and a tradeoff between computational complexity and relevance of the inference procedure has to be sought on the basis of such symmetry in order to perform model selection. Finally, we have shown how differential geometry can be used to understand the consistency of the inverse problem, and how the special features associated with the large N limit have a clear geometric interpretation in terms of *distance* and *volume*. In this language, criticality of the inferred model is related to the strong divergence of the number of datasets which can be described through a small shift of the inferred parameters. Finally, we have presented the application of these ideas to two datasets, a synthetic one describing a self-excitatory point process and an empirical one describing transactions in a financial market. We used those datasets in order to illustrate our ideas by separating genuine features of the inferred model and spurious ones, finding that the dataset describing financial transactions can be well-described by a fully-connected ferromagnet in which interactions play a prominent role with respect to external driving factors.

Appendix A

Binary Inference

A.1 Maximum entropy principle

Consider a set of data $\hat{\mathbf{s}} = \{s^{(t)}\}_{t=1}^T$, and a family of operators ϕ . The maximum entropy principle states that among all probability distributions \mathbf{p} such that $\langle \phi \rangle = \bar{\phi}$ (where as usual $\bar{\phi} = \frac{1}{T} \sum_{t=1}^T \phi_{\mu, s^{(t)}}$), the one which maximizes the Shannon entropy $S(\mathbf{p})$ is given by the statistical model (2.1)

$$p_s^* = \frac{1}{Z(\mathbf{g})} \exp \left(\sum_{\mu=1}^M g_\mu^* \phi_{\mu, s} \right) \quad (\text{A.1})$$

in which each of the g_μ^* is seen as a Lagrange multiplier enforcing the condition $\langle \phi_\mu \rangle = \bar{\phi}_\mu$.

This principle is often invoked in order to justify the model (2.1) as the simplest (i.e., with higher entropy) one which is able to explain a given set of empirical averages [85]. Indeed it should be observed that this principle doesn't completely solve the problem of selecting the most appropriate model in order to explain data $\hat{\mathbf{s}}$, rather it converts it into the problem of selecting the best set of observables $\bar{\phi}$. In both cases a family ϕ has to be specified, and this has to be done on the basis of some *a priori* information (e.g., which operators are likely to be contained in the model), or

according to the specific goal of the inference problem which one is trying to solve (e.g., which observables are considered relevant for a particular application).

Proof. The proof of this result amounts to solve the constrained optimization problem

$$\mathbf{p}^* = \arg \max_{\mathbf{p}} \left[S(\mathbf{p}) + (g_0 + 1) + \sum_{\mu=1}^M g_{\mu} (\langle \phi_{\mu} \rangle - \bar{\phi}_{\mu}) \right], \quad (\text{A.2})$$

in which the Lagrange multipliers $\{g_{\mu}\}_{\mu=1}^M$ constrain the averages $\langle \phi_{\mu} \rangle$ to their empirical values, while g_0 enforces the normalization. By differentiation with respect to p_s , one can easily obtain equation (2.1). The conditions for the existence and the uniqueness of such solution are the same ones required in order to solve the inverse problem, and are described in section 2.2. \square

A.2 Concavity of the free energy

Consider the free energy $F(\mathbf{g})$ defined as in section 2.1. We want to prove that it is a concave function by showing that the susceptibility matrix $\hat{\chi}$ defined in equation (2.6) is positive semidefinite.

Proof. First one can show that

$$\chi_{\mu,\nu} = -\frac{\partial^2 F}{\partial g_{\mu} \partial g_{\nu}} = \sum_s (\phi_{\mu,s} - \langle \phi_{\mu} \rangle) (\phi_{\nu,s} - \langle \phi_{\nu} \rangle) p_s \quad (\text{A.3})$$

which allows to proof that for any vector \mathbf{x} , the quadratic form $\mathbf{x}^T \hat{\chi} \mathbf{x}$ is greater or equal than zero. In fact one has that

$$\sum_{\mu,\nu>0} x_{\mu} \chi_{\mu,\nu} x_{\nu} = \sum_s p_s \left[\sum_{\mu>0} x_{\mu} (\phi_{\mu,s} - \langle \phi_{\mu} \rangle) \right] \left[\sum_{\nu>0} x_{\nu} (\phi_{\nu,s} - \langle \phi_{\nu} \rangle) \right] \quad (\text{A.4})$$

$$= \left\langle \left[\sum_{\mu>0} x_{\mu} (\phi_{\mu,s} - \langle \phi_{\mu} \rangle) \right]^2 \right\rangle \geq 0. \quad (\text{A.5})$$

Additionally, if the operators $\phi_{\mu,s}$ are minimal in the sense defined in section 2.1 above expression has to be strictly larger than zero for $\mathbf{x} \neq \mathbf{0}$. In fact if $\mathbf{x}^T \hat{\chi} \mathbf{x} = 0$, then it must hold for each state s that

$$\sum_{\mu>0} (\phi_{\mu,s} - \langle \phi_{\mu} \rangle) x_{\mu} = 0 , \quad (\text{A.6})$$

which by minimality of ϕ implies that $x_{\mu} = 0$ for each μ . \square

A.3 Small deviations of the empirical averages

We want to show that given a statistical model (ϕ, \mathbf{g}) , equations (2.23) and (2.24) hold for the empirical averages $\bar{\phi}$.

Proof. For the averages, it is sufficient to show that due to the factorization property of $P_T(\hat{\mathbf{s}}|\mathbf{g})$ one has

$$\langle \bar{\phi}_{\mu} \rangle_T = \frac{1}{T} \sum_{t=1}^T \langle \phi_{\mu,s^{(t)}} \rangle_T = \frac{1}{T} \sum_{t=1}^T \langle \phi_{\mu} \rangle = \langle \phi_{\mu} \rangle , \quad (\text{A.7})$$

while for the covariances one can write

$$\langle \bar{\phi}_{\mu} \bar{\phi}_{\nu} \rangle_T - \langle \bar{\phi}_{\mu} \rangle_T \langle \bar{\phi}_{\nu} \rangle_T = \frac{1}{T^2} \sum_{t,t'=1}^T [\langle \phi_{\mu,s^{(t)}} \phi_{\nu,s^{(t')}} \rangle_T - \langle \phi_{\mu,s^{(t)}} \rangle_T \langle \phi_{\nu,s^{(t')}} \rangle_T] . \quad (\text{A.8})$$

By noting that due to independence all terms with $t \neq t'$ vanish from previous expression, one recovers equation (2.24). \square

A.4 Sanov theorem

We want to prove Sanov theorem (2.35), which states that given a probability distribution \mathbf{p} and a compact set of probability densities $\mathcal{M} \subseteq \mathcal{M}(\Omega)$, one has that the

empirical frequencies $\bar{\mathbf{p}}$ sampled from $P_T(\hat{\mathbf{s}}|\mathbf{p})$ obey the large deviation principle

$$\lim_{\delta \rightarrow 0} \lim_{T \rightarrow \infty} -\frac{1}{T} \log \text{Prob}(\bar{\mathbf{p}} \in \mathcal{M}') = D_{KL}(\mathbf{q}^*||\mathbf{p}) . \quad (\text{A.9})$$

where $\mathbf{q}^* = \arg \min_{\mathbf{q} \in \mathcal{M}} D_{KL}(\mathbf{q}||\mathbf{p})$ and \mathcal{M}' is the compact set $\mathcal{M}' = \{\mathbf{p}' = \mathbf{p} + \delta\mathbf{p} \in \mathcal{M}(\Omega) \mid \mathbf{p} \in \mathcal{M}, \delta\mathbf{p} \in [-\delta, \delta]^{|\Omega|}\}$

Proof. We will provide a simple combinatorial proof of Sanov theorem along the lines of [60], which requires some preliminary definitions. Given an empirical frequency $\bar{\mathbf{q}}$, we denote with $\hat{\mathbf{s}}(\bar{\mathbf{q}})$ the set $\hat{\mathbf{s}}(\bar{\mathbf{q}}) = \{\hat{\mathbf{s}} \in \Omega^T \mid \bar{q}_s = \frac{1}{T} \sum_{t=1}^T \delta_{s,s(t)}\}$ of empirical datasets compatible with $\bar{\mathbf{q}}$. We also define the set of all possible empirical frequencies as $\overline{\mathcal{M}}_T(\Omega)$. For those sets it holds that:

- The cardinality of $\hat{\mathbf{s}}(\bar{\mathbf{q}})$ is bound by

$$\frac{1}{\mathcal{P}_1(T)} e^{TS(\bar{\mathbf{q}})} \leq |\hat{\mathbf{s}}(\bar{\mathbf{q}})| \leq \mathcal{P}_2(T) e^{TS(\bar{\mathbf{q}})} , \quad (\text{A.10})$$

where $\mathcal{P}_1(T), \mathcal{P}_2(T)$ are polynomials in T with positive coefficients. This descends from applying Stirling bounds on the factorial to the exact relation

$$|\hat{\mathbf{s}}(\bar{\mathbf{q}})| = \frac{T!}{\prod_s (T\bar{q}_s)!} . \quad (\text{A.11})$$

and plugging the definition of Shannon entropy (2.11) in the resulting expression.

- The cardinality of $\overline{\mathcal{M}}_T(\Omega)$ is bound by

$$|\overline{\mathcal{M}}_T(\Omega)| \leq (T+1)^{|\Omega|} . \quad (\text{A.12})$$

because each configuration s is visited a number of times between 0 and T .

- Due to compactness of \mathcal{M} and continuity of $D_{KL}(\mathbf{q}||\mathbf{p})$, one has that $\min_{\mathbf{q} \in \mathcal{M}} D_{KL}(\mathbf{q}||\mathbf{p})$ exists and is attained in the (unique, due to convexity) point $\mathbf{q}^* \in \mathcal{M}$.

By using those properties, we can find an upper bound for the large deviation function as follows:

$$\begin{aligned}
 \text{Prob}(\bar{\mathbf{p}} \in \mathcal{M}') &= \sum_{\bar{\mathbf{q}} \in \overline{\mathcal{M}_T(\Omega)} \cap \mathcal{M}'} \text{Prob}(\bar{\mathbf{p}} = \bar{\mathbf{q}}) \\
 &= \sum_{\bar{\mathbf{q}} \in \overline{\mathcal{M}_T(\Omega)} \cap \mathcal{M}'} \sum_{\hat{\mathbf{s}} \in \hat{\mathbf{s}}(\bar{\mathbf{q}})} P_T(\hat{\mathbf{s}}|\mathbf{p}) \\
 &\leq \sum_{\bar{\mathbf{q}} \in \overline{\mathcal{M}_T(\Omega)} \cap \mathcal{M}'} \mathcal{P}_2(T) e^{TS(\bar{\mathbf{q}})} e^{-T[S(\bar{\mathbf{q}}) + D_{KL}(\bar{\mathbf{q}}||\mathbf{p})]} \\
 &\leq (T+1)^{|\Omega|} \mathcal{P}_2(T) e^{-TD_{KL}(\mathbf{q}^*||\mathbf{p})}. \tag{A.13}
 \end{aligned}$$

where $\mathbf{q}^* = \arg \min_{\mathbf{q} \in \mathcal{M}'} D_{KL}(\mathbf{q}||\mathbf{p})$. This trivially implies

$$\lim_{T \rightarrow \infty} \frac{1}{T} \log \text{Prob}(\bar{\mathbf{p}} \in \mathcal{M}) \leq -D_{KL}(\mathbf{q}^*||\mathbf{p}). \tag{A.14}$$

By taking the limit $\delta \rightarrow 0$, one recovers $\mathbf{q}^* \rightarrow \mathbf{q}^*$. For the lower bound, one needs to notice that for any δ it is possible to find a sufficiently large T and a $\delta\mathbf{p} \in [-\delta, \delta]^{|\Omega|}$ such that $\bar{\mathbf{q}}^* \in \overline{\mathcal{M}_T(\Omega)} \cap \mathcal{M}'$ is close enough to \mathbf{q}^* (due to density of rational numbers into real numbers), so that $|D_{KL}(\bar{\mathbf{q}}^*||\mathbf{p}) - D_{KL}(\mathbf{q}^*||\mathbf{p})| < \epsilon$ with ϵ arbitrary. Then one can write that

$$\text{Prob}(\bar{\mathbf{p}} \in \mathcal{M}') \geq \text{Prob}(\bar{\mathbf{p}} = \bar{\mathbf{q}}^*) = \sum_{\hat{\mathbf{s}} \in \hat{\mathbf{s}}(\bar{\mathbf{q}}^*)} P_T(\hat{\mathbf{s}}|\mathbf{p}) \geq \frac{1}{\mathcal{P}_1(T)} e^{-TD_{KL}(\bar{\mathbf{q}}^*||\mathbf{p})}, \tag{A.15}$$

which due to the arbitrariness of ϵ allows to prove the lower bound

$$\lim_{T \rightarrow \infty} \frac{1}{T} \log \text{Prob}(\bar{\mathbf{p}} \in \mathcal{M}) \geq -D_{KL}(\bar{\mathbf{q}}^*||\mathbf{p}) \geq -D_{KL}(\mathbf{q}^*||\mathbf{p}) - \epsilon. \tag{A.16}$$

□

A.5 Cramér-Rao bound

Cramér-Rao bound states that given a statistical model (ϕ, \mathbf{g}) with $F(\mathbf{g})$ strictly convex and an unbiased estimator of \mathbf{g} denoted as \mathbf{g}^* , the covariance matrix of \mathbf{g}^* under the measure $\langle \dots \rangle_T$ is bound according to equation (2.38).

