ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Physics

Approximate inference on graphical models: message-passing, loop-corrected methods and applications

Giovanni Catania
* * * * * *

Supervisor
Prof. Alfredo Braunstein

Doctoral examination committee
Prof. Cristopher Moore, Santa Fe Institute for Complex Systems, New Mexico (USA)
Prof. Federico Ricci Tersenghi, Università Roma La Sapienza, Italy

Politecnico di Torino

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . .

Giovanni Catania
Torino, July 15, 2021

Summary

Probabilistic graphical models provide an unified framework to analyze physical systems of many interacting degrees of freedom, by combining elements of graph and probability theory. Interacting physical systems display a variety of collective phenomena, depending on the mutual dependencies between units and the consequent topological structure of their interactions, whose first analysis led to the foundation of statistical mechanics and the study of phase transitions. A paradigmatic example is given by classic spin models defined on graphs or lattices, historically developed to describe the microscopic origin of magnetism; at the same time, they provide a general description of a wide class of phenomena in several research fields, from biology, neuroscience, computer science and econophysics. In this perspective, the formalism of graphical models provides a common framework and set of methodologies to analyze interacting systems in virtually any field of pure and applied science. In general, the difficulty of analyzing high-dimensional system is that the presence of interactions makes any computation unfeasible in practice, as the volume of the configuration space grows exponentially with the system size. In this sense, the term *approximate inference* in the manuscript's title refers to the generic problem of estimating relevant features of a probabilistic graphical model, such as its marginal distributions.

The main goal of this thesis is to introduce a novel class of approximation schemes to estimate marginal probabilities on discrete (spin) models, called Density Consistency. This method shares similarities with other message-passing schemes commonly employed in statistical physics and inference, such as Belief Propagation and Expectation Propagation. The novelty introduced by Density Consistency relies on a simple way to encode approximate loop corrections coming from all the cycles in the graph, by exploiting a refined Gaussian approximation.

The structure of the manuscript resembles the research path I carried out during my three years of PhD, and it is divided in two parts. After a brief introduction presented in the first chapter, Part I focuses on the Density Consistency method: in particular, I discuss a generic derivation on probabilistic graphical models of binary degrees of freedom, using the factor graph representation. Its properties and its relations to other advanced mean field methods are discussed, and its performances are evaluated on finite size systems. Furthermore, I present an analytic theory for the ferromagnetic Ising model in the thermodynamic limit, providing a closed form expression for the critical temperature. Finally, Density Consistency is applied to the Inverse Ising problem in statistical inference and its performances are compared to other state-of-the-art techniques.

Part II of the manuscript contains a standalone chapter, in which I discuss another project I contributed to during my last year of PhD, somehow prompted by the COVID-19 pandemic. By using message-passing techniques in a Bayesian inference framework, we developed an on-line epidemic mitigation protocol in order to detect the individuals with the highest risk to be infected, starting from the knowledge about their contacts, that can be registered using digital contact tracing applications. Probabilistic inference provide a criterion to selectively isolate individuals with the highest risk, so that an effective epidemic suppression can be achieved while avoiding global containment measures, with consequent and well-known economic and social drawbacks.

This project has been carried out jointly with another group of researchers based in Paris (France) and Lausanne (Switzerland).

I summarize below the contents of each chapter and the list of co-authored papers, all of them being covered in the manuscript.

Thesis Outline

Chapter 1: **Introduction**
I present some basic notions of equilibrium statistical physics, graph theory and statistical inference, in order to provide a common background knowledge and notation used through the whole manuscript.

Part I **Loop corrections in spin models through Density Consistency**

Chapter 2: **Approximate methods in statistical physics**
I discuss several state-of-the-art approximations in statistical physics and high-dimensional inference, with a special focus on message-passing techniques such as Belief Propagation. An additional detailed derivation of Expectation Propagation is presented, that is required to better understand Density Consistency.

Chapter 3: **Density Consistency**
This is the core chapter of the manuscript, where Density Consistency is derived on graphical models of binary spins and its main properties are discussed. The contents of this chapter are included in Paper A.

Chapter 4: **Results: forward problem**
I evaluate the performances of Density Consistency in comparison to other approximations. In particular, the first part presents a series of results on finite size systems, while in the second DC is used to derive a quasi-analytic solution for the ferromagnetic Ising model in the thermodynamic limit. The contents of this chapter are included in Paper A.

Chapter 5: **The Inverse Ising problem**
This chapter focuses on the Inverse Ising Problem in statistical physics. After presenting some background motivations and formulating the problem in a Bayesian framework, I derive an approximate solution for the maximum likelihood estimator using Density Consistency, whose performances on synthetic data are compared with other state-of-the-art techniques. The contents of this chapter are included in Paper B.

Chapter 6: **Conclusions and future perspectives**
I summarize the main findings obtained so far on Density Consistency. In addition, I describe some future directions and possible applications of Density Consistency to be yet investigated.

Part II **Bayesian inference approaches for epidemic mitigation**

Chapter 7: **Bayesian inference approaches for epidemic mitigation**
This chapter describes a Bayesian-inference guided mitigation protocol for epidemic spreading processes from contact tracing data. The main method used to perform the approximate inference is Belief Propopagation; results are validated on top of a compartmental model designed to describe a realistic spreading of SARS-CoV-2 in a population. The contents of this chapter are related to Paper C.

Coauthored papers and pre-prints

Paper A Alfredo Braunstein, Giovanni Catania and Luca Dall'Asta,
Loop corrections in spin models through density consistency,
Physical Review Letters 123, 020604 (2019)
arXiv:1810.10602.

Paper B Alfredo Braunstein, Giovanni Catania, Luca Dall'Asta and Anna Paola Muntoni,
A Density Consistency approach to the inverse Ising problem,
Journal of Statistical Mechanics: theory and experiment, 033416 (2021)
arXiv:2010.13746

Paper C Antoine Baker, Indaco Biazzo, Alfredo Braunstein, Giovanni Catania, Luca Dall'Asta, Alessandro Ingrosso, Florent Krzakala, Fabio Mazza, Marc Mézard, Anna Paola Muntoni, Maria Refinetti, Stefano Sarao Mannelli and Lenka Zdeborová,
Epidemic mitigation by statistical inference from contact tracing data,
arXiv:2009.09422, accepted on Proceedings of the National Academy of Sciences

Contents

Chapter 1

Introduction

This opening chapter reviews some basic concepts in statistical physics and its connection to inference problems, that will be used in the rest of the thesis. In addition, some elementary notions about graph theory will be recalled in Sec. 1.2, with a particular attention to the factor graph representation that will be needed in Chapters 2-3. Finally, a very brief introduction to Bayesian inference is presented in Section 1.3.

1.1 Statistical Physics and thermodynamics

Statistical physics was born at the end of 19^{th} century as an attempt to provide a microscopic interpretation of classical laws of thermodynamics. Thermodynamics deals with global properties of physical systems, that can be described in terms of few relevant macroscopic quantities. A macroscopic physical system is characterized by an enormous ($\sim N_A = 10^{23}$ where N_A is Avogadro's constant) number of *interacting* degrees of freedom, so that deterministic laws of classic mechanics are unfeasible to describe the behaviour of each microscopic object. In this perspective, statistical physics relies on a probabilistic approach where each microscopic configuration is associated to a certain probability to be observed under suitable external conditions; in this way the macroscopic behaviour is obtained through a statistical, averaged description of the phenomena occurring at the microscopic level.
Historically, the first microscopic model developed is the one of the ideal gas, where macroscopic quantities like pressure or internal energy can be derived from the collective microscopic motion of each gas particle that can freely move inside a box with fixed volume. In this case, the assumption that degrees of freedom are independent makes the computation of thermodynamic variables easy: however, it fails to capture the correct behaviour of the gas at low temperature (for instance, when a phase transition to the liquid occurs upon lowering the external temperature). The presence of interactions make most statistical physics models not solvable exactly, so that approximation methods are needed to estimate their collective behaviour. In the rest of the manuscript, we will deal with classical statistical physics, where the degrees of freedom are real (or eventually discrete) variables: this means that the quantum nature of the microscopic degrees of freedom (described by wave-functions in infinite-dimensional Hilbert spaces) is not taken into account. The whole information about the microscopic interactions between the degrees of freedom is encoded in the