Proof. First, it is necessary to prove that to prove that, after defining $V_\mu = \frac{\partial \log P_T(\hat{\mathbf{s}}|g)}{\partial g_\mu}$, and using equation (2.15) one has

$$\langle V_\mu \rangle_T = \langle T [\bar{\phi}_\mu - \langle \phi_\mu \rangle] \rangle_T = 0, \quad (\text{A.17})$$

where we also used equation (2.23) (i.e., $\langle \phi_\mu \rangle = \langle \phi_\mu \rangle_T$). Then, it is possible to compute the covariance

$$\begin{aligned} \text{Cov}(V_\mu, g_\nu^* - g_\nu) &= \langle V_\mu [g_\nu^* - g_\nu] \rangle_T - \langle V_\mu \rangle_T \langle g_\nu^* - g_\nu \rangle_T \\ &= \langle V_\mu g_\nu^* \rangle_T - \langle V_\mu \rangle_T g_\nu - \langle V_\mu \rangle_T \langle g_\nu^* - g_\nu \rangle_T \\ &= \sum_{\hat{\mathbf{s}}} \left[\frac{1}{P_T(\hat{\mathbf{s}}|g)} \frac{\partial P_T(\hat{\mathbf{s}}|g)}{\partial g_\mu} g_\nu^* \right] P_T(\hat{\mathbf{s}}|g) = \frac{\partial g_\nu}{\partial g_\mu} \\ &= \delta_{\mu,\nu} \end{aligned} \quad (\text{A.18})$$

and exploit Cauchy-Schwartz inequality, which implies that for any pair of vectors \mathbf{x}, \mathbf{y} it holds that

$$(\mathbf{x}^T \langle \mathbf{V}[\mathbf{g} - \mathbf{g}^*]^T \rangle_T \mathbf{y})^2 \leq \langle (\mathbf{x}^T \mathbf{V})^2 \rangle_T \langle ([\mathbf{g}^* - \mathbf{g}]^T \mathbf{y})^2 \rangle_T \quad (\text{A.19})$$

Equation (A.17) fixes the value of the left-hand side term of equation (A.19) to be $\mathbf{x}^T \mathbf{y}$, while the right-hand side can be expanded into

$$\langle (\mathbf{x}^T \mathbf{V})^2 \rangle_T \langle ([\mathbf{g}^* - \mathbf{g}]^T \mathbf{y})^2 \rangle_T = T (\mathbf{x}^T \hat{\chi} \mathbf{x}) (\mathbf{y}^T \langle [\mathbf{g}^* - \mathbf{g}] [\mathbf{g}^* - \mathbf{g}]^T \rangle_T \mathbf{y}), \quad (\text{A.20})$$

where we used that $\langle \mathbf{V} \mathbf{V}^T \rangle_T = T \hat{\chi}$ due to equation (2.24). Finally, by choosing the arbitrary vector \mathbf{x} to be $\mathbf{x} = \hat{\chi}^{-1} \mathbf{y} / T$ ($\hat{\chi}$ is invertible due to strict concavity of $F(\mathbf{g})$), it holds for any \mathbf{y} that

$$\frac{1}{T} (\mathbf{y}^T \hat{\chi}^{-1} \mathbf{y}) \leq (\mathbf{y}^T \langle [\mathbf{g}^* - \mathbf{g}] [\mathbf{g}^* - \mathbf{g}]^T \rangle_T \mathbf{y}) \quad (\text{A.21})$$

which proves the thesis (2.38). \square

A.6 Convergence of the inferred couplings

Given a set of empirical frequencies $\bar{\mathbf{p}}$, we want to prove that for a generic set of models described by an operator set ϕ the mean and the covariances of couplings \mathbf{g} defining a probability distribution \mathbf{p} , weighted by the measure provided by the posterior $P_T(\mathbf{g}|\bar{\mathbf{p}})$ are given in the limit of large T by equations (2.36) and (2.37).

Proof. To calculate them, we first notice that, by defining

$$\mathcal{Z}(\bar{\phi}) = \int d\mathbf{g} e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})} = \int d\mathbf{g} e^{T \sum_{\mu=0}^M g_{\mu} \bar{\phi}_{\mu}} \quad (\text{A.22})$$

it is possible to write

$$\frac{\partial \mathcal{Z}(\bar{\phi})}{\partial \bar{\phi}_{\mu}} = T \int d\mathbf{g} g_{\mu} e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})} \quad (\text{A.23})$$

$$\frac{\partial^2 \mathcal{Z}(\bar{\phi})}{\partial \bar{\phi}_{\mu} \partial \bar{\phi}_{\nu}} = T^2 \int d\mathbf{g} g_{\mu} g_{\nu} e^{-TD_{KL}(\bar{\mathbf{p}}|\mathbf{p})}, \quad (\text{A.24})$$

so that the calculation of the generating function $\log \mathcal{Z}(\bar{\phi})$ allows to find the required momenta of \mathbf{g} . In the limit of large T it is possible to perform a saddle-point estimation of the function $\mathcal{Z}(\bar{\phi})$ around the minimum of the convex function $D_{KL}(\bar{\mathbf{p}}||\mathbf{p})$ (or, equivalently, the maximum of the concave free energy $F(\mathbf{g})$), which requires the expansion of the Kullback-Leibler divergence. This procedure yields

$$\begin{aligned} \mathcal{Z}(\bar{\phi}) &= \int d\mathbf{g} \exp \left[T \left(F(\mathbf{g}^*) + \sum_{\mu=1}^M g_{\mu}^* \bar{\phi}_{\mu} + \frac{1}{2} \sum_{\mu,\nu} \frac{\partial^2 F(\mathbf{g})}{\partial g_{\mu} \partial g_{\nu}} (g_{\mu} - g_{\mu}^*)(g_{\nu} - g_{\nu}^*) + \dots \right) \right] \\ &\xrightarrow{T \rightarrow \infty} e^{-TS(\bar{\phi})} \sqrt{\frac{2\pi}{T \det \hat{\chi}}}, \end{aligned} \tag{A.25}$$

where – as shown in section 2.2 – the maximum likelihood estimator \mathbf{g}^* can be defined as the minimizer of the Kullback Leibler divergence. The differentiation of $\log \mathcal{Z}(\bar{\phi})$ finally leads to

$$\frac{1}{T} \frac{\partial \log \mathcal{Z}(\bar{\phi})}{\partial \bar{\phi}_{\mu}} \xrightarrow{T \rightarrow \infty} -\frac{\partial S(\bar{\phi})}{\partial \bar{\phi}_{\mu}} = g_{\mu}^* \tag{A.26}$$

$$\frac{1}{T^2} \frac{\partial^2 \log \mathcal{Z}(\bar{\phi})}{\partial \bar{\phi}_{\mu} \partial \bar{\phi}_{\nu}} \xrightarrow{T \rightarrow \infty} -\frac{1}{T} \frac{\partial^2 S(\bar{\phi})}{\partial \bar{\phi}_{\mu} \partial \bar{\phi}_{\nu}} = \frac{\chi_{\mu,\nu}^{-1}}{T}, \tag{A.27}$$

where we used equation (2.31) and (2.32) to express the derivatives of the entropy $S(\bar{\phi})$. □

Appendix B

High-dimensional inference

B.1 The fully-connected ferromagnet: saddle-point calculation

We want to prove that the free energy $F(h, J)$ of a fully connected ferromagnet described by the probability density (3.23) can be written as in (3.24).

Proof. This can be shown by noting that by using Stirling formula and approximating the sum with an integral one can write

$$\begin{aligned} Z(h, J) &= e^{-J/2} \sum_{N_+=0}^N \delta \left[Nm - \left(\frac{N + N_+}{2} \right) \right] \binom{N}{N_+} \exp \left[N \left(\frac{Jm^2}{2} + hm \right) \right] \\ &\xrightarrow{N \rightarrow \infty} e^{-J/2} \int_{-1}^1 dm \exp \left[N \left(\frac{Jm^2}{2} + hm + s(m) \right) \right], \end{aligned} \quad (\text{B.1})$$

with $s(m) = -\frac{1+m}{2} \log \frac{1+m}{2} - \frac{1-m}{2} \log \frac{1-m}{2}$. For (h, J) independent of N , above integral can be evaluated by saddle-point, and is dominated by the (absolute) minimum $m_{s.p.}(h, J)$ of the function $f_{h,J}(m) = -\frac{Jm^2}{2} - hm - s(m)$. By substituting

$F(h, J) = -\log Z(h, J)$ one finds

$$F(h, J) \xrightarrow{N \rightarrow \infty} \frac{J}{2} + N f_{h,J}(m_{s.p.}(h, J)) - \frac{1}{2} \log \frac{2\pi}{N \partial_m^2 f_{h,J}(m_{s.p.}(h, J))}, \quad (\text{B.2})$$

where $m_{s.p.}(h, J)$ satisfies the saddle-point equation

$$m = \tanh(Jm + h). \quad (\text{B.3})$$

Instead for large, finite N , $J > 1$ independent of N and $0 \leq h \propto 1/N$ equation (B.3) has two minima m_+ and m_- , whose contribution can be kept into account through

$$\begin{aligned} Z_+ + Z_- &= Z_+ \left(1 + \frac{Z_-}{Z_+}\right) = Z_+ e^{\log(1+Z_-/Z_+)} \\ &= Z_+ e^{-F_{trans}} \end{aligned} \quad (\text{B.4})$$

which yields the last term of equation (3.24). □

B.1.1 The leading contribution F_0 .

The main features of the model can be described by keeping into account the term $F_0(h, J)$, which is the only one in equation (3.24) proportional to N . It is given by

$$F_0(h, J) = N f_{h,J}(m_{s.p.}(h, J)), \quad (\text{B.5})$$

where

$$f_{h,J}(m) = -hm - \frac{Jm^2}{2} + \left(\frac{1+m}{2} \log \frac{1+m}{2} + \frac{1-m}{2} \log \frac{1-m}{2} \right), \quad (\text{B.6})$$

and $m_{s.p.}(h, J)$ is defined as the absolute minimum of the function $f_{h,J}(m)$, hence it satisfies the transcendental equation

$$m = \tanh(Jm + h) . \quad (\text{B.7})$$

The contribution of $F_0(h, J)$ to the ensemble averages is

$$\left\langle \sum_i s_i \right\rangle_0 = -\frac{\partial F_0}{\partial h} = N m_{s.p.} \quad (\text{B.8})$$

$$\left\langle \frac{1}{N} \sum_{i<j} s_i s_j \right\rangle_0 = -\frac{\partial F_0}{\partial J} = N \frac{m_{s.p.}^2}{2} , \quad (\text{B.9})$$

while the one to the susceptibility matrix $\hat{\chi}$ is given by

$$\hat{\chi}_0 = N \chi_{s.p.} \begin{pmatrix} 1 & m_{s.p.} \\ m_{s.p.} & m_{s.p.}^2 \end{pmatrix} , \quad (\text{B.10})$$

where $\chi_{s.p.} = \partial m_{s.p.} / \partial h$. Its eigenvalues are given by $N \chi_{s.p.} (0, 1 + m_{s.p.}^2)$.