Hamiltonian H, that can be generically written as a sum of k-body functional interactions:

$$H\left(x_1, \ldots x_N\right) = \sum_{k=1}^{N} H^{(k)}\left(x_1, \ldots, x_N\right) \tag{1.1}$$

$$H^{(k)}\left(x_1, \ldots, x_N\right) = \sum_{i_1, \ldots i_k} \tilde{J}_{i_1, \ldots i_k} \phi^{(k)}_{i_1, \ldots, i_k}\left(x_{i_1}, \ldots, x_{i_k}\right) \tag{1.2}$$

where each variable x_i denotes a degree of freedom (they might be canonical coordinates or momenta, rotational or vibrational degrees of freedom associated to complex molecules, magnet dipoles, and so on). In most cases, one assumes that a certain physical system described by a Hamiltonian of the type (1.1) is in equilibrium with a *heat bath* (also called reservoir) at a fixed temperature T. Therefore, microscopic degrees of freedom are allowed to exchange energy with the bath, that is assumed to be un-modified from these interactions (this is consistent with an assumption of an infinitely large reservoir). These settings define the so-called *canonical ensemble*. The probability distribution describing the probability of observing each microscopic configuration in equilibrium with the reservoir is expressed by the Boltzmann law:

$$p\left(x_1, \ldots, x_N\right) = \frac{1}{Z} e^{-\beta H\left(x_1, \ldots, x_N\right)} \tag{1.3}$$

where the quantity $\beta \hat{=} 1/k_B T$ is called *inverse temperature*. The above expression has a very simple interpretation: configurations with low energy are more likely to be observed with respect to high-energy ones, and states with same energy have the same probability to be observed. Moreover, the ratio between the probability of high-energy configuration w.r.t. low-energy ones depends on the temperature T, being an exponential decreasing function w.r.t. β. In this perspective, two extreme cases can be distinguished: at infinite temperature ($\beta \to 0$), the Boltzmann's law (1.3) becomes a uniform measure over all the configurations; on the other hand, in the zero-temperature limit ($\beta \to \infty$), the equilibrium distribution becomes peaked over configurations with minimum energy (*ground states*), all the others having a null measure. The constant Z, called *partition function*, ensures the correct normalization of (1.3) and it can be computed by integrating over all the possible microscopic configurations:

$$Z = \int dx_1 \ldots dx_N \exp\left[-\beta H\left(x_1, \ldots, x_N\right)\right] \tag{1.4}$$

The full information about the macroscopic behaviour of the physical system is encoded into the partition function (1.4). Apart from a constant prefactor, its logarithm defines the Helmoltz free energy, that is equivalent to the one defined in standard thermodynamics:

$$F \hat{=} -\frac{1}{\beta} \log Z \equiv U - TS \tag{1.5}$$

where U is the internal energy and S is the entropy. It is easy to show that these quantities can be expressed as suitable expectation values over the Boltzmann measure (1.3):

$$U = \int d\boldsymbol{x} p\left(\boldsymbol{x}\right) H\left(\boldsymbol{x}\right) \tag{1.6}$$

$$S = -k_B \int d\boldsymbol{x} p\left(\boldsymbol{x}\right) \log p\left(\boldsymbol{x}\right) \tag{1.7}$$

where (1.7) is the expression derived by Shannon in information theory [38, 125]. The above distribution describes an ensemble of particles interacting through the Hamiltonian H, subject to thermal flucutations at inverse temperature β. The analogy with information theory allows to derive the canonical ensemble distribution from the maximum-entropy principle [38]. The main goal of statistical physics is to provide a description of the macroscopic physical system by evaluating its free energy, from which any physical quantity can be computed. For instance, the average energy and entropy can be obtained by:

$$U = \frac{\partial \left(\beta F\right)}{\partial \beta} \qquad S = \frac{\partial F}{\partial T} \tag{1.8}$$

There are only few examples in which an exact evaluation of F can be performed, as in general the configuration space grows exponentially with the system size. The simplest case occurs for non-interacting models, i.e. where the Hamiltonian contains only $1-$body terms and the Boltzmann law (1.3) factorizes over single-node marginals, so that an exact computation can be carried out (e.g. in the ideal gas). The interesting cases where degrees of freedom interact under suitable potentials are intractable in almost all cases. Interacting models give rise to a variety of collective phenomena, that can be described mathematically by evaluating the free energy F in the *thermodynamic limit*, i.e. when the number of degrees of freedom goes to infinity. The existence of collective phenomena determines different macroscopic behaviours, corresponding to changes in the free energy under external conditions. The point(s) at which the system changes its global behaviour by a small perturbation of the control parameter (e.g. temperature) define a *phase transition*: mathematically, they correspond to points at which the free energy shows a non-analytic behaviour in the thermodynamic limit.

1.1.1 Discrete models and binary spins

In this thesis, we will mainly focus on systems with discrete degrees of freedom, so that each of them takes values on a finite alphabet $\mathcal{X} = \{a_1, \ldots, a_q\}$. The simplest case corresponds to a bimodal support, i.e $|\mathcal{X}| = 2$: without loss of generality, the two states can be taken as symmetric, namely $\sigma_i \in \{-1,1\}$. In this setting, the generic $p-$body functional interactions in (1.1) can be simplified as the product of the p degrees of freedom that participate to the interaction:

$$H\left(\sigma_1, \ldots, \sigma_N\right) = \sum_i \tilde{J}_i \sigma_i + \sum_{i,j} \tilde{J}_{ij} \sigma_i \sigma_j + \sum_{i,j,k} \tilde{J}_{ijk} \sigma_i \sigma_j \sigma_k + \ldots \tag{1.9}$$

Discrete degrees of freedom defined on $\{-1,1\}$ are typically called *binary spins* in statistical physics: indeed they represent the classical counterpart of $\frac{1}{2}-$spin models in quantum mechanics, used to describe the behaviour of the electron's magnetic dipoles in crystal (or disordered) structures. In particular, the scalar degree of freedom σ_i can be related to the projection of the spin-momentum operator along one of the three spatial axis. The simplest - and yet highly non-trivial - example of an interacting model of classic spins can be obtained from (1.9) by keeping only pairwise interactions, and it is known in the literature as the *Ising model*. For convenience, we rewrite its Hamiltonian by specifying the 1-body terms (also called external fields and denoted with h_i) and 2-body couplings, denoted in the following with J_{ij}:

$$H\left(\sigma_1, \ldots, \sigma_N\right) = -\sum_i h_i \sigma_i - \sum_{i,j} J_{ij} \sigma_i \sigma_j \tag{1.10}$$

In a real physical system, the strength of interactions typically depend on the distance between the two spins, that assumed to occupy the sites of a certain topology (e.g. a lattice): as a consequence, interactions can be neglected if the distance between them is large enough; the simplest case corresponds to retaining only the interactions between the "closest" spins (also referred to as nearest neighbours), so that each spin interacts with a small subset of the other variables: as a consequence, not all the J_{ij} are present in (1.10). A more mathematical description of the interaction topology is based on graph theory, discussed in the next section. In general, depending on the sign of the couplings we distinguish three very different scenarios:

- if all the couplings are positive, i.e. $J_{ij} > 0$, the model is *ferromagnetic*
- if all the couplings are negative, i.e. $J_{ij} < 0$ the model is called *antiferromagnetic*
- if couplings are both positive and negative, the model is called *spin glass*