The role of Gaussian fluctuations

The term $F_{fluct}(h, J)$ allows to compute the eigenvalue decomposition for the matrix $\hat{\chi}$, whose smallest eigenvalue receives a contribution which grows in N , and is related to the Gaussian integral (B.1). It results

$$F_{fluct}(h, J) = -\frac{1}{2} \log \left(\frac{2\pi}{N \partial_m^2 f_{h,J}(m_{s.p.}(h, J))} \right) . \quad (\text{B.11})$$

The contribution of $F_{fluct}(h, J)$ to the solution of the direct problem is

$$\left\langle \sum_i s_i \right\rangle_{fluct} = -\chi_{s.p.}^2 \frac{m}{(1-m^2)^2} \quad (\text{B.12})$$

$$\left\langle \frac{1}{N} \sum_{i<j} s_i s_j \right\rangle_{fluct} = -\frac{\chi_{s.p.}^2}{2} \left(J - \frac{1-3m^2}{(1-m^2)^2} \right) \quad (\text{B.13})$$

and

$$\hat{\chi}_{fluct} = \chi_{s.p.}^4 (1-m_{s.p.}^2)^{-3} \hat{\mathbf{a}}(J, m), \quad (\text{B.14})$$

with

$$a_{11}(J, m) = -(1 - J - 3Jm^2) \quad (\text{B.15})$$

$$a_{12}(J, m) = a_{2,1}(J, m) = -(3 - 3J - 3Jm^2) \quad (\text{B.16})$$

$$\begin{aligned} a_{22}(J, m) &= 1 - 2J + J^2 - 11m^2 + 14Jm^2 \\ &\quad - 3J^2m^2 - 4Jm^4 + 3J^2m^4 - J^2m^6 \end{aligned} \quad (\text{B.17})$$

B.1.2 Transition line and metastability

The function $f_{h,J}(m)$ may display either one or two local minima according to the value of the couplings h and J . In the case $h \geq 0$ that we are considering, whenever two local minima m_+ and m_- are present, one has $m_{s.p.} = m_+$ with $\delta f_{h,J} = f_{h,J}(m_-) - f_{h,J}(m_+) \geq 0$. The contribution of the *state* m_- to the saddle point integral vanishes in the large N limit as long as $\delta f_{h,J}$ is finite, but for $\delta f_{h,J} \approx 1/N$, the contribution of the m_- cannot be neglected, and requires the introduction of a term in the free energy of the form

$$F_{trans}(h, J) = -\log \left(1 + e^{-N\delta f_{h,J}} \sqrt{\frac{J - \frac{1}{1-m_+^2}}{J - \frac{1}{1-m_-^2}}} \right). \quad (\text{B.18})$$

For small enough values of h , the values of the minima become $m_+ = -m_-$, and above term can be written as

$$F_{trans}(h, J) = -\log(1 + e^{-2Nhm_{s.p.}}) . \quad (\text{B.19})$$

Hence, this term describes the region of the coupling space which we call *transition line*, where $hm_{s.p.} \ll 1/N$. The contribution to the averages and to the generalized susceptibility of this term is given by

$$\left\langle \sum_i s_i \right\rangle_{trans} = -N [1 - \tanh(Nhm_{s.p.})](h\chi_{s.p.} + m_{s.p.}) \quad (\text{B.20})$$

$$\left\langle \frac{1}{N} \sum_{i<j} s_i s_j \right\rangle_{trans} = N h m_{s.p.} \chi_{s.p.} [1 - \tanh(Nhm_{s.p.})] \quad (\text{B.21})$$

and

$$\hat{\chi}_{trans} = N^2 \hat{\mathbf{b}}(h, J, m) . \quad (\text{B.22})$$

The matrix $\hat{\mathbf{b}}(h, J, m)$ (whose explicit form is not particularly illuminating) can be obtained by deriving above averages with respect to h and J .

Determinant of the generalized susceptibility

The term $\sqrt{\det \hat{\chi}}$ is shown in chapter 5 to be relevant in order to count the number of distinguishable statistical models inside a given region of the space (h, J) . It can be calculated at leading order in N by keeping into account the different contributions to the free energy $F(h, J)$. The region in which $|h| \gg 1/N$ is described by $F \xrightarrow{N \rightarrow \infty} F_0 + F_{fluct}$, and it results

$$\begin{aligned} \det \hat{\chi} &\xrightarrow{N \rightarrow \infty} \det(\hat{\chi}_0 + \hat{\chi}_{fluct}) \xrightarrow{N \rightarrow \infty} \det \hat{\chi}_0 + \frac{N}{2} \chi_{s.p.}^3 \\ &= \frac{N}{2} \chi_{s.p.}^3 , \end{aligned} \quad (\text{B.23})$$

while the region $h \ll 1/N$ is dominated by the contribution $F_0 + F_{trans}$, implying

$$\det \hat{\chi} \xrightarrow{N \rightarrow \infty} \det(\hat{\chi}_0 + \hat{\chi}_{trans}) \xrightarrow{N \rightarrow \infty} N^3 \left(\frac{m_{s.p.}^4 \chi_{s.p.}}{\cosh^2(Nhm_{s.p.})} \right) + O(N^2). \quad (\text{B.24})$$

B.1.3 Marginal polytope for a fully connected ferromagnet

We want to characterize the marginal polytope $\mathcal{G}(\phi)$ for the fully connected ferromagnet (3.23), that is, the set of empirical averages $(m, c) \in \mathbb{R}^2$ compatible with at least one probability density $p \in \mathcal{M}(\Omega)$.

Proof. Due to density of the empirical frequency \bar{p} in the space $\mathcal{M}(\Omega)$, we will consider the large T limit of a sequence of observations $\{m^{(t)}\}_{t=1}^T$. Fixed any $m \in [-1, 1]$, one needs to require

$$m = \frac{1}{T} \sum_{t=1}^T m^{(t)} \quad (\text{B.25})$$

and ask for a possible arrangement of the sequence $\{m^{(t)}\}_{t=1}^T$ compatible with a correlation c , that is,

$$c = \frac{1}{T} \sum_{t=1}^T \frac{(m^{(t)})^2 - 1/N}{1 - 1/N}, \quad (\text{B.26})$$

where, after easy combinatorics, we used the fact that the correlation $c^{(t)}$ measured in the observation number t depends just upon the total magnetization $m^{(t)}$. Finding a solution to this problem is easy due to convexity of $\sum_t (m^{(t)})^2$. In particular by taking the limit $T \rightarrow \infty$ a solution can be found for any m , while then the minimum and the maximum value of c are given respectively by $\frac{m^2 - 1/N}{1 - 1/N}$ and 1. Interestingly, the same result can be obtained with more simplicity by exploiting the necessary condition $\text{Var}[\sum_i s_i] \geq 0$. Notice also that for large N the connected correlation coefficient $c - m^2$ is bound from below by $(m^2 - 1)/N$: equivalently no large system, subject to whatever type of interaction, can be globally anti-correlated. \square

Appendix C

Convex optimization

In this appendix we will briefly remind part of the theory which has been developed in order to solve unconstrained minimization problems of convex functions of the form $H(\mathbf{g}) : \mathbb{R}^M \rightarrow \mathbb{R}$, addressing the interested reader to [21] for a more complete analysis.

C.1 Differentiable target

Consider a convex, differentiable function $H(\mathbf{g}) : \mathbb{R}^M \rightarrow \mathbb{R}$. Then for each point \mathbf{g} it exists a *gradient* $\nabla H(\mathbf{g}) = \left(\frac{\partial}{\partial g_1}, \dots, \frac{\partial}{\partial g_M} \right) H(\mathbf{g})$ and a positive semi-definite *Hessian* matrix $\hat{\chi}(\mathbf{g})$ with elements $\chi_{\mu,\nu} = \partial_\mu \partial_\nu H(\mathbf{g})$. Then the following properties hold:

1. The gradient is a global under-estimator of $H(\mathbf{g})$, namely for any \mathbf{g}' one has that

$$H(\mathbf{g}) \geq H(\mathbf{g}') + \nabla H(\mathbf{g}')^T (\mathbf{g} - \mathbf{g}') . \quad (\text{C.1})$$

2. The gradient defines a *descent direction* $\mathbf{v} = -\nabla H(\mathbf{g})$, which means that for all \mathbf{g} it exists an ϵ such that

$$H(\mathbf{g} - \epsilon \nabla H(\mathbf{g})) \leq H(\mathbf{g}) \quad (\text{C.2})$$

3. The Hessian defines the descent direction $\mathbf{v} = -\hat{\chi}^{-1}(\mathbf{g})\nabla H(\mathbf{g})$. Algorithms exploiting this property usually go under the name of *Newton's methods*.

These properties are simple consequences of differentiability and convexity of $H(\mathbf{g})$, and allow to solve the problem of its minimization. The first property implies that given a \mathbf{g} such that $\nabla H(\mathbf{g}) = \mathbf{0}$, \mathbf{g} is a global minimum of $H(\mathbf{g})$. If this equation can be explicitly solved, the minimum can be found. Indeed if, as it often is the case, the condition $\nabla H(\mathbf{g}) = \mathbf{0}$ is non-analytically solvable, it is possible to exploit properties 2. and 3. in order to build iterative algorithms which decrease the target function $H(\mathbf{g})$ at each step. In particular, iterative algorithms exploiting property 2. are expected to achieve linear convergence to the minimum, while more sophisticated algorithms (Newton methods) constructed by using the Hessian can achieve quadratic convergence. More efficient schemes (quasi-Newton methods) such as the L-BFGS approximation [49, 23] exploit an approximation for the Hessian in order to save memory and computational power by exploiting successive updates of the gradient. We present in the following an example of a simple algorithm which can be used to minimize a convex differentiable $H(\mathbf{g})$, which we use mainly as a proof of principle for the solvability of this type of problem. Secondly, the efficiency of the Boltzmann learning algorithm presented in section 3.1 is rooted in the gradient descent method.

C.1.1 Gradient descent algorithm

Given a convex, differentiable $H(\mathbf{g})$ and starting point $\mathbf{g}^{(0)}$, we consider a sequence $\{\mathbf{g}^{(k)}\}_{k=1}^K$ built according to the iterative scheme

$$\mathbf{g}^{(k+1)} = \mathbf{g}^{(k)} - \epsilon_k \nabla H(\mathbf{g}^{(k)}), \quad (\text{C.3})$$

where we introduced the *schedule* $\{\epsilon_k\}_{k=1}^K$. Suppose that each of the ϵ_k is chosen in order to satisfy the (Armijo) condition

$$H(\mathbf{g}^{(k+1)}) \leq H(\mathbf{g}^{(k)}) - \epsilon_k \beta \|\nabla H(\mathbf{g})\|^2, \quad (\text{C.4})$$

for a given $0 < \beta < 1$, by considering the initial value $\epsilon_k = 1$ and iterating the map $\epsilon_k \leftarrow \epsilon_k / \tau$ for $\tau < 1$ until (C.4) is satisfied. Then it holds that either $\min_{\mathbf{g} \in \mathbb{R}^M} H(\mathbf{g}) = -\infty$ or $\lim_{k \rightarrow \infty} \|\nabla H(\mathbf{g})\|^2 = 0$, that is, if a minimum exists, the sequence $\{\mathbf{g}^{(k)}\}_{k=1}^K$ can approximate it with arbitrary precision.

Remark C.1. *Searching the optimal ϵ_k is usually called a line search, and the procedure that we introduce to find it is guaranteed to find an ϵ_k satisfying (C.4) if the maximum eigenvalue of $\hat{\chi}$ is bounded by a given χ_{\max} . In particular the convexity of $H(\mathbf{g})$ and a straightforward application of Taylor theorem allow to prove that any ϵ_k in the interval*

$$0 \leq \epsilon_k \leq \frac{2(1 - \beta)}{\chi_{\max}} \quad (\text{C.5})$$

satisfies the Armijo condition (C.4).

Proof. In order to prove that the convergence of the algorithm, one can use (C.4) to iteratively build the inequality

$$H(\mathbf{g}^{(K)}) \leq H(\mathbf{g}^{(K-1)}) - \beta \epsilon_k \|\nabla H(\mathbf{g}^{(K-1)})\|^2 \leq H(\mathbf{g}^{(0)}) - \beta \sum_{k=0}^{K-1} \epsilon_k \|\nabla H(\mathbf{g}^{(k)})\|^2. \quad (\text{C.6})$$

Then, as the succession $H(\mathbf{g}^{(K)}) - H(\mathbf{g}^{(0)})$ is strictly decreasing in K , it has a limit. Such limit can be either $-\infty$ (in which case $H(\mathbf{g})$ has no minimum) or can be finite. The finiteness of the limit implies that

$$H(\mathbf{g}^{(\infty)}) - H(\mathbf{g}^{(0)}) = \lim_{K \rightarrow \infty} \sum_{k=0}^{K-1} \epsilon_k \|\nabla H(\mathbf{g}^{(k)})\|^2 \quad (\text{C.7})$$

which leads to

$$\lim_{K \rightarrow \infty} \epsilon_K \|\nabla H(\mathbf{g}^{(K)})\|^2 = 0. \quad (\text{C.8})$$

□

Notice that the rate of convergence in K of this algorithm can be rather slow, which is the reason why more sophisticated algorithms are commonly used to perform this task (see [21]).