The above categorization can actually be extended to Hamiltonians including high order interactions as in 1.1: for instance, the spin-glass limit where the all possible $k-$spin couplings (up to the N-body term) are considered and all the k-th order interaction tensors are sampled from a Gaussian distribution defines the Random Energy Model, developed by Derrida in [41].
Each coupling in the Hamiltonian defines a "soft" constraint (at $T > 0$) that favours configurations where the product $J_{ij}\sigma_i\sigma_j$ is positive. This is true for any k-body term: for instance, an external field h_i favours configurations where σ_i is aligned to it. In the ferromagnetic case, nearest neighbours spin prefer to be aligned, so that there is a competition between the energetic term and the thermal fluctuations each spin is subject to. Depending on the graph topology, this competition determines a different global behaviour in the thermodynamic limit, controlled by the external temperature T: in particular, below a *critical* value T_c, the energetic term dominates the contribution to the free energy, so that the system is characterized by a global spin ordering (*ferromagnetic phase*); conversely, at high temperatures the thermal fluctuations destroy such ordering, so that the free energy is dominated by configurations with no alignment between nearest-neighobour spins (*paramagnetic phase*). In a homogenous model, i.e. when all the interaction terms in (1.10) are equal, this phase transition can be characterized in terms of a unique order parameter, namely the magnetization m:

$$m\left(\beta\right) = \frac{1}{N}\langle\sum_i \sigma_i\rangle = -\frac{1}{\beta}\frac{\partial F}{\partial h} \tag{1.11}$$

where $\langle\cdot\rangle$ denotes the ensemble average w.r.t. the Boltzmann measure (1.3) with the Ising Hamiltonian (1.10). An abrupt change of the order parameter m determines a ferromagnetic-paramagnetic phase transition: the free energy computed as a function of m shows indeed a different behaviour in the two regimes: at $T > T_c$ it has a unique minimum at $m = 0$; at $T < T_c$, the point $m = 0$ becomes a maximum and two symmetric minima (in the absence of an external field) appear at $m = \pm m_0\left(T\right)$. This behaviour signals a *spontaneous symmetry breaking*, so that in the thermodynamic limit the system is observed in just one of the two minima.
Conversely, with $J_{ij} < 0$ the two spins try to minimize the energy by having oppisite sign ($\uparrow\downarrow$). Spin glasses encode both positive and negative couplings, that are typically assumed to be randomly drawn from a certain distribution (for instance, a Gaussian). Combining all the interactions together, it might happen in the antiferro / spin glass scenario that no configuration satisfies all the constraints induced by the couplings. This behaviour determines a *frustration* in the model. The ground-state free energy of such systems typically displays an exponential number of minima, corresponding to those configurations that minimize the number of un-satisfied constraints. The main issue is that these configurations depend on the particular instance of the disordered couplings. The common approach used in this case is to analyze the typical properties of the free energy with respect to the distribution of the interactions, that are assumed to be varying on a time-scale much larger with respect to thermal fluctuations (this feature is referred to as *quenched disorder*). The corresponding free energy is obtained by averaging with respect to the parameters distribution, whose computation requires highly non trivial analytic tools (among them, the replica method [89]).
Historically, the ferromagnetic model is the first one introduced in the seminal works by Lenz [80] and Ernst Ising [66] on a one-dimensional chain. However, the first exact solution displaying a ferromagnetic-paramagnetic phase-transition at finite T was obtained by Onsager on the square lattice [102], in the absence of external fields. The general case in $d \geq 3$ still lacks for an exact

solution in the thermodynamic limit: in this case, the equilibrium behaviour of the model can be analyzed both by using numerical techniques (such as Monte Carlo sampling), mean-field like theories (as we will discuss in the next Chapter) or eventually using field theories (such as conformal bootstrap).

With regard to spin glass models, a first mean-field theory was developed by Sherrington and Kirkpatrick [126], who introduced an analytic solution for the fully connected spin glass with random couplings (or SK model, named after the authors). A huge step forward to understand the nature of the spin glass phase at low temperatures was developed by Parisi in a series of works [89, 106], postulating how the symmetry of the paramagnetic phase is broken into a hierarchical multi-valley structure, which turns out to be exact in the fully connected spin glass.

However, in the present manuscript we will not deal with ensemble properties [1], all the discussion being presented at the *single instance* level, i.e. for a given realization of the Hamiltonian.

As an overall final remark, the success of discrete (or spin) models in statistical physics relies on their ability to describe a large variety of cooperative phenomena observed several fields of pure and applied science: for instance, the Ising model can be used to describe phase separations in binary mixtures [140], DNA thermal denaturation [136], cancer growth [131], urban segregation [122] or financial markets [150].

Any discrete model with pairwise interactions can be defined on a generic topology, depending on how the interactions are arranged. Mathematically, the topological connections between degrees of freedom can be described in terms of a *graph*, as discussed in the next section.

1.2 A brief introduction to Graph Theory

Graphs are mathematical structures encoding pairwise relations/interactions between pairs of objects. A graph $G = (V, E)$ is defined by a (ordered) set $V = \{1, \ldots, N\}$ of vertices - also called nodes - and by a set $E = \{(i, j) \mid i, j \in V, i \neq j\}$ of node pairs, called links or edges. Each node is associated to a degree of freedom and links represent the interactions or relations between the two nodes connected by the link. The topological connections between nodes can be encoded in the so-called *adjacency matrix* of the graph, a $N \times N$ square matrix whose entries are defined as follows:

$$\mathcal{A}_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases} \tag{1.12}$$

so that all the non-zero elements of $\mathcal{A}$ correspond to an edge (i, j). Nodes connected between an edge are tipycally referred to as *adjacent* or *nearest neighbours* (n.n.), a nomenclature that will be extensively used in the rest of the manuscript. Graphs can be *directed* or *undirected*: in the latter case, if a node i is connected to another j it is also true that j is connected to i; as a consequence, the adjacency matrix defined by (1.12) is symmetric by construction. On directed graphs, each edge is an *ordered* pair of nodes, i.e. it is characterized by a direction on which the link is defined; as a consequence, the corresponding adjacency matrix is not symmetric. Figure 1.1 shows two examples of an undirected (left) graph and a directed (right) graph with five nodes. The neighborhood of a node i is the set of nodes directely connected to i by a link, namely $\partial i = \{j \in V \mid (i, j) \in E\}$. We also define the degree of a certain node i as the number of nodes in its neighborhood, i.e. $d_i = |\partial i|$. On directed graphs, one should distinguish between inner and outer links with respect to a node i, where the inner (resp. outer) edges in the neighborood are defined as the number of links that point to (resp. start from) node i. A *walk* (or *path*) in

[1] Here the word "ensemble" is intended with respect to the distribution of interaction strenghts in the Hamiltonian.

a graph is a sequence of edges connecting two nodes: in particular, given a sequence of vertices $\{k_0, \dots, k_{n+1}\}$ with $k_0 = i$ and $k_{n+1} = j$, a walk between i and j is defined by the set of edges $\{(k_a, k_{a+1}) \in E, a = 0, \dots, n\}$, n being the path's length. Walks are closed if the starting and the ending node coincide, i.e. $i \equiv j$; closed walks are typically called *loops*. An important distinction that will be used in the rest of the manuscript is made between *sparse* and *dense* graphs. A dense graph is characterized by a number of edges that is "close" to its maximum number: the extreme case is achieved when each node is connected to all the others, so that $|E| = \binom{N}{2} \sim O\left(N^2\right)$ and the corresponding adjacency matrix is given by $\mathcal{A}_{ij} = (1 - \delta_{ij})$; this topology is called *fully-connected*. Conversely, sparse graphs are characterized by a "low" number of edges w.r.t. the fully-connected limit, so that $|E| \sim O\left(N^\alpha\right)$ with $\alpha < 2$. At fixed number of vertices, the graph containing the smallest number of edges while still connecting all the nodes by at least one link is called *tree* (or acyclic graph): in this case, the number of edges is simply equal to the number of vertices minus one, i.e. $|E| = N - 1$; tree graphical models are particularly relevant in statistical physics since they can be efficiently solved using message-passing techniques, as we will discuss in the next Chapter. The graph formalism just introduced allows to define models with pairwise interactions on any topology. From now on, we will refer to the Ising model as any binary spin model with 1 and 2-body couplings, defined on a generic graph $G = (V, E)$:

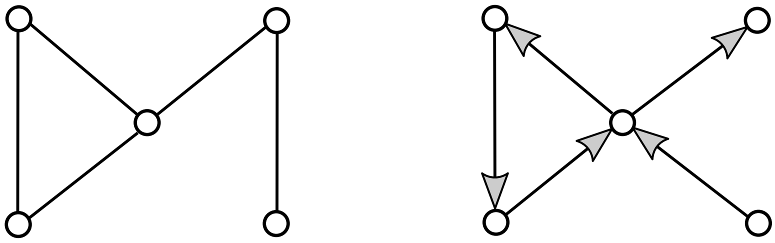

Figure 1.1: Example of an undirected graph (left) and directed graph (right) with $N = 5$ vertices. The arrows in the directed graph indicate the direction of the links.

$$H(\boldsymbol{\sigma}) = -\sum_{(i,j)\in E} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \tag{1.13}$$

$$p(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left[\sum_{(i,j)\in E} \beta J_{ij}\sigma_i\sigma_j + \beta\sum_i h_i\sigma_i\right] \tag{1.14}$$

where the couplings are assumed to be non-zero for all the nodes i, j connected by a link in the graph, and we used a vectorial notation $\boldsymbol{\sigma} = \{\sigma_1, \dots, \sigma_N\}$ to indicate the full set of variables.

In the rest of the thesis, we will make use of several graph architectures with different properties, depending on the presence of short/long loops. A simple distinction can be made between graphs that *locally* behave as trees and graphs with short loops. For instance, random graphs generated according to the Erdős-ény model [46] or with constant degree (also called random regular graphs) can be considered as locally tree-like: the reason is that the typical length of loops increases with the number of nodes as $O(\log N)$. Among the architectures that contain short loops we will mainly deal with lattices, i.e. graphs with constant degree (apart from eventual open boundary conditions) whose edges distribution is represented by a regular structure with no topological randomness (linear chains, regular planar graphs, hypercubic lattices, and so on). Lattice models are among the most studied objects in condensed matter physics, mainly because of the analogies with the typical atomic arrangment in solid crystals; moreover, the regular structure of these graphs allows (in some cases) to carry out analytic calculations, by exploiting additional model symmetries [128]. Somehow in between the two above classes, we will also deal with other architectures inspired by real-world networks, such as small-world [137] and scale-free graphs [8].

1.2.1 Factor graphs

Graphs defined as in the previous section are well-suited to describe models with pairwise interactions. However, many physical systems or analogous models in biology or computer science can display high-order relations between degrees of freedom. A natural way of generalizing graphical models previously introduced to this setting exploits the factor graph representation [99]. The factor graph is a powerful construction that allows to describe the factorization of a certain multivariate probability distribution in terms of "local" objects, encoding for arbitrary high-order interactions.

A factor graph is defined by a set of *variable* nodes $V = \{i\}_{i=1}^{N}$, a set of *factor* (or *check*) nodes $F = \{a\}_{a=1}^{M}$ and a set of edges $E = \{(i, a) \mid i \in V, a \in F\}$: edges always connect nodes of the two different sets, so that the resulting graph is bi-partite. Each node $i \in V$ represents a real (or discrete) degree of freedom x_i; conversely, each factor node $a \in F$ is associated to a certain non-negative function ψ_a (also called *compatibility* or *potential* function) that takes into account the local mutual dependencies between the subset of variable nodes connected to a by a link $(i, a) \in E$. A toy example of a factor graph is shown in Figure 1.2.

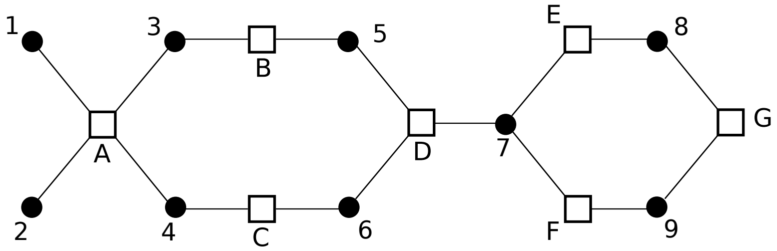

Figure 1.2: Toy example of a factor graph with 9 variable nodes and 7 factor nodes

The joint probability distribution over the factor graph $G = (V, F, E)$ with prescribed compatibility functions $\{\psi_a\}_{a\in F}$ can be expressed as:

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{a\in F}\psi_a(\boldsymbol{x}_{\partial a}) \tag{1.15}$$

The above expression defines a *probabilistic graphical model* (also called *markov random field*) [78, 99, 134], and it will be the central object under investigation through the rest of this thesis. The potential functions ψ_a can, in principle, have a very general expression depending on the specific interactions of the problem investigated, so that the factor graph formalism can be used to describe a wide range of models in inference, machine learning, combinatorial optimization. In the latter case, (1.15) can be used to describe the set of solutions of constraint satisfaction problems like k-SAT [99]. In statistical physics, any model of interacting degrees of freedom described by a Hamiltonian of the type (1.1) shows indeed such a factorization property, as a consequence of the locality of physical interactions. In particular, the Boltzmann law (1.3) can be expressed into the form (1.15), by identifying each k-body interaction term in the Hamiltonian (1.1) with a compatibility function, thanks to the following map:

$$\tilde{J}_{i_1,\ldots i_k}\phi^{(k)}_{i_1,\ldots,i_k}(x_{i_1},\ldots,x_{i_k}) = -\frac{1}{\beta}\log\psi_a(x_{i_1},\ldots,x_{i_k}) \tag{1.16}$$

where $\partial a = \{i_1, \dots, i_k\}$. If all the factor nodes have degree 2, the factor graph reduces to a simple graph, so that $a \equiv (i, j)$ and the potential functions $\psi_a \equiv \psi_{ij}$ describe the mutual local interaction between two nodes. In this setting, Eq. 1.15 defines a *pairwise* graphical model.

1.3 Statistical Inference

The word *inference* denotes a generic process that allows to retrieve information from a certain amount of data. Nowadays, inference problems appear in any field of science, from signal processing to computational biology, from neuroscience to artificial intelligence, and - particularly relevant in the *post-Covid world* - in the context of epidemic spreading. The increasing availaibility of large-scale datasets in the last decades is challenging the scientific community to develop computational efficient tools to perform inference in high-dimensional systems.

Since data are always affected by some noise, its is necessary to rely on a probabilistic description: therefore, with the name *statistical inference* we typically refer to the process of deducing some information about a probabilstic model by the analysis of data, whose most common approach is based on the Bayesian framework. A Bayesian inference problem aims at deriving features of some unknown set of variables $\{x_i\}_{i=1,\dots,N}$, by knowing some measurements of another set of quantities $\{y_\mu\}_{\mu=1,\dots,M}$; furthermore, we assume to know a model (or hypothesis) $\mathcal{H}$ describing the deterministic relation between the two sets of variables.