C.2 Non-differentiable target

If a convex function is not differentiable in all of its domain the solution of the minimization problem is technically more complicated, but it is still possible to take advantage of the convexity property in order to build efficient minimization algorithms (see [20, 21]). Consider a convex function $H(\mathbf{g}) : \mathbb{R}^M \rightarrow \mathbb{R}$. Then one can define a *sub-gradient* as any global underestimator of $H(\mathbf{g})$, namely $\mathbf{v} \in \mathbb{R}^M$ is a subgradient of $H(\mathbf{g})$ in \mathbf{g}' if for any \mathbf{g} one has

$$H(\mathbf{g}) \geq H(\mathbf{g}') + \mathbf{v}^T(\mathbf{g} - \mathbf{g}'). \quad (\text{C.9})$$

The set of all the sub-gradients of $H(\mathbf{g})$ in \mathbf{g}' is called the *sub-differential* of $H(\mathbf{g})$, and is denoted with $\tilde{\nabla}H(\mathbf{g})$. One can show that

- $\tilde{\nabla}H(\mathbf{g})$ is non-empty if $H(\mathbf{g})$ is locally convex and bounded around \mathbf{g} .
- $\tilde{\nabla}H(\mathbf{g})$ is closed and convex.
- The sub-differential is *additive*, so that $\tilde{\nabla}[H_1(\mathbf{g}) + H_2(\mathbf{g})] = \tilde{\nabla}H_1(\mathbf{g}) + \tilde{\nabla}H_2(\mathbf{g})$.
- The sub-differential has the *scaling* property $\tilde{\nabla}\lambda H(\mathbf{g}) = \lambda\tilde{\nabla}H(\mathbf{g})$ for $\lambda > 0$.
- If $H(\mathbf{g})$ is differentiable, then $\tilde{\nabla}H(\mathbf{g}) = \{\nabla H(\mathbf{g})\}$.

This properties characterize the sub-differential as a notion generalizing the ordinary differential, which is suitable to solve problems involving non-differentiable functions. In particular the properties shown above for differentiable functions generalize to:

1. If $\mathbf{0} \in \tilde{\nabla}H(\mathbf{g})$ then \mathbf{g} is a global minimum of $H(\mathbf{g})$.
2. The direction $\mathbf{v} = -\epsilon\tilde{\nabla}H(\mathbf{g})$ is not in general a descent direction.

This implies that in order to minimize a non-differentiable function it is still possible to find the points whose sub-differential is equal to zero, but that a naive sub-gradients descent similar to (C.3) is not guaranteed to find a solution.

An example: the absolute value

Consider the function $H(g) : \mathbb{R} \rightarrow \mathbb{R}$ defined as $H(g) = H_d(g) + |g|$, with $H_d(g)$ convex and differentiable. Then the sub-differential of $H(\vec{g})$ is given by

$$\tilde{\nabla}H(g) = \nabla H_d(g) + \text{sgn}(g) \tag{C.10}$$

where

$$\text{sgn}(g) = \begin{cases} \text{sign}(g) & \text{if } g \neq 0 \\ [-1, 1] & \text{if } g = 0 \end{cases} . \tag{C.11}$$

which is minimum for

$$\begin{aligned} x = 0 & \quad \text{if } |\nabla H_d(0)| \leq 1 \\ x \gtrless 0 & \quad \text{if } \nabla H_d(x) = \mp 1 . \end{aligned} \tag{C.12}$$

The notion of sub-gradient also allow us to generalize the gradient descent algorithm to non-differentiable functions, as shown in the following.

C.2.1 Sub-gradient descent algorithm

Consider a convex $H(\mathbf{g})$, a starting point $\mathbf{g}^{(0)}$, and a sequence $\{\mathbf{g}^{(k)}\}_{k=1}^K$ built according to the iterative scheme

$$\mathbf{g}^{(k+1)} = \mathbf{g}^{(k)} - \epsilon_k \mathbf{v}^{(k)}, \quad (\text{C.13})$$

where $\mathbf{v}^{(k)} \in \tilde{\nabla}H(\mathbf{g}^{(k)})$ is a sub-gradient in $\mathbf{g}^{(k)}$, and where we introduced the schedule $\{\epsilon_k\}_{k=1}^K$. Then, one can show that if $H(\mathbf{g})$ has a minimum \mathbf{g}^* , then

$$H(\mathbf{g}^{best}) - H(\mathbf{g}^*) \leq \frac{R^2 + G^2 \sum_{k=1}^K \epsilon_k^2}{2 \sum_{k=1}^K \epsilon_k} \quad (\text{C.14})$$

where $\mathbf{g}^{best} = \arg \min_{\mathbf{g} \in \mathbf{g}^{(k)}} H(\mathbf{g})$, while R and G enforce respectively a bound of the initial distance from the minimum $\|\mathbf{g}^{(1)} - \mathbf{g}^*\|^2 \leq R$ and the Lipschitz bound $\frac{|H(\mathbf{g}) - H(\mathbf{g}')|}{\|\mathbf{g} - \mathbf{g}'\|} \leq G^1$. In particular, by choosing $\epsilon_k \propto 1/k$, one can show that in that case

$$\lim_{k \rightarrow \infty} H(\mathbf{g}^{best}) - H(\mathbf{g}^*) = 0 \quad (\text{C.15})$$

Proof. To prove this result, it is necessary to consider the Euclidean distance to the minimum \mathbf{g}^* , which due to the property (C.9) satisfies

$$\begin{aligned} \|\mathbf{g}^{(K)} - \mathbf{g}^*\|^2 &= \|\mathbf{g}^{(K-1)} - \epsilon_{K-1} \mathbf{v}^{(K-1)} - \mathbf{g}^*\|^2 \\ &= \|\mathbf{g}^{(K-1)} - \mathbf{g}^*\|^2 - 2\epsilon_{K-1} (\mathbf{g}^{(K-1)} - \mathbf{g}^*)^T \mathbf{v}^{(K-1)} + \epsilon_k^2 \|\mathbf{v}^{(K-1)}\|^2 \\ &\leq \|\mathbf{g}^{(K-1)} - \mathbf{g}^*\|^2 - 2\epsilon_{K-1} (H(\mathbf{g}^{(K-1)}) - H(\mathbf{g}^*)) + \epsilon_k^2 \|\mathbf{v}^{(K-1)}\|^2, \end{aligned} \quad (\text{C.16})$$

¹Although the hypothesis of Lipschitz bounded $H(\mathbf{g})$ is not strictly required, for the sake of clarity we have chosen to present the algorithm in this simpler form.

so that one can recursively build the inequality

$$\begin{aligned}
 \|\mathbf{g}^{(K)} - \mathbf{g}^*\|^2 &\leq \|\mathbf{g}^{(0)} - \mathbf{g}^*\|^2 - 2 \sum_{k=0}^{K-1} \epsilon_k (H(\mathbf{g}^{(k)}) - H(\mathbf{g}^*)) + \sum_{k=0}^{K-1} \epsilon_k^2 \|\mathbf{v}^{(k)}\|^2 \\
 &\leq R^2 - 2[H(\mathbf{g}^{best}) - H(\mathbf{g}^*)] \sum_{k=0}^{K-1} \epsilon_k + G^2 \sum_{k=0}^{K-1} \epsilon_k^2. \tag{C.17}
 \end{aligned}$$

Finally, by using $\|\mathbf{g}^{(K)} - \mathbf{g}^*\|^2 \geq 0$, one can rearrange the terms and obtain the bound (C.14). □

Notice that in this case the sequence ϵ_k is not optimized on-line, rather it is fixed at the beginning of the algorithm. This is because the sub-gradient doesn't specify necessarily a descent direction, hence the sub-gradient descent may increase the function $H(\mathbf{g})$, requiring the values \mathbf{g}^{best} and $H(\mathbf{g}^{best})$ to be stored at each iteration step.

Appendix D

Complete families

D.1 Rate of convergence for the complete inverse problem

Consider a statistical model \mathbf{p} in which all states have strictly positive probability (i.e. it exists a $p_{min} \neq 0$ such that $\forall s \quad p_{min} \leq p_s$). We want to show how within inference scheme (4.14) the inferred couplings concentrate around their actual values at fixed N in the limit $T \rightarrow \infty$.

Proof. The expression for g_μ^* is:

$$g_\mu^* = \frac{1}{|\Omega|} \sum_s \phi_{\mu,s} \log \bar{p}_s, \quad (\text{D.1})$$

while the probability to observe a given set empirical frequencies $\bar{\mathbf{p}}$ out of the measure of T samples is given by the multinomial distribution described in section 2.1.4. Its mean and correlations are sufficient to completely determine the convergence for large enough values of T . In particular one finds that

$$\langle g_\mu^* \rangle_T \xrightarrow{T \rightarrow \infty} \frac{1}{|\Omega|} \sum_s \phi_{\mu,s} \log \langle \bar{p}_s \rangle = g_\mu, \quad (\text{D.2})$$

while the fluctuations of the inferred couplings are equal to

$$\text{Var}(g_\mu^*) = \langle (g_\mu^*)^2 \rangle_T - \langle g_\mu^* \rangle_T^2 \xrightarrow{T \rightarrow \infty} \frac{1}{|\Omega|^2} \sum_{s,s'} \phi_{\mu,s} \phi_{\mu,s'} \frac{\text{Cov}(\bar{p}_s, \bar{p}_{s'})}{p_s p_{s'}} \quad (\text{D.3})$$

$$= \frac{1}{T} \left[\left(\frac{1}{|\Omega|^2} \sum_s \frac{1}{p_s} \right) - \delta_{\mu 0} \right], \quad (\text{D.4})$$

which is the result shown in equation (4.15). This can be generalized to the case in which the set of states with strictly positive probabilities is a subset $\mathcal{I} \subset \Omega$, so that one can define the set of *regular* operators $\phi^{reg} = \{\phi_\mu \in \phi \mid \sum_{s \in \mathcal{I}} \phi_{\mu,s} = 0\}$. The same proof as above can be performed for regular operators on the estimator

$$g_\mu^{*reg} = \frac{1}{|\Omega|} \sum_{s \in \mathcal{I}} \phi_{\mu,s} \log \bar{p}_s, \quad (\text{D.5})$$

finding the result described in equation (??). □

D.2 Factorization property for tree-like models

In this section we prove a fundamental property of statistical models whose the interaction structure is loop-less, which we call *trees* and rigorously define as follows.¹

Definition D.1. Consider a statistical model (ϕ, \mathbf{g}) of the form (2.1), with $g_\mu \neq 0$ for all $g_\mu \in \mathbf{g}$. Then the set ϕ is called a *tree* if it is not possible to find a *cycle* connecting any set of vertices, i.e., it doesn't exist a closed path $\{i_1, \dots, i_{L-1}, i_L = i_1\} \in V^L$ such that for each couple $\{i_n, i_{n+1}\}$ there exist an operator $\phi_{i_n, i_{n+1}} \in \phi$ depending on both s_{i_n} and $s_{i_{n+1}}$, with $\phi_{i_n, i_{n+1}} \neq \phi_{i_m, i_{m+1}}$ for all $n \neq m \in \{1, \dots, L-1\}$.

For trees we will show along the lines of [54] that the following factorization property holds.

¹This definition corresponds to what is often referred in literature as a *forest*, while the word *tree* is typically reserved to each connected component of a forest. For simplicity we will disregard such difference, and make no distinction among trees and forests.

Theorem D.1. Consider a tree-like statistical model (ϕ, \mathbf{g}) . Then its associated probability density \mathbf{p} can be written as

$$p(s) = \prod_{\mu=1}^M p^{\partial\phi_\mu}(s^{\partial\phi_\mu}) \prod_{i \in V} p^{\{i\}}(s^i)^{1-|\partial i|}, \quad (\text{D.6})$$

where $\partial i = \{\phi_\mu \in \phi \mid \phi_\mu(s) \text{ depends upon } s_i\}$ while $\partial\phi = \{i \in V \mid \phi(s) \text{ depends upon } s_i\}$.