Bayesian inference allows to express the probability distribution of the variables $\boldsymbol{x}$ to be inferred conditioned to the observed data $\boldsymbol{y}$, and it can be derived from law of conditional probability [87] (or, more directly, from the Bayes theorem):

$$p\left(\boldsymbol{x} \mid \boldsymbol{y}, \mathcal{H}\right) = \frac{p\left(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{H}\right) p_0\left(\boldsymbol{x}\right)}{p\left(\boldsymbol{y}, \mathcal{H}\right)} \tag{1.17}$$

Each of the quantities in (1.17) has a speficic meaning and nomenclature. The term $p\left(\boldsymbol{x} \mid \boldsymbol{y}, \mathcal{H}\right)$ at the left-hand side is called *posterior* distribution, representing the probability of the unknown quantites conditioned to the observed data: it quantifies our belief *after* observing the data. Conversely, $p_0\left(\boldsymbol{x}\right)$ is the *prior* distribution, encoding additional *a-priori* information we assume to know about the quantities to be inferred: in other words, it represents our belief *before* observing the data. The term $p\left(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{H}\right)$ is the *likelihood*, quantifiying the probability of observing the data for given values of the unknown parameters, and it has to be interpreted as a function of the latters at given data. In practice, the likelihood represents as a stochastic function, encoding how the observed quantities are related to the unknown ones with the additional presence of noise. Finally, $p\left(\boldsymbol{x} \mid \mathcal{H}\right)$ is the *evidence*, and it is a constant w.r.t. the unknown variables $\boldsymbol{x}$, that ensures the normalization of the posterior.

In many applications, the model is given by external information, i.e. laws describing the relation between $\boldsymbol{x}$ and $\boldsymbol{y}$. A simple example is given by Linear Estimation Problems (discussed in the next chapter). The unknown vector $\boldsymbol{x}$ might as well describe the parameters of a certain statistical model describing an input-output relation, as it happens in linear regression or, more in general, in supervised learning.

There are also situations where we only have empirical measurements about some quantity of interests, without a stastitical model describing their mutual dependencies: in these cases, it is necessary to exploit an *effective* statistical description, such that the resulting model is the "fairest" one to describe the data. The most common - and rigorous - approach relies on the maximum entropy principle [125], that allows to reconstruct the least-biased distribution (in terms of entropy) subject to additional constraints in such a way to be compatible with the data. We will come back to this issue in Chapter 5 for the inverse Ising problem.

1.3.1 Connection to Statistical Mechanics

The Bayesian approach can be directly linked to statistical mechanics by rewriting the posterior distribution in terms of the Boltzmann law:

$$p^{\beta}\left(\boldsymbol{x} \mid \boldsymbol{y}, \mathcal{H}\right)=\frac{1}{p\left(\boldsymbol{y}, \beta\right)} \exp\left[\beta \log p\left(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{H}\right)+\beta \log p_{0}\left(\boldsymbol{x}\right)\right]=\frac{e^{-\beta H(\boldsymbol{x} \mid \boldsymbol{y})}}{Z\left(\boldsymbol{y}, \beta\right)} \tag{1.18}$$

where in the right-most hand side H is the Hamiltonian, to be considered as a function over the $\boldsymbol{x}$ components, each one representing a microscopic dregree of freedom. Further notice that the additional parameter β plays the role of a fictious temperature, providing an exact mapping to the statistical physics picture. In particular, the MAP estimator coincides with the ground state of the Hamiltonian in (1.18), at it can be computed by letting $\beta \to \infty$. In the language of disordered systems, the data $\boldsymbol{y}$ play the role of quenched disorder. Moreover, in this notation the evidence is nothing but the partition function Z of the model. Typically, the prior distribution is factorized over the $\boldsymbol{x}$ components, so that the second term in (1.18) is interpreted as a 1-body interaction term, while all the remaining interacting part is encoded into the (log) likelihood function.

The connection between statistical physics and inference dates back to the seminal works by Jaynes [68] and Shannon [125], and nowadays the two branches are more and more connected, as developments into one of the two find applications in - or give more understanding to - the other. This connection allows to interpret common estimators used in Bayesian inference in statistical physics terms: for instance, the maximum-a-posteriori (MAP) estimator is equivalent to the ground state of the Hamiltonian in (1.18), and it can be computed in the limit $\beta \to \infty$; conversely, marginal probabilities computed over the posterior coincide with equilibrium expectation values over the Botlzmann measure at the right hand side.

The connection to statistical physics further allows to address more information-theoretic questions: for instance, under which conditions the information encoded in the data is sufficient to retrieve the unknown quantities. This type of questions can be formulated in terms of algorithmic phase transitions, separating *easy* regimes where inference is possible from *hard* regimes where it is not, typically depending on the fraction between the number of observations M and the number of variables to be inferred N. Furthermore, somehow in between the two aforementioned phases one should distinguish regimes where inference is computationally feasible from where it is not [149].

We finally remark that many problems in statistical inference can be considered as *dual* or *inverse* with respect to what typically done in statistical physics: indeed, in the latter case one typically starts from a known model with the aim of computing some relevant observables; in the former the process is reversed, so that we want to gain insights of a unknown model by knowing some observations. This connection will be transparent in Chapter 5 where we will discuss the inverse Ising problem.

Part I

Loop corrections in discrete graphical models through Density Consistency

Chapter 2

Approximate methods in statistical physics

This chapter reviews some approximation methods commonly used in statistical physics. In particular, Section 2.1 introduces a variational approach to compute free energy approximations in probabilistic graphical models, based on the Gibbs free energy minimization. This setting allows to derive the Mean-Field method and the Bethe Approximation, discussed respectively in Sections 2.1.1 and 2.2. Section 2.3 summarizes other techniques that improve the Bethe Approximation on loopy graphs. Finally, 2.4 describes Expectation Propagation in the context of high-dimensional inference. Most of the contents of this chapter will be needed to better understand the derivation of Density Consistency in Chapter 3, and to make numerical comparisons in Chapter 4.

2.1 Variational methods

The free energy is a fundamental quantity in statistical physics, as it contains all the information required to describe the macroscopic behaviour of a physical system at equilibrium. Knowing the free energy and its dependency on the model parameters (specified by the Hamiltonian) and on external control quantities (e.g. temperature), any observable can be computed by performing suitable derivatives on it. However, there are only a few examples in which this quantity can be computed exactly, for instance in very homogeneous models in the thermodynamic limit. For a generic physical system of N interacting degrees of freedom, the computation of the free energy is an intractable problem: in particular, when the variables take values on a discrete finite alphabet $\mathcal{X}$, its computation scales exponentially with the system size N as $O\left(|\mathcal{X}|^N\right)$. In the language of computational complexity, evaluating the free energy is typically a $\sharp$P-complete problem [95, 99]. For this reason, in the statistical physics community a lot of effort has been devoted to design tractable approximations to the free energy. In the following, we will discuss variational methods in the generic context of probabilistic graphical model discussed in Section 1.2.1. We will mainly restrict to models of discrete variables, denoted with $\sigma_i \in \mathcal{X}$, even if the same reasoning can be applied to continuous degrees of freedom. Let us consider the following graphical model, whose probability density is given by:

$$p\left(\boldsymbol{\sigma} \mid \beta, \boldsymbol{\theta}\right) = \frac{1}{Z} \exp\left[-\beta H\left(\boldsymbol{\sigma} \mid \boldsymbol{\theta}\right)\right] = \frac{1}{Z} \prod_a \psi_a\left(\boldsymbol{\sigma}_{\partial a} \mid \beta, \boldsymbol{\theta}\right) \tag{2.1}$$

where $\beta = T^{-1}$ and we set $k_B = 1$ for simplicity. The above equivalence holds for any model defined by a Hamiltonian H with local interaction terms, included into $\boldsymbol{\theta}$, so that the Boltzmann distribution can always be written as a product of local terms ψ_a. We recall the definition of the Helmholtz free energy:

$$F(\beta, \boldsymbol{\theta}) = -\frac{1}{\beta} \log Z(\beta, \boldsymbol{\theta}) \tag{2.2}$$

In the following, we will drop the dependency on β and on the model parameters $\boldsymbol{\theta}$ to simplify the notation. Variational methods can be designed starting from a functional expression for the free energy, tipycally referred to as the *Gibbs variational free energy* , and defined as follows:

$$\mathcal{F}[q] = \mathcal{U}[q] - T\mathcal{S}[q]. \tag{2.3}$$

The quantities $\mathcal{U}$ and $\mathcal{S}$ refer respectively to the variational internal energy and the variational entropy:

$$\mathcal{U}[q] = \sum_{\boldsymbol{\sigma}} q(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}) = -\frac{1}{\beta} \sum_a \sum_{\boldsymbol{\sigma}_{\partial a}} q(\boldsymbol{\sigma}) \log \psi_a(\boldsymbol{\sigma}_{\partial a}) \tag{2.4}$$

$$\mathcal{S}[q] = -\sum_{\boldsymbol{\sigma}} q(\boldsymbol{\sigma}) \log q(\boldsymbol{\sigma}) \tag{2.5}$$

where the right hand side of (2.4) follows from (1.16). In the above expressions, q is a *trial* probability distribution, and the square brackets in (2.3) indicate that the quantity $\mathcal{F}$ is a functional of q. The variational principle states that the physical system under investigation is described at equilibrium by the distribution q^* minimizing (2.3), subject to the constraint that q^* has to be normalized to 1. Constraints can be included by means of Lagrange multipliers and using the method of constrained optimization. Let us consider the following modified Gibbs free energy, where the normalization constraint is included by using a Lagrange multiplier λ:

$$\mathcal{F}_\lambda[q] = \mathcal{F}[q] + \lambda \left(\sum_{\boldsymbol{\sigma}} q(\boldsymbol{\sigma}) - 1 \right) \tag{2.6}$$

It is straightforward to show that minimizing (2.6) w.r.t. $q(\boldsymbol{\sigma})$ leads to the Boltzmann distribution (2.1) and the Gibbs free energy minimum corresponds to the Heltmholtz free energy:

$$p = \underset{q}{\operatorname{argmin}}\, \mathcal{F}_\lambda[q] = \frac{1}{Z} e^{-\beta H(\boldsymbol{\sigma})} \tag{2.7}$$

$$F = \min_q \mathcal{F}_\lambda[q] = \mathcal{F}_\lambda[p] \tag{2.8}$$

Equivalently, one can show that inserting the Boltzmann law as trial distribution into (2.3) leads to the Helmholtz free energy (2.2). However, this formal justification of the variational principle does not help in computing the free energy. Indeed, in deriving (2.7), we just assumed that the trial probability distribution has to be normalized. In statistical mechanics, many approximations rather exploit *factorized* trial distributions to approximate the true free energy, and the minimization procedure leads in general to a system of self-consistent equations to be solved iteratively do determine the approximate trial probabilities. Using the above reasoning, the variational principle can also be rephrased as:

$$\mathcal{F}[q] = F + \frac{1}{\beta} D_{KL}(q \,||\, p) \tag{2.9}$$

where D_{KL} is the Kullback-Leiber divergence between two probability distributions [38]:

$$D_{KL}(q \,||\, p) = \sum_{\boldsymbol{\sigma}} q(\boldsymbol{\sigma}) \log \frac{q(\boldsymbol{\sigma})}{p(\boldsymbol{\sigma})} \tag{2.10}$$

In this way, the Gibbs free energy can be used to compute upper bounds to the true free energy, since $D_{KL}(q \,||\, p) \geq 0$, the equality being satisfied only when $q(\boldsymbol{\sigma}) \equiv p(\boldsymbol{\sigma})$.

2.1.1 Mean Field

The simplest variational approach can be obtained by assuming that the trial probability distribution is factorized over single-node marginals: this approximation is known as naïve Mean-Field (nMF, or simply Mean-Field), and it represents the first step in analyzing any model in statistical physics, by relying on the simple assumption that the degress of freedom are uncorrelated. The MF trial distribution has the following expression:

$$q^{MF}\left(\boldsymbol{\sigma}\right)=\prod_{i\in V}q_i\left(\sigma_i\right) \tag{2.11}$$

where each q_i is the single-node marginal over node i, also called *belief*. By plugging the above trial expression into the Gibbs Free energy functional (2.3), it is straightforward to show that the corresponding Mean-Field variational free energy is given by:

$$\mathcal{F}\left[q^{MF}\right]=-\frac{1}{\beta}\sum_{a\in F}\sum_{\sigma_i,i\in\partial a}\log\psi_a\left(\boldsymbol{\sigma}_{\partial a}\right)\prod_{i\in\partial a}q_i\left(\sigma_i\right)+T\sum_{i\in V}\sum_{\sigma_i}q_i\left(\sigma_i\right)\log q_i\left(\sigma_i\right) \tag{2.12}$$

Note that while the Gibbs free energy is a function of the full joint probability distribution, the mean-field free energy is only a functional of single-node beliefs and it contains a polynomial amount of terms w.r.t. the system size N. After adding a set of normalization constraints for each single-node belief, we are left with the following minimization:

$$F^{MF}=\min_{\{q_i\}_{i\in V}}\left\{\mathcal{F}\left[q^{MF}\right]+\sum_{i\in V}\lambda_i\left(\sum_{\sigma_i}q_i\left(\sigma_i\right)-1\right)\right\} \tag{2.13}$$

where we introduced a set of Lagrange multipliers $\{\lambda_i\}_{i\in V}$ and F^{MF} corresponds to the mean-field approximation to the true free energy. Minimizing the above expression leads to a set of self-consistent equations to determine the marginals q_i. On the other hand, minimization w.r.t. the multipliers λ_i ensures that the single-node beliefs are correctly normalized. Once the self-consistent equations are solved, one can plug in their expression into 2.12 to get the approximate Mean-Field free energy F^{MF}. The main drawback of the Mean-field approximation is that it neglects all the information about correlations, so that the joint distribution of any set of variables (or equivalently, their expectation value) is trivially factorized:

$$\langle\sigma_{i_1},\ldots,\sigma_{i_n}\rangle_{q^{MF}}=\prod_{k=1}^{n}\langle\sigma_{i_k}\rangle_{q^{MF}} \tag{2.14}$$

Mean field theory for the Ising model

As a simple example, we breifly discuss the mean-field theory for the ferromagnetic Ising model. In this case, the degrees of freedom take two symmetric values, i.e. $\sigma_i\in\{-1,1\}$. This allows to parametrize the single-node beliefs q_i by a unique real quantity, namely the magnetization m_i, which is nothing but the expectation value of σ_i over q_i:

$$q_i\left(\sigma_i\right)=\frac{1+\sigma_i m_i}{2} \tag{2.15}$$

$$\langle\sigma_i\rangle_{q_i}=q_i\left(+1\right)-q_i\left(-1\right)=m_i \tag{2.16}$$

Plugging in the parametrization (2.15) into (2.12) allows to rewrite the Mean-Field variational free energy with an explicit dependence over the magnetizations:

$$F^{MF}\left(\{m_i\}_{i\in V}\right) = -\sum_i h_i m_i - \sum_{(i,j)\in E} J_{ij} m_i m_j - \frac{1}{\beta}\sum_{i=1}^{N}\left[\mathcal{H}\left(\frac{1+m_i}{2}\right) + \mathcal{H}\left(\frac{1-m_i}{2}\right)\right] \quad (2.17)$$

where $\mathcal{H}(x) = -x\log x$. Note that F^{MF} is not anymore a functional but rather a function of the full set of magnetizations $\{m_i\}_{i=1}^{N}$. Its minimum can be found by setting

$$\frac{\partial F}{\partial m_i} = 0 \quad \forall i = 1,\ldots,N \quad (2.18)$$

which leads to the following set of self-consistent equations for the magnetizations:

$$m_i = \tanh\left[\beta\left(h_i + \sum_{j\in\partial i} J_{ij} m_j\right)\right] \quad \forall i = 1,\ldots,N \quad (2.19)$$