Proof. The theorem can be proved by induction on the number of operators M . Consider the case $M = 1$ in which just one operator is present ($\phi = \{\phi\}$). Then, it is trivial to see that equation (D.6) holds due to

$$p(s) \propto \exp[g\phi(s)] \propto p^{\partial\phi}(s^{\partial\phi}) \prod_{i \in V \setminus \partial\phi} p^{\{i\}}(s_i). \quad (\text{D.7})$$

Let then property (D.6) hold for the case of M operators, and consider a statistical model in which $|\phi| = M + 1$. Then, as (ϕ, \mathbf{g}) is a tree, it is possible to consider without loss of generality an operator $\phi_\mu \in \phi$ such that $|\partial j| = 1$ for all $j \in \partial\phi_\mu$ but at most a single variable. Suppose that such variable exists, and label it as i . Then by defining the cluster $\Gamma = \{j \in V \mid j \notin \partial\phi_\mu\} \cup \{i\}$, a straightforward application of Bayes rule yields

$$\begin{aligned} p(s) &= p^\Gamma(s^\Gamma) p^{V \setminus \Gamma}(s^{V \setminus \Gamma} | s^\Gamma) \\ &= p^\Gamma(s^\Gamma) \frac{p^{\partial\phi_\mu}(s^{\partial\phi_\mu})}{p^{\{i\}}(s^{\{i\}})}. \end{aligned} \quad (\text{D.8})$$

Additionally, the marginal $p^\Gamma(s^\Gamma)$ can be written in the form

$$p^\Gamma(s^\Gamma) \propto \exp \left(\sum_{\phi_\nu \in \phi \setminus \partial\phi_\mu} g_\nu \phi_\nu(s^\Gamma) \right) \underbrace{\sum_{s_j \mid j \notin \Gamma} \exp(g_\mu \phi_\mu(s^{\partial\phi_\mu}))}_{\equiv \psi(s_i)}. \quad (\text{D.9})$$

Then it is possible to reabsorb the $\psi(s_i)$ factor inside a new operator obtained by the following change on a generic $\phi_\rho \in \partial i \setminus \phi_\mu$:

$$\phi_\rho(s^{\partial\phi_\rho}) \rightarrow \phi'_\rho(s^{\partial\phi_\rho}) = \phi_\rho(s^{\partial\phi_\rho}) + \frac{\log \psi(s_i)}{g_\rho}. \quad (\text{D.10})$$

The statistical model describing the reduced problem for Γ spins can thus be described by using M operators, so that it is possible to use the inductive hypothesis to show that

$$p^\Gamma(s^\Gamma) = p^{\{i\}}(s_i)^{1-(|\partial i|-1)} \prod_{\phi_\nu \in \phi \setminus \{\phi_\mu\}} p^{\partial\phi_\nu}(s^{\partial\phi_\nu}) \prod_{j \in V \setminus \{i\}} p^{\{j\}}(s_j)^{1-|\partial j|}. \quad (\text{D.11})$$

Above expression can finally be plugged into equation (D.8) so to obtain equation (D.6). In order to prove the thesis (D.6) in full generality it is nevertheless necessary to perform an analogous derivation in the case in which no such i variable exist, an exercise which for the sake of conciseness we leave to the reader. \square

D.3 Factorization property of the one-dimensional periodic chain

Consider a one-dimensional periodic chain of size N , range R and periodicity ρ defined by a complete, orthogonal set of operators ϕ and a set of translation operators \mathbf{T} . We want to show that for such chain it holds the factorization property

$$p(s) = \prod_{n=0}^{N/\rho-1} \frac{p^{\Gamma_n}(s^{\Gamma_n})}{p^{\gamma_n}(s^{\gamma_n})}. \quad (\text{D.12})$$

where the sets Γ_n and γ_n are defined as in section 4.2.4.

Proof. To obtain this result, one needs to define a two-dimensional model defined by the log-probability

$$\begin{aligned} \log p_\lambda(s, t) = & -\log Z(\mathbf{g}) + \sum_{n=0}^{N/\rho-1} \sum_{\mu \in \phi} g_\mu \phi_\mu(s_{1+n\rho}^n, \dots, s_{R+n\rho}^n) \\ & + \lambda \sum_{n=0}^{N/\rho-1} \sum_{i=(n+1)\rho+1}^{n\rho+R} [(t_i^n - s_i^n)^2 + (t_i^n - s_i^{n+1})^2], \end{aligned} \quad (\text{D.13})$$

in which the configuration space contains the degrees of freedom are $s_i^n \in \{-1, 1\}$ (with $n = 0, \dots, N/\rho - 1$ and $i = 1 + n\rho, \dots, R + n\rho$) and $t_i^n \in \{-1, 1\}$ (with $n = 0, \dots, N/\rho - 1$ and $i = 1 + (n + 1)\rho, \dots, R + n\rho$). The model is sketched in figure D.1, in which it is possible to appreciate the connection with the original one-dimensional chain. In particular, the interaction mediated by λ controls the strength of the bonds in the auxiliary dimension (labeled by n), so that in the limit $\lambda \rightarrow \infty$ the model describes the original chain, with the obvious identification $s_i^n \rightarrow s_i$ and $t_i^n \rightarrow s_i$. By defining the row variables $\underline{s}^n = \{s_i^n\}_{i=1+n\rho}^{i=R+n\rho}$ and $\underline{t}^n = \{t_i^n\}_{i=1+(n+1)\rho}^{i=R+n\rho}$, one

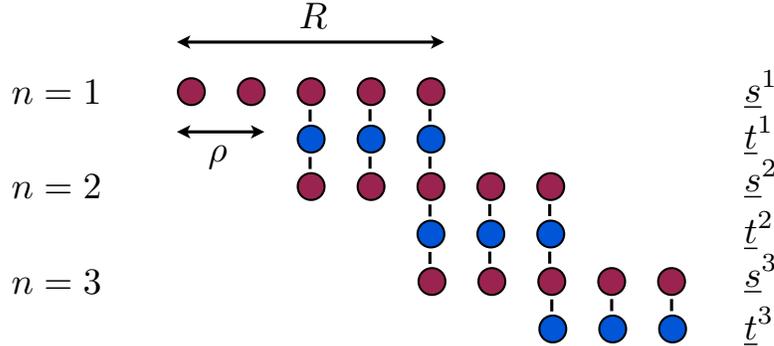


Figure D.1: Two dimensional auxiliary model $p_\lambda(s)$ associated with the original distribution $p(s)$ describing a one-dimensional periodic chain.

can see that the log-probability for the two dimensional model can be written as

$$\log p_\lambda(s, t) = -\log Z_\lambda(\mathbf{g}) - \sum_{n=0}^{N/\rho-1} \left[\mathcal{H}_\lambda^n(\underline{s}^n) + \mathcal{H}_\lambda^{n,n}(\underline{s}^n, \underline{t}^n) + \mathcal{H}_\lambda^{n,n+1}(\underline{t}^n, \underline{s}^{n+1}) \right], \quad (\text{D.14})$$

hence the distribution over the degrees of freedom \underline{s}^n and \underline{t}^n and whose log-probability is given by (D.14) defines a *tree*, because only successive row of variables interact². For such a model, one can straightforwardly generalize the result of appendix D.2 to the case of the non-binary variables \underline{s}^n and \underline{t}^n to show that the full measure $p_\lambda(s, t)$ can be decomposed into the product of the marginals

$$p_\lambda(s, t) = \frac{\prod_n p_\lambda^{\Gamma_n \cup \gamma_n}(\underline{s}^n, \underline{t}^n) p_\lambda^{\gamma_n \cup \Gamma_{n+1}}(\underline{t}^n, \underline{s}^{n+1})}{\prod_n p_\lambda^{\Gamma_n}(\underline{s}^n) p_\lambda^{\gamma_n}(\underline{t}^n)}, \quad (\text{D.15})$$

where Γ_n and γ_n are analogously defined in the case of the two-dimensional model. By taking the $\lambda \rightarrow \infty$ limit, the identification

$$p_\lambda^{\Gamma_n \cup \gamma_n}(\underline{s}^n, \underline{t}^n) \xrightarrow{\lambda \rightarrow \infty} p^{\Gamma_n}(s_{n\rho+1}, \dots, s_{n\rho+R}) \quad (\text{D.16})$$

$$p_\lambda^{\Gamma_n}(\underline{s}^n) \xrightarrow{\lambda \rightarrow \infty} p^{\Gamma_n}(s_{n\rho+1}, \dots, s_{n\rho+R}) \quad (\text{D.17})$$

$$p_\lambda^{\gamma_n}(\underline{t}^n) \xrightarrow{\lambda \rightarrow \infty} p^{\gamma_n}(s_{(n+1)\rho+1}, \dots, s_{n\rho+R}). \quad (\text{D.18})$$

allows to recover the factorization property which had to be proven. \square

²Periodic boundary conditions enforce the presence of a single loop of length N , so that the model is not exactly a tree. Nevertheless, for N large enough and for \mathbf{g} sufficiently distant from critical points of the model, if any, the presence of such loop can be neglected.

Appendix E

Geometry

E.1 Geodesics

We want to find that the condition which a curve $\gamma : [a, b] \in \mathbb{R} \rightarrow \mathcal{M}(\phi)$ has to satisfy order to minimize a functional $\ell(\gamma)$ of the form

$$\ell(\gamma) = \int_a^b dt \sqrt{\chi_{\mu,\nu} \frac{d\gamma_\mu}{dt} \frac{d\gamma_\nu}{dt}} , \quad (\text{E.1})$$

(in which summations on repeated indices are implicit) is given by equation (5.7).

Proof. In order for γ to be a minimum, it needs to extremize the functional $\ell(\gamma)$, so that by constructing the variation $\gamma \rightarrow \gamma + \delta\gamma$ we can impose $\delta\ell(\gamma + \delta\gamma) - \ell(\gamma) = \delta\ell(\gamma) = 0$. This implies

$$\delta\ell(\gamma) = \int_a^b dt \left(\chi_{\mu,\nu} \frac{d\gamma_\mu}{dt} \frac{d\gamma_\nu}{dt} \right)^{-1/2} \left(\frac{1}{2} \partial_\rho \chi_{\mu,\nu} \frac{d\gamma_\mu}{dt} \frac{d\gamma_\nu}{dt} \delta\gamma_\rho + \chi_{\mu,\nu} \frac{d\gamma_\mu}{dt} \frac{d}{dt} \delta\gamma_\nu \right) = 0 . \quad (\text{E.2})$$

By changing variable to

$$dt = \left(\chi_{\mu,\nu} \frac{d\gamma_\mu}{dt} \frac{d\gamma_\nu}{dt} \right)^{-1/2} du , \quad (\text{E.3})$$

one obtains

$$\delta\ell(\gamma) = \int_{u_a}^{u_b} du \left(\frac{1}{2} \partial_\rho \chi_{\mu,\nu} \frac{d\gamma_\mu}{du} \frac{d\gamma_\nu}{du} \delta\gamma_\rho + \chi_{\mu,\nu} \frac{d\gamma_\mu}{du} \frac{d}{du} \delta\gamma_\nu \right) = 0, \quad (\text{E.4})$$

which after integration by parts and some manipulation reads

$$\delta\ell(\gamma) = - \int_{u_a}^{u_b} du \left[\chi_{\mu,\rho} \frac{d^2\gamma_\mu}{du^2} + \frac{1}{2} (\partial_\nu \chi_{\mu,\rho} + \partial_\mu \chi_{\nu,\rho} - \partial_\rho \chi_{\mu,\nu}) \frac{d\gamma_\mu}{du} \frac{d\gamma_\nu}{du} \right] \delta\gamma_\rho = 0. \quad (\text{E.5})$$

Imposing the integrand of above expression to be equal to zero and composing with the inverse Fisher information $\hat{\chi}^{-1}$ yields equation (5.7). \square

E.2 Property of the maximum likelihood estimator

We want to prove that, given a probability density \mathbf{p} defined by a statistical model (ϕ, \mathbf{g}) , for any empirical dataset of length T generated by \mathbf{p} producing empirical averages $\bar{\phi}$, the probability of the maximum likelihood estimator $\mathbf{g}^*(\bar{\phi})$ taking a given value \mathbf{g}' satisfies

$$\lim_{\delta\mathbf{g} \rightarrow 0} \lim_{T \rightarrow \infty} -\frac{1}{T} \log \text{Prob}(\mathbf{g}^*(\bar{\phi}) = \mathbf{g}' + \delta\mathbf{g}) = D_{KL}(\mathbf{p}' || \mathbf{p}), \quad (\text{E.6})$$

being \mathbf{p}' the density associated with the statistical model (ϕ, \mathbf{g}') .