Notice that there is no need to add the normalization constraints in the minimization in (2.17) as the single-node marginals (2.15) are already normalized by construction. In the Mean-Field approximation, degrees of freedom turn out to be uncorrelated, but each of them is subject to an effective local field resulting from the combined action of its neighbours. Eq. (2.19) can be iteratively solved to provide an estimate of the equilibrium behaviour of the model. In particular, on ferromagnetic models defined on hypercubic lattices, the above expression can be further simplified by assuming a constant coupling J among nearest neighbour spins:

$$m = \tanh\left[2d\beta J m\right] \quad (2.20)$$

where $2d$ is the degree of each node and d is the dimensionality of the lattice, and we set $h = 0$ for simplicity. At thermodynamic limit, the mean-field theory predicts a second-order phase transition at a critical temperature $T_c = 1/\beta_c = 2dJ$ (in zero field), where a spontaneous non-zero magnetization arises below T_c. This means that mean field theory is able to capture the presence of a spontaneous symmetry breaking, marking the onset of a paramagnetic-ferromagnetic transition. However, the mean field theory for the Ising model is exact only in the fully connected limit: in the above expression, this can be obtained by letting the number of dimensions d go to infinity, and by properly rescaling the couplings as $J \to J/2d$ in order to have an intensive free energy density w.r.t. N. The universal behaviour of the model given by its critical exponents is also known to be predicted by the Mean Field theory for $d \geq 4$ ($d_u = 4$ is known as the *upper critical dimension* for the Ising model).
As a final comment, we remark that, despite the correlations are not taken into account under the Mean Field approximation, one can use Linear Response theory to estimate their contribution [57, 71]: we will come back to this point in the context of the Inverse Ising Problem discussed in Chapter 5.

2.2 The Bethe Approximation and Belief Propagation

A natural way to go beyond the Mean-field approximation can be obtained by considering a trial distribution factorized over function-node marginals rather than variable-nodes. This approximation is named after Bethe and Peierls who first introduced it in the context of lattice ferromagnetic models [14, 109]. A more general approach on factor graph models was developed by Yedidia, Weiss, Freeman in a series of works [143, 145, 146]. For reasons that will be clear in the following, the derivation of the Bethe approximation is typically carried out starting from models on acyclic graphs (trees), and then extended to generic loopy graphs. Following this approach, we start from a graphical model defined by (2.1) on an acyclic factor graph, whose probability distribution can be expressed as follows:

$$p\left(\boldsymbol{\sigma}\right)=\frac{\prod_{a}q_{a}\left(\boldsymbol{\sigma}_{\partial a}\right)}{\prod_{i\in V}q_{i}\left(\sigma_{i}\right)^{\left|\partial i\right|-1}}\tag{2.21}$$

where

$$q_{a}\left(\boldsymbol{\sigma}_{\partial a}\right)=\sum_{\boldsymbol{\sigma}_{\setminus\partial a}}p\left(\boldsymbol{\sigma}\right)\tag{2.22}$$

$$q_{i}\left(\boldsymbol{\sigma}_{i}\right)=\sum_{\boldsymbol{\sigma}_{\setminus i}}p\left(\boldsymbol{\sigma}\right)\tag{2.23}$$

are respectively the factor and node marginals. If the graphical model does not contain loops, Eq. (2.21) is an *exact* expression, and (2.22)-(2.23) represent the true marginals over factor/variable nodes. The denominator in Eq. (2.21) needs to be introduced in order to remove the effect of the overcounting of each single-node's contribution to the factor probabilities q_a: indeed, each variable σ_i appears in a number $|\partial i|$ of factor-node marginals, so that it is overcounted exactly $|\partial i|-1$ times in the numerator. Plugging in (2.21) into the Gibbs variational free energy, we get the following expression for the *Bethe variational free energy*:

$$\begin{aligned}\mathcal{F}^{BA}\left[q\right]=&-\frac{1}{\beta}\sum_{a\in F}\sum_{\sigma_{i},i\in\partial a}\log\psi_{a}\left(\boldsymbol{x}_{\partial a}\right)q_{a}\left(\boldsymbol{\sigma}_{\partial a}\right)+T\sum_{a\in F}\sum_{\sigma_{i},i\in\partial a}q_{a}\left(\boldsymbol{\sigma}_{\partial a}\right)\log q_{a}\left(\boldsymbol{\sigma}_{\partial a}\right)\\&+T\sum_{i\in V}\left(1-\left|\partial i\right|\right)\sum_{\sigma_{i}}q_{i}\left(\sigma_{i}\right)\log q_{i}\left(\sigma_{i}\right)\end{aligned}\tag{2.24}$$

The above expression, involves only a polinomial amount of terms to be computed w.r.t. N, as in the Mean-Field case: in particular, the first term in (2.24) corresponds to the energetic part of the free energy, while the others refer respectively to the entropy of each function-node and variable-node marginals, the latters being multiplied by $1-|\partial i|$ to avoid overcounting in the entropic contribution of factor-node marginals, as previously discussed. If the graph is acyclic, the Bethe free energy coincides with the true free energy of the model when the $\{q_a\}$ and $\{q_i\}$ coincide to the true marginals, and its stationary points can be computed by a polinomial iterative algorithm known as Belief Propagation, discussed in the next section.

2.2.1 Belief Propagation

The constrained minimization of (2.24) results in a set of self-consistent equation that can be solved in polynomial time by using an iterative scheme known as Belief Propagation (BP). It was first introduced by Gallager in the context of decoding algorithms and then generalized by Pearl in [108]. However, the connection between the Belief Propagation equations and the Bethe free energy was first clarified in [70], and further generalized by Yedidia and collaborators in a series of seminal works [143–146]. In the following, we will follow the latters' approach, described also in [99]. The minimization of the Bethe free energy can be carried out with respect to a set of node and factor beliefs, by enforcing the normalization constraints for each of them, plus an additional set of local consistency conditions between the two beliefs:

$$\sum_{\boldsymbol{\sigma}_{\partial a}} q_a\left(\boldsymbol{\sigma}_{\partial a}\right)=1 \quad \forall a \in F \tag{2.25}$$

$$\sum_{\boldsymbol{\sigma}_i} q_i\left(\sigma_i\right)=1 \quad \forall i \in V \tag{2.26}$$

$$\sum_{\boldsymbol{\sigma}_{\partial a \setminus i}} q_a(\boldsymbol{\sigma}_{\partial a})=q_i\left(\sigma_i\right) \qquad \forall a \in F, i \in \partial a \tag{2.27}$$

As previously done with the Mean-Field approximation, these constraints can be included through a suitable set of Lagrange multipliers. Thefore, let us define the following constrained free energy, denoted with $\mathcal{L}$:

$$\mathcal{L}^{BA}=\mathcal{F}^{BA}+\sum_i \gamma_i\left[\sum_{\sigma_i} q_i\left(\sigma_i\right)-1\right]+\sum_i \sum_{a\in\partial i} \sum_{\sigma_i} \lambda_{ai}\left(\sigma_i\right)\left[q_i\left(\sigma_i\right)-\sum_{\boldsymbol{\sigma}_{\partial a\setminus i}} q_a\left(\boldsymbol{\sigma}_{\partial a}\right)\right] \tag{2.28}$$

Notice that the constraints (2.25) do not need to be inserted in the above expression, as they follow directly from the other two conditions. Setting to 0 the derivatives of (2.28) with respect to the beliefs q_i, q_a gives the equations for the beliefs at the fixed point as functions of the Lagrange multipliers $\left\{\gamma_a\right\}_{a\in F}, \left\{\gamma_i\right\}_{i\in V}, \left\{\lambda_{ai}\right\}_{a\in F}^{i\in\partial a}$; on the other hand, imposing the stationarity of (2.28) over the multipliers enforces the normalization and consistency conditions (2.25)-(2.26)-(2.27). The connection between the Belief Propagation update equations and the stationary points of (2.6) is obtained by identifying the Lagrange multipliers λ_{ai} with the following quantities [99, 143]:

$$\lambda_{ai}\left(\sigma_i\right)=\log\nu_{i\to a}\left(\sigma_i\right)=\log\prod_{b\in\partial i\setminus a} m_{b\to i}\left(\sigma_i\right) \tag{2.29}$$

Inserting the above expression into the saddle-point equations for (2.27), one gets the Belief Propagation fixed point equations for the site and factor beliefs, shown below:

$$q_a\left(\boldsymbol{\sigma}_{\partial a}\right)=\frac{1}{Z_a}\psi_a\left(\boldsymbol{\sigma}_{\partial a}\right)\prod_{i\in\partial a}\nu_{i\to a}\left(\sigma_i\right) \qquad \forall a\in F \tag{2.30}$$

$$q_i\left(\sigma_i\right)=\frac{1}{Z_i}\prod_{a\in\partial i} m_{a\to i}\left(\sigma_i\right) \qquad \forall i\in V \tag{2.31}$$

The quantities $\nu_{i\to a}\left(\sigma_i\right), m_{a\to i}\left(\sigma_i\right)$ are called *messages*, and represent the variational parameters used to update BP equations until convergence. Their update rules follow directly from (2.29) and by imposing the consistency condition (2.27) on (2.30)-(2.31). After some straighforward algebra, we get:

$$m_{a\to i}^{(\tau+1)}\left(\sigma_i\right)\propto\sum_{\boldsymbol{\sigma}_{\partial a\setminus i}}\psi_a\left(\boldsymbol{\sigma}_{\partial a}\right)\prod_{j\in\partial a\setminus i}\nu_{j\to a}^{(\tau)}\left(\sigma_j\right) \tag{2.32}$$

$$\nu_{i\to a}^{(\tau+1)}\left(\sigma_i\right)\propto\prod_{b\in\partial i\setminus a} m_{b\to i}^{(\tau)}\left(\sigma_i\right) \tag{2.33}$$

where τ is an iteration number. Eqs. (2.32)-(2.33) are known as the belief propagation update equations: in both cases, the symbol $\propto$ states that the two sides of the equation must be equal apart from a constant factor, to be computed by imposing the normalization of the left-hand sides. A graphical representation of the variable/factor node beliefs and the update rules for the message is shown in Figure 2.1-2.2, respectively. We remark that, in principle, one could

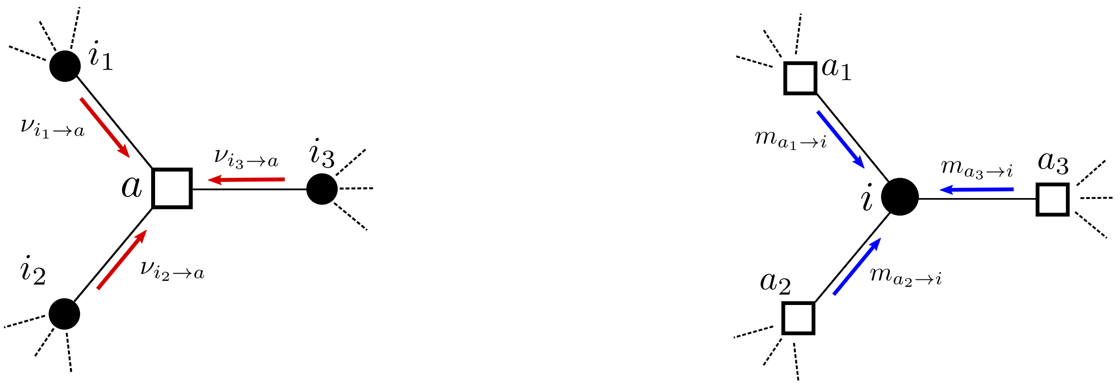

Figure 2.1: Left: graphical representation of a factor node belief $q_a(\boldsymbol{\sigma}_{\partial a})$ given by (2.30), where $\partial a = \{i_1, i_2, i_3\}$; red arrows represent variable-to-factor messages $\nu_{i\to a}(\sigma_i)$, and the compatibility function ψ_a is implicitly included into the white square representing factor node a. Right: graphical representation of a variable node belief $q_i(\sigma_i)$ given by Eq. (2.31), where $\partial i = \{a_1, a_2, a_3\}$; blue arrows represent factor-to-variable message $m_{a\to i}(\sigma_i)$.

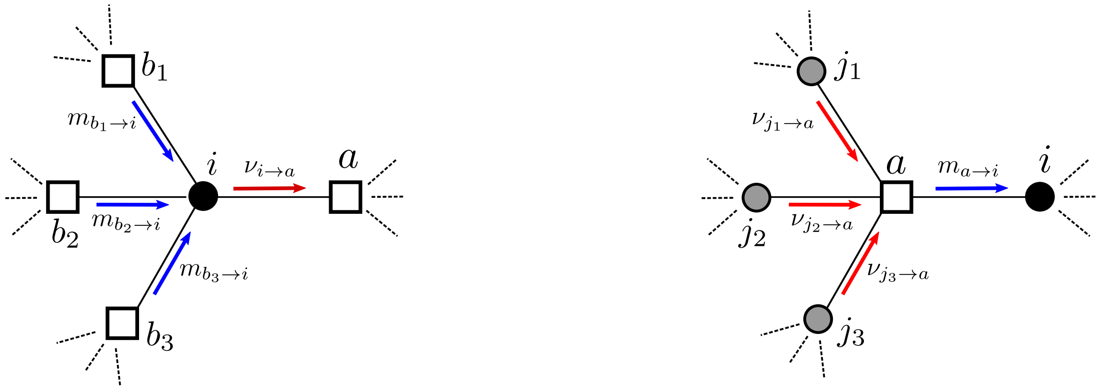

Figure 2.2: Left: graphical representation of the update rule for the variable-to-factor message $\nu_{i\to a}(\sigma_i)$ as given by Eq. (2.33), where $\partial i\backslash a = \{b_1, b_2, b_3\}$. Right: graphical representation of the update rule for the factor-to-variable message $m_{a\to i}(\sigma_i)$ given by (2.32), where $\partial a\backslash i = \{j_1, j_2, j_3\}$; the variable nodes traced over $(\partial a\backslash i)$ are represented as gray dots. In both panels, red and blue arrows represent respectively variable-to-factor messages $\nu_{i\to a}$ and factor-to-variable message $m_{a\to i}$ as in Figure 2.1.

write a single set of self-consistent equations by choosing just one of the two sets of messages, either $\{m_{a\to i}\}$, i.e. the ingoing messages to variable nodes, or $\{\nu_{i\to a}\}$, i.e. the outgoing messages. The set of equations (2.33)-(2.32) can be iteratively solved up to numerical convergence w.r.t. $\{\nu_{i\to a}, m_{a\to i}\}_{i\in V}^{a\in F}$: once a fixed point is found, the marginal beliefs can be computed using (2.30)-(2.31). The Belief propagation algorithm runs on a polinomial time w.r.t to the system size N, provided that the graph is sparse. For instance, if both factor and variable nodes have a degree $\sim O(1)$ w.r.t. N, the computational cost per iteration scales like $O(2|E|)$, $|E|$ being the number of edges in the factor graph. Notice that the message update rules and the beliefs definition can be used to re-write the Bethe Free energy only in terms of messages:

$$\mathcal{F}^{BA}\left[\{\nu_{i\to a}, m_{a\to i}\}_{i\in V}^{a\in F}\right] = -\sum_{a}\log Z_a - \sum_{i}\log Z_i + \sum_{(ia)}\log Z_{ia} \tag{2.34}$$

where

$$Z_a = \sum_{\boldsymbol{\sigma}_{\partial a}} \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \prod_{i\in\partial a} \nu_{i\to a}\left(\sigma_i\right) \tag{2.35}$$

$$Z_i = \sum_{\sigma_i} \prod