Proof. To prove this relation, we first need to define the set $\mathcal{M}(\phi, \mathbf{g}')$ of probability distributions *compatible* with \mathbf{g}' , defined by

$$\mathcal{M}(\phi, \mathbf{g}') = \left\{ \mathbf{q} \in \mathcal{M}(\Omega) \left| \forall \phi_\mu \in \phi, \sum_s q_s \phi_{\mu,s} = \sum_s \phi_{\mu,s} \exp \left(\sum_{\mu=0}^M g'_\mu \phi_{\mu,s} \right) = \langle \phi_\mu \rangle_{\mathbf{g}'} \right. \right\} \quad (\text{E.7})$$

It can be shown that:

1. $\mathcal{M}(\phi, \mathbf{g}')$ is compact.
2. $\bar{\mathbf{p}} \in \mathcal{M}(\phi, \mathbf{g}')$, if and only if $\mathbf{g}^*(\bar{\phi}) = \mathbf{g}'$.
3. Due to continuity of the functions $\mathbf{g}^*(\bar{\phi})$ and $\bar{\phi}(\bar{\mathbf{p}})$ it holds

$$\lim_{\bar{\mathbf{q}} \rightarrow \bar{\mathbf{p}}} \mathbf{g}^*(\bar{\phi}(\bar{\mathbf{q}})) = \mathbf{g}^*(\bar{\phi}(\bar{\mathbf{p}})) \quad (\text{E.8})$$

Then, Sanov theorem (section 2.2.4) applied to the set $\mathcal{M}(\phi, \mathbf{g})$ implies that

$$\lim_{\delta \rightarrow 0} \lim_{T \rightarrow \infty} -\frac{1}{T} \log \text{Prob}[\bar{\mathbf{p}} \in \mathcal{M}'(\phi, \mathbf{g}')] = D_{KL}(\mathbf{q}^* || \mathbf{p}) \quad (\text{E.9})$$

where $\mathbf{q}^* = \arg \min_{\mathbf{q} \in \mathcal{M}(\phi, \mathbf{g}')} [D_{KL}(\mathbf{q} || \mathbf{p})]$. In order to find the minimum, one can show that for $\mathbf{q} \in \mathcal{M}(\phi, \mathbf{g}')$ it holds

$$D_{KL}(\mathbf{q} || \mathbf{p}) = -S(\mathbf{q}) + F(\mathbf{g}) + \sum_{\mu=1}^M g_{\mu} \langle \phi \rangle_{\mathbf{g}'}, \quad (\text{E.10})$$

where only the term $-S(\mathbf{q})$ depends on the distribution \mathbf{q} . Then the maximum entropy principle (appendix A.1) states that the density $\mathbf{q} \in \mathcal{M}(\phi, \mathbf{g}')$ maximizing $S(\mathbf{q})$ is the statistical model described (ϕ, \mathbf{g}') , whose associated probability density has been called \mathbf{p}' . Then one has

$$\lim_{\delta \rightarrow 0} \lim_{T \rightarrow \infty} -\frac{1}{T} \log \text{Prob}[\bar{\mathbf{p}} \in \mathcal{M}'(\phi, \mathbf{g}')] = D_{KL}(\mathbf{p}' || \mathbf{p}) \quad (\text{E.11})$$

finally, by using property 2. of $\mathcal{M}(\phi, \mathbf{g}')$ and the continuity property 3. one has that for δ sufficiently small, $\text{Prob}[\bar{\mathbf{p}} \in \mathcal{M}'(\phi, \mathbf{g}')] is arbitrarily close to $\text{Prob}[\mathbf{g}^*(\bar{\phi}) - \mathbf{g}' \in \delta \mathbf{g}]$, which together with (E.11) proves the thesis (5.11). $\square$$

E.3 Expansion of the Kullback-Leibler divergence

We want to prove that, given a pair of statistical models (ϕ, \mathbf{g}) and (ϕ, \mathbf{g}') and an accuracy parameter ϵ , for large T they are indistinguishable if condition (5.14) holds.

Proof. If \mathbf{g} and \mathbf{g}' are indistinguishable, then corollary (5.11) implies that, for large T , $D_{KL}(\mathbf{p}'||\mathbf{p}) \leq \epsilon/T$. As $D_{KL}(\mathbf{p}'||\mathbf{p}) = 0 \Leftrightarrow \mathbf{p} = \mathbf{p}'$, one can expand $D_{KL}(\mathbf{p}'||\mathbf{p})$ around the point $\mathbf{g}' = \mathbf{g}$, obtaining

$$D_{KL}(\mathbf{p}'||\mathbf{p}) \approx D_{KL}(\mathbf{p}||\mathbf{p}) + \sum_{\mu=1}^M \frac{\partial D_{KL}(\mathbf{p}'||\mathbf{p})}{\partial g'_\mu} \Bigg|_{\mathbf{g}'=\mathbf{g}} (g'_\mu - g_\mu) \quad (\text{E.12})$$

$$+ \frac{1}{2} \sum_{\mu,\nu=1}^M \frac{\partial^2 D_{KL}(\mathbf{p}'||\mathbf{p})}{\partial g'_\mu \partial g'_\nu} \Bigg|_{\mathbf{g}'=\mathbf{g}} (g'_\mu - g_\mu)(g'_\nu - g_\nu). \quad (\text{E.13})$$

It is easy to see that for the probability distributions \mathbf{p} and \mathbf{p}' associated respectively to \mathbf{g} and \mathbf{g}' it holds equation (2.14), which reads

$$D_{KL}(\mathbf{p}'||\mathbf{p}) = F(\mathbf{g}') - F(\mathbf{g}) + \sum_{\mu=1}^M (g'_\mu - g_\mu) \langle \phi_\mu \rangle_{\mathbf{p}'} . \quad (\text{E.14})$$

As equation (E.14) implies $D_{KL}(\mathbf{p}||\mathbf{p}) = 0$, $\partial_\mu D_{KL}(\mathbf{p}'||\mathbf{p})|_{\mathbf{g}'=\mathbf{g}} = 0$ and

$\partial_\nu \partial_\mu D_{KL}(\mathbf{p}'||\mathbf{p})|_{\mathbf{g}'=\mathbf{g}} = \chi_{\mu,\nu}$, equation (5.14) is proven. \square

E.4 Volume of indistinguishability

Given the space $\mathcal{M}(\phi)$ identified by the minimal operator set ϕ , we want to show that the volume of the space of indistinguishable distributions around a point \mathbf{g} is given by equation (5.15), where T is the length of the dataset and $\epsilon > 0$ is the accuracy parameter.

Proof. The volume $\mathcal{V}_{T,\epsilon}(\mathbf{g})$ is given by

$$\mathcal{V}_{T,\epsilon}(\mathbf{g}) = \int_{\mathcal{M}_{ind}} d\mathbf{g} , \quad (\text{E.15})$$

while property (5.14) characterizes the region of indistinguishability \mathcal{M}_{ind} around \mathbf{g} as $\mathcal{M}_{ind} \xrightarrow{T \rightarrow \infty} \{\mathbf{p}' \in \mathcal{M}(\Omega) \mid \frac{1}{2}(\mathbf{g}' - \mathbf{g})^T \hat{\boldsymbol{\chi}}(\mathbf{g}' - \mathbf{g}) \leq \frac{\epsilon}{T}\} \subseteq \mathcal{M}(\phi)$. We also need to require that T large enough in order to neglect the variations of $\boldsymbol{\chi}$ in \mathcal{M}_{ind} , so that we can treat it as constant in \mathbf{g}' . Due to symmetry of $\hat{\boldsymbol{\chi}}$, the components of Fisher information matrix can be decomposed as $\chi_{\mu,\nu} = \sum_{\lambda=1}^M u_{\mu,\lambda} \chi_{\lambda} u_{\nu,\lambda}$, while due to minimality of ϕ the eigenvalues χ_{λ} are strictly positive, suggesting the change of coordinates

$$\eta_{\lambda} = \sum_{\mu=1}^M (g'_{\mu} - g_{\mu}) u_{\mu,\lambda} \sqrt{\chi_{\lambda}} . \quad (\text{E.16})$$

Then the region \mathcal{M}_{ind} is mapped into the spherical region

$\mathcal{M}_{ind} = \{\mathbf{p}' \in \mathcal{M}(\Omega) \mid \frac{1}{2}\boldsymbol{\eta}^T \boldsymbol{\eta} \leq \frac{\epsilon}{T}\}$ so that the volume becomes

$$\mathcal{V}_{T,\epsilon}(\mathbf{g}) = \frac{1}{\sqrt{\det \hat{\boldsymbol{\chi}}}} \int_{\mathcal{M}_{ind}} d\boldsymbol{\eta} , \quad (\text{E.17})$$

where $1/\sqrt{\det \hat{\boldsymbol{\chi}}} \neq 0$ is the Jacobian of transformation (E.16). It is then sufficient to remind that the volume of a sphere of radius $\sqrt{\frac{2\epsilon}{T}}$ in M dimensions is given by

$$\int_{\mathcal{M}_{ind}} d\boldsymbol{\eta} = \left(\frac{\pi^{\frac{M}{2}}}{\Gamma(\frac{M}{2} + 1)} \right) \left(\frac{2\epsilon}{T} \right)^{\frac{M}{2}} \quad (\text{E.18})$$

to prove equation (5.15). □

E.5 Estimation of the empirical observables for an Hawkes point process

Consider a fully connected Hawkes process defined as in (5.39), characterized by exogenous intensity μ and kernel parameters parameters α and β . We will show that the qualitative features of the fully connected pairwise model associated through the binning functions (5.47) and (5.48) can be obtained by using an approximate scheme. More precisely, given a realization of a fully connected Hawkes point-process \mathbf{X} and a bin size $\delta\tau$ we will calculate the quantities

$$m_i = \frac{1}{T} \sum_{t=1}^T s_i^{(t)}(\mathbf{X}, \delta\tau) \quad (\text{E.19})$$

$$\delta c_{ij} = \frac{N}{T} \left[\left(\sum_{t=1}^T s_i^{(t)}(\mathbf{X}, \delta\tau) s_j^{(t)}(\mathbf{X}, \delta\tau) \right) - m_i m_j \right]. \quad (\text{E.20})$$

First, one can easily notice (expanding the minimum inside the binning functions) that any correlation function of the quantities $b_i^{(t)}$ and $s_i^{(t)}$ can be linked to the properties the Hawkes processes under convolution. In particular, one has for the first two momenta

$$\mathbb{E}[b_i^{(t)}(\mathbf{X}, \delta\tau)] = f_i(\delta\tau) \quad (\text{E.21})$$

$$\mathbb{E}[b_i^{(t)}(\mathbf{X}, \delta\tau) b_j^{(t)}(\mathbf{X}, \delta\tau)] = f_i(\delta\tau) + f_j(\delta\tau) - f_{i+j}(\delta\tau), \quad (\text{E.22})$$

where $f_i(\delta\tau)$ is the average number of events of type i during time $\delta\tau$ in the stationary state, while $f_{i+j}(\delta\tau)$ is the average number of events of type i or j , which is associated with the convolution $X_{i+j} = X_i + X_j$. Thus, to calculate the quantities (E.19) and

(E.20) one needs to calculate

$$\begin{aligned}\mathbb{E}[b_c^{(t)}] &= \sum_{K=0}^{\infty} \text{Prob}[\delta X_c(t \delta \tau) \geq 1, \delta X_{\setminus c}(t \delta \tau) = K] \\ &= 1 - \sum_{K=0}^{\infty} \text{Prob}[\delta X_c(t \delta \tau) = 0, \delta X_{\setminus c}(t \delta \tau) = K],\end{aligned}\quad (\text{E.23})$$

where $\delta X_c(t \delta \tau) = X_c(\delta \tau(t + 1)) - X_c(\delta \tau t)$, while $c \in \{i, j, i + j\}$ and $\setminus c \in \{V \setminus \{i\}, V \setminus \{j\}, V \setminus \{i + j\}\}$ refer to the channels which one needs to take into account to calculate magnetizations and correlations. Above probability can be computed by taking into account that:

- The convolution of a set of Hawkes processes is a Hawkes process.
- Probability (E.23) can be reduced via convolution to the probability of a 2-variate Hawkes processes describing channel c and the environment $\setminus c$.

The parameter set describing such convolution is given for $c = i$ by $\boldsymbol{\mu} = \mu(1, N - 1)$, β unchanged and

$$\hat{\boldsymbol{\alpha}} = \alpha \begin{pmatrix} 0 & 1 \\ N - 1 & N - 2 \end{pmatrix}.\quad (\text{E.24})$$

The one describing the case $c = i + j$ has $\boldsymbol{\mu} = \mu(2, N - 2)$, β unchanged and

$$\hat{\boldsymbol{\alpha}} = \alpha \begin{pmatrix} 1 & 2 \\ N - 2 & N - 3 \end{pmatrix}.\quad (\text{E.25})$$

With this in mind, one can expand probability (E.23) in term of the intensities and obtain

$$\mathbb{E}[b_c^{(t)}] = 1 - \sum_{K=0}^{\infty} \frac{1}{K!} \int_0^{\delta \tau} d\tau_K \dots \int_0^{\delta \tau} d\tau_1 e^{-\int_0^{\delta \tau} du \lambda_c(u) + \lambda_c(u)} \lambda_{\setminus c}(\tau_k) \prod_{k=0}^K \lambda_{\tau_k}^{\setminus c}.\quad (\text{E.26})$$

In principle, one should plug the initial conditions in the stochastic intensities $\lambda(\tau)$ inside previous formula and compute the integral. For example, if one supposes the initial intensities to correspond to the stationary state intensities, one should insert into (E.26) the following expression

$$\lambda_c(\tau) = \mu_c + e^{-\beta\tau}(\bar{\lambda}_c - \mu_c) + \sum_{k=1}^K \alpha_{c,\setminus c} e^{-\beta(\tau-\tau_k)}\theta(\tau - \tau_k) \quad (\text{E.27})$$

$$\lambda_{\setminus c}(\tau) = \mu_{\setminus c} + e^{-\beta\tau}(\bar{\lambda}_{\setminus c} - \mu_{\setminus c}) + \sum_{k=1}^K \alpha_{\setminus c,\setminus c} e^{-\beta(\tau-\tau_k)}\theta(\tau - \tau_k) \quad (\text{E.28})$$

and perform explicitly the integral. This is very hard to do analytically, so that we consider an approximate scheme in which it is possible to obtain a qualitatively correct result for the averages, motivated by the fact that in both the cases that we consider ($c = i$ and $c = i + j$) we have that $\alpha_{\setminus c,c}, \alpha_{\setminus c,c} \gg \alpha_{c,c}, \alpha_{c,\setminus c}$ and $\mu_{\setminus c} \gg \mu_c$. This regime justifies the approximation in which the trajectory $\lambda_{\setminus c}(\tau)$ is described by the *deterministic* function

$$\lambda_{\setminus c}(\tau) = L_{\setminus c}^0 \psi(\delta t) + L_{\setminus c} [1 - \psi(\delta t)] , \quad (\text{E.29})$$

where

$$L_{\setminus c}(\tau)^0 = \left[\left(\boldsymbol{\delta} - \frac{\boldsymbol{\alpha}}{\beta} \right)^{-1} \right]_{\setminus c,c} \mu_c + \left[\left(\boldsymbol{\delta} - \frac{\boldsymbol{\alpha}}{\beta} \right)^{-1} \right]_{\setminus c,\setminus c} \mu_{\setminus c} \quad (\text{E.30})$$

is the average intensity of channel $\setminus c$ in the stationary state in which channel c is free to produce events, while

$$L_{\setminus c} = \left(1 - \frac{\alpha_{\setminus c,\setminus c}}{\beta} \right)^{-1} \mu_{\setminus c} \quad (\text{E.31})$$

is the average in the stationary state in which channel c is conditioned in order not to produce events. Finally $\psi(\delta\tau)$ is a generic function such that $\psi(0) = 1$ and $\psi(\infty) = 0$. Then, one can insert this approximation into equation (E.26), supposing that the number of events K is deterministic and concentrated around its average number.

Then we have

$$\mathbb{E}[b_c(\delta t)] = 1 - \sum_k \text{Prob}[\delta X_c(\delta\tau) = 0, \delta X_{\setminus c}(\delta\tau) = K] \quad (\text{E.32})$$

$$\approx 1 - e^{-\int du [L_c^0 \psi(u) + L_{\setminus c}(1-\psi(u))]} . \quad (\text{E.33})$$

If for example we suppose that $\psi(\tau) = e^{-\beta\tau}$, so that the relaxation dynamics for the intensity is ruled by the same parameter β controlling the dynamics, we get

$$\begin{aligned} \mathbb{E}[b_i^{(t)}] &\xrightarrow{N \rightarrow \infty} 1 - \exp\left(-\frac{\mu \delta t}{1 - \alpha/\beta}\right) & (\text{E.34}) \\ \mathbb{E}[b_j^{(t)}] &\xrightarrow{N \rightarrow \infty} 1 - \exp\left(-\frac{2\mu \delta t}{1 - \alpha/\beta}\right) \\ N \left(\mathbb{E}[b_i^{(t)} b_j^{(t)}] - \mathbb{E}[b_i^{(t)}] \mathbb{E}[b_j^{(t)}] \right) &\xrightarrow{N \rightarrow \infty} \frac{2\alpha\mu e^{-2\mu\delta t/(1-\alpha/\beta)} [e^{-\beta\delta t} - 1 + \beta\delta t]}{(\alpha - \beta)^2} . \end{aligned}$$

This information can be exploited to compute m and δc , which after using the rule $s_c^{(t)} = 2b_c^{(t)} - 1$ result

$$m = 1 - 2e^{-\mu\delta t/(1-\alpha/\beta)} \quad (\text{E.35})$$

$$\delta c = \left(\frac{8\alpha\mu e^{-2\mu\delta t/(1-\alpha/\beta)} [e^{-\beta\delta t} - 1 + \beta\delta t]}{(\alpha - \beta)^2} \right) . \quad (\text{E.36})$$

This result provides a simple qualitative picture, whose degree of inaccuracy lies in the choice of the function $\psi(\tau)$, and in the hypothesis that the trajectory of the stochastic intensity concentrates around a deterministic function. Nevertheless, this approximation captures some of the features that we find by computing magnetization and correlations for various realizations of Hawkes processes for various bin sizes, as shown in figure E.1. Notice in particular the qualitative features of the model correctly reproduced in this scheme, namely (i) correlations drop to zero for small bin sizes (Epps effect) or values $\delta\tau$ larger than the average inter-event time, (ii) correlations increase with the interaction parameter α and are zero for the Poisson case $\alpha = 0$. The

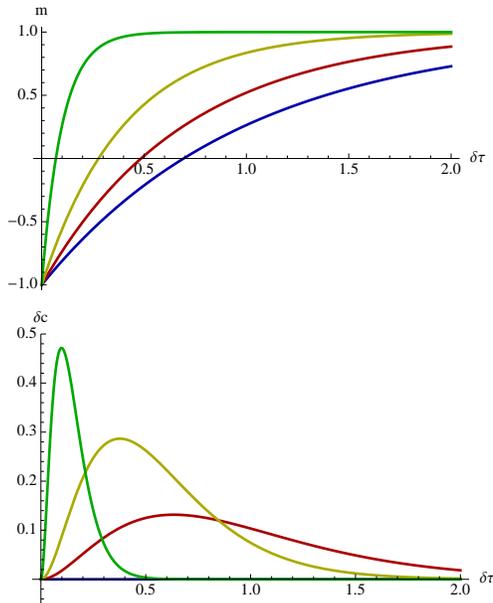


Figure E.1: Approximate values of average magnetization m and rescaled connected correlation δc associated with a fully-connected pairwise model used to describe a fully-connected Hawkes process. We consider in particular models for which $\mu = 1$, $\beta = 2$ and $\alpha = 0, 0.3, 0.6, 0.9$ (respectively blue, red, yellow and green line). The corrected qualitative features of the model are captured in this approximate scheme.

magnetizations calculated in this way correspond instead to the exact value. In figure E.2 we plot the ensemble averages of the model and the average inferred couplings (h^*, J^*) in the case of a fully connected pairwise model for various choices of the bin size and of the interaction parameter α . Finally, notice that this approximation is able to capture the finiteness of δc , which implies that the description of data in term of a fully connected ferromagnet doesn't lead to a degenerate representation of the model (see section 3.3).

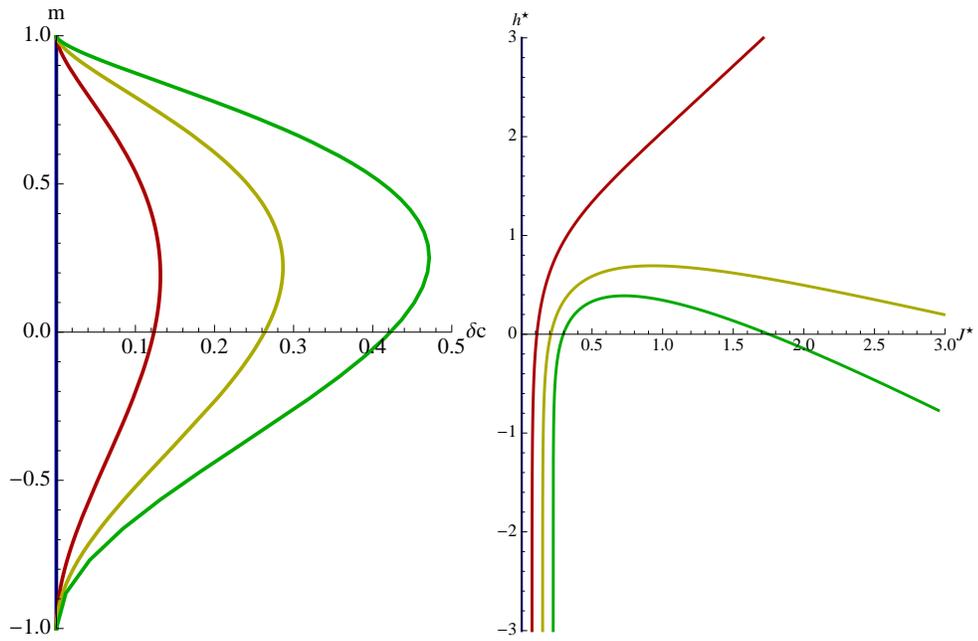


Figure E.2: Approximate values of the empirical averages (m, c) and of the inferred couplings (h^*, J^*) obtained by using a fully connected pairwise model to fit a set of Hawkes point process, for the same choice of models and color conventions as in the previous plot, parametrically plotted as a function of the bin size $\delta\tau$.

Bibliography

- [1] Findings regarding the market events of may 6, 2010. Tech. rep., U.S. Commodity Futures Trading Commission and the U.S. Securities and Exchange Commission.
- [2] Google translate. <http://translate.google.com>.
- [3] Kaggle official website. <http://www.kaggle.com>.
- [4] Netflix official website. <http://www.netflixprize.com/>.
- [5] Yahoo finance. <http://yahoo.finance.com>.
- [6] ACKLEY, D., HINTON, G., AND SEJNOWSKI, T. A learning algorithm for boltzmann machines. *Cogn. Sci.* 9 (1985), 147.
- [7] AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory (1973)*, vol. 1, Springer Verlag, pp. 267–281.
- [8] ALMEIDA, J., AND THOULESS, D. Stability of the sherrington-kirkpatrick solution of a spin glass model. *J. Phys. A-Math. Gen.* 11 (1978), 983–990.
- [9] AMARI, S. *Differential Geometrical Methods In Statistics*. Springer, 1985.
- [10] AMARI, S., BARNDORFF-NIELSEN, O., KASS, R., LAURITZEN, S., AND RAO, C. *Differential Geometry in Statistical Inference*. Institute of Mathematical Statistics, 1987.
- [11] AMARI, S., AND NAGAOKA, H. *Methods of information geometry*, vol. 191. American Mathematical Society, 2007.
- [12] AURELL, E., AND EKEBERG, M. Inverse ising inference using all the data. *Phys. Rev. Lett.* 108, 9 (2012), 090201.
- [13] BALASUBRAMANIAN, V. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005, ch. MDL, Bayesian inference, and the geometry of the space of probability distributions.
- [14] BAUWENS, L., AND HAUTSCH, N. *Handbook of Financial Time Series*. Springer, 2009, ch. Modelling financial high frequency data using point processes, pp. 953–979.

- [15] BAXTER, R. *Exactly solved models in statistical mechanics*. Academic Press, 1982.
- [16] BOUCHAUD, J. Economics needs a scientific revolution. *Nature* 455, 7217 (2008), 1181.
- [17] BOUCHAUD, J., FARMER, J. D., AND LILLO, F. *Handbook of Financial Markets: Dynamics and Evolution*. Elsevier, 2008, ch. How Markets Slowly Digest Changes in Supply and Demand, pp. 57–156.
- [18] BOUCHAUD, J., AND POTTERS, M. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.
- [19] BOWSHER, C. Modelling security market events in continuous time: Intensity based, multivariate point process models. Tech. Rep. 2002-W22, Nuffield College, Oxford, 2002.
- [20] BOYD, S. Subgradient methods. *Lecture notes*, <http://www.stanford.edu/class/ee364b/lectures.html> (2010).
- [21] BOYD, S., AND VANDENBERGHE, L. *Convex optimization*. Cambridge University Press, 2004.
- [22] BRAUNSTEIN, A., PAGNANI, A., WEIGT, M., AND ZECCHINA, R. Inference algorithms for gene networks: a statistical mechanics analysis. *J. Stat. Mech.* 2008, 12 (2008), P12001.
- [23] BYRD, R., LU, P., NOCEDAL, J., AND ZHU, C. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comp.* 16, 5 (1995), 1190–1208.
- [24] COCCO, S., LEIBLER, S., AND MONASSON, R. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc. Natl. Acad. Sci. U.S.A.* 106 (2009), 14058.
- [25] COCCO, S., AND MONASSON, R. Adaptive cluster expansion for inferring boltzmann machines with noisy data. *Phys. Rev. Lett.* 106, 9 (2011), 090601.
- [26] COCCO, S., AND MONASSON, R. Adaptive cluster expansion for the inverse ising problem: Convergence, algorithm and tests. *J. Stat. Phys.* 147, 2 (2012), 252–314.
- [27] COVER, T., THOMAS, J., WILEY, J., ET AL. *Elements of information theory*, vol. 6. Wiley Online Library, 1991.
- [28] DACOROGNA, M., GENÇLAY, R., MÜLLER, U., OLSEN, R., AND PICTET, O. *An Introduction to High-Frequency Finance*. Academic Press, 2001.

-
- [29] DE LACHAPELLE, D., AND CHALLET, D. Turnover, account value and diversification of real traders: evidence of collective portfolio optimizing behavior. *New J. Phys.* 12 (2010), 075039.
- [30] DE MARTINO, A., AND MARSILI, M. Statistical mechanics of socio-economic systems with heterogeneous agents. *J. Phys. A-Math. Gen.* 39 (2006), R465.
- [31] DONOHO, D. Compressed sensing. *IEEE T. Inform Theor.* 52, 4 (2006), 1289–1306.
- [32] EPPS, T. Comovements in stock prices in the very short run. *J. Amer. Stat. Ass.* 74 (1979), 291–298.
- [33] FAMA, E. Efficient capital markets: A review of theory and empirical work. *J. Financ.* 25, 2 (1970), 383–417.
- [34] FELLER, W. *An introduction to probability theory and its applications*. John Wiley & Sons, 1950.
- [35] GORI, G., AND TROMBETTONI, A. The inverse ising problem for one-dimensional chains with arbitrary finite-range couplings. *J. Stat. Mech.* 2011 (2011), P10021.
- [36] HASTINGS, W. Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.
- [37] HAWKES, A. Point spectra of some mutually exciting point processes. *J. R. Statist. Soc. B* 33 (1971), 438–443.
- [38] HAWKES, A. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 1 (1971), 83–90.
- [39] HINTON, G. A practical guide to training restricted boltzmann machines. Tech. rep., Univ. Toronto, 2010.
- [40] HUANG, K. *Statistical Mechanics*. John Wiley & Sons, 1987.
- [41] ISING, E. Beitrag zur theorie des ferromagnetismus. *Z. Phys. A-Hadron. Nucl.* 31, 1 (1925), 253–258.
- [42] JAEGER, F., VERTIGAN, D., AND WELSH, D. On the computational complexity of the jones and tutte polynomials. *Math. Proc. Cambridge* 108, 01 (1990), 35–53.
- [43] JERRUM, M., AND SINCLAIR, A. Polynomial-time approximation algorithms for the ising model. *Lect. Notes. Comput. Sc.* (1990), 462–475.
- [44] JOULIN, A., LEFEVRE, A., GRUNBERG, D., AND BOUCHAUD, J. Stock price jumps: news and volume play a minor role. *Arxiv preprint arxiv:0803.1769* (2008).

-
- [45] KAPPEN, H., AND RODRIGUEZ, F. Efficient learning in boltzmann machines using linear response theory. *Neural. Comput.* 10 (1998), 1137–1156.
- [46] KIRMAN, A. *Complex economics: individual and collective rationality*. Routledge, 2010.
- [47] KRAUTH, W. Introduction to monte carlo algorithms. *Lect. Notes. Phys.* (1998), 1–35.
- [48] LILLO, F., MORO, E., VAGLICA, G., AND MANTEGNA, N. Specialization and herding behavior of trading firms in a financial market. *New J. Phys.* 10 (2008), 043019.
- [49] LIU, D., AND NOCEDAL, J. On the limited memory bfgs method for large scale optimization. *Math. Program.* 45, 1 (1989), 503–528.
- [50] MACKAY, D. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.
- [51] MANTEGNA, N., AND STANLEY, E. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, 1999.
- [52] MARINARI, E., AND VAN KERREBROECK, V. Intrinsic limitations of the susceptibility propagation inverse inference for the mean field ising spin glass. *J. Stat. Mech.* 2010 (2010), P02008.
- [53] MASTROMATTEO, I., AND MARSILI, M. On the criticality of inferred models. *J. Stat. Mech.* 2011 (2011), P10012.
- [54] MÉZARD, M., AND MONTANARI, A. *Information, Physics and Computation*. Oxford University Press, 2009.
- [55] MÉZARD, M., AND MORA, T. Constraint satisfaction problems and neural networks: a statistical physics perspective. *J. Physiol. Paris* 103 (2009), 107–113.
- [56] MÉZARD, M., PARISI, G., AND VIRASORO, M. *Spin glass theory and beyond*. World scientific Singapore, 1987.
- [57] MONASSON, R. The mean-field ising model. *Lecture notes*, <http://www.phys.ens.fr/~monasson/> (2010).
- [58] MORA, T., AND BIALEK, W. Are biological systems poised at criticality? *J. Stat. Phys.* (2011), 1–35.
- [59] MORO, E., VICENTE, J., MOYANO, L., GERIG, A., FARMER, J. D., VAGLICA, G., LILLO, F., AND MANTEGNA, N. Market impact and trading profile of hidden orders in stock markets. *Phys. Rev. E* 80 (2009), 066102.

-
- [60] MÖRTERS, P. Large deviation theory and applications. *Lecture notes*, <http://people.bath.ac.uk/maspm/> (2008).
- [61] MYUNG, I., AND BALASUBRAMANIAN, V. Counting probability distributions: Differential geometry and model selection. *Proc. Natl. Acad. Sci. U.S.A.* *97* (2000), 11170.
- [62] PLEFKA, T. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *J. Phys. A-Math. Gen.* *15* (1982), 1971–1978.
- [63] RAVIKUMAR, P., WAINWRIGHT, M., AND LAFFERTY, J. High-dimensional ising model selection using ℓ_1 -regularized logistic regression. *Ann. Stat.* *38*, 3 (2010), 1287–1319.
- [64] RICCI-TERSENGHI, F. The bethe approximation for solving the inverse ising problem: a comparison with other inference methods. *J. Stat. Mech.* *2012*, 08 (2012), P08015.
- [65] RISSANEN, J. Universal coding, information, prediction, and estimation. *IEEE T. Inform Theor.* *30*, 4 (1984), 629–636.
- [66] RISSANEN, J. Stochastic complexity and modelling. *Ann. Stat.* *14* (1986), 1080.
- [67] ROUDI, Y., AURELL, E., AND HERTZ, J. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.* *3*, 22 (2009), 1–15.
- [68] ROUDI, Y., TYRCHA, J., AND HERTZ, J. The ising model for neural data: Model quality and approximate methods for extracting functional connectivity. *Phys. Rev. E* *79* (2009), 051915.
- [69] RUAL, J., VENKATESAN, K., HAO, T., HIROZANE-KISHIKAWA, T., DRICOT, A., LI, N., BERRIZ, G., GIBBONS, F., DREZE, M., AYIVI-GUEDEHOUSOU, N., ET AL. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* *437*, 7062 (2005), 1173–1178.
- [70] RUSSELL, S., AND NORVIG, P. *Artificial intelligence: a modern approach*. Prentice Hall, 2010.
- [71] SCHMIDT, M., AND MURPHY, K. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)* (2010).
- [72] SCHNEIDMAN, E., BERRY II, M., SEGEV, R., AND BIALEK, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* *440* (2006), 1007–1012.
- [73] SCHWARZ, G. Estimating the dimension of a model. *Ann. Stat.* *6*, 2 (1978), 461–464.

-
- [74] SESSAK, V., AND MONASSON, R. Small-correlation expansions for the inverse ising problem. *J. Phys. A-Math. Theor.* 42 (2009), 055001.
- [75] SHENDURE, J., AND JI, H. Next-generation dna sequencing. *Nat. Biotechnol.* 26, 10 (2008), 1135–1145.
- [76] SHLENS, J., FIELD, G., GAUTHIER, J., GRIVICH, M., PETRUSCA, D., SHER, A., LITKE, A., AND CHICHILNISKY, E. The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* 26, 32 (2006), 8254–8266.
- [77] SOCOLICH, M., LOCKLESS, S., RUSS, W., LEE, H., GARDNER, K., AND RANGANATHAN, R. Evolutionary information for specifying a protein fold. *Nature* 437 (2005), 512–518.
- [78] STEPHENS, G., MORA, T., TKACIK, G., AND BIALEK, W. Thermodynamics of natural images. *Arxiv preprint arXiv:0806.2694* (2008).
- [79] TANAKA, T. Mean field theory of boltzmann machine learning. *Phys. Rev. E* 58 (1998), 2302.
- [80] TELLER, E., METROPOLIS, N., AND ROSENBLUTH, A. Equation of state calculations by fast computing machines. *J. Chem. Phys* 21, 13 (1953), 1087–1092.
- [81] THOULESS, D., ANDERSON, P., AND PALMER, R. Solution of 'solvable model of a spin glass'. *Philos. Mag.* 35, 3 (1977), 593–601.
- [82] TKACIK, G., SCHNEIDMAN, E., BERRY II, M., AND BIALEK, W. Ising models for networks of real neurons. *Arxiv preprint arXiv:q-bio/0611072v1* (2006).
- [83] TRICHET, J. Reflections on the nature of monetary policy non-standard measures and finance theory.
- [84] TYRCHA, J., ROUDI, Y., MARSILI, M., AND HERTZ, J. Effect of nonstationarity on models inferred from neural data. *Arxiv preprint arXiv:1203.5673* (2012).
- [85] WAINWRIGHT, M., AND JORDAN, M. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1, 1-2 (2008), 1–305.
- [86] WAINWRIGHT, M., RAVIKUMAR, P., AND LAFFERTY, J. High-dimensional graphical model selection using ℓ_1 -regularized logistic regression. *Adv. Neur. In.* 19 (2006), 1465–1472.
- [87] WEIGT, M., WHITE, R., SZURMANT, H., HOCH, J., AND HWA, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* 106 (2009), 67.

- [88] WHEELER, D., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y., MAKHIJANI, V., ROTH, G., ET AL. The complete genome of an individual by massively parallel dna sequencing. *Nature* 452, 7189 (2008), 872–876.
- [89] ZAMPONI, F. Mean field theory of spin glasses. *Arxiv preprint arXiv:1008.4844* (2010).

Notation

List of Symbols

N	System size
Ω	Configuration space
\mathbf{p}	Probability density
ϕ	Operator set
\mathbf{g}	Coupling vector
\mathbf{g}^*	Estimator of a coupling vector \mathbf{g}
$\mathcal{M}(\Omega)$	Space of probability densities in Ω
$\mathcal{M}(\phi)$	Space of statistical models associated to the operator set ϕ
M	Cardinality of the set of operators ϕ
V	Vertex set (the set of spins $\{1, \dots, N\}$)
E	Edge set
Γ	Cluster (a generic subset of V)
\mathbf{p}^Γ	Marginal probability associated to cluster Γ
F	Free energy
$\langle \phi \rangle$	Ensemble averages
$\hat{\chi}$	Generalized susceptibility matrix
S	Shannon entropy
$D_{KL}(\mathbf{p} \mathbf{q})$	Kullback-Leibler divergence between distributions \mathbf{p} and \mathbf{q}
T	Size of the empirical dataset
$\hat{\mathbf{s}}$	Empirical dataset
$\bar{\phi}$	Empirical averages
$\bar{\mathbf{p}}$	Empirical frequencies
$\mathcal{G}(\phi)$	Marginal polytope (set of possible empirical averages) associated to the operator set ϕ
$P_T(\hat{\mathbf{s}} \mathbf{g})$	Likelihood function
$P_T(\mathbf{g} \hat{\mathbf{s}})$	Posterior function
$P_0(\mathbf{g})$	Prior
$\langle \dots \rangle$	Average on the measure defined by \mathbf{p}
$\langle \dots \rangle_T$	Average on the measure defined by $P_T(\hat{\mathbf{s}} \mathbf{g})$

List of Subscripts

i, j, \dots	Spin index
μ, ν, \dots	Operator index
s, s', \dots	Configuration index
t, t', \dots	Sampled configuration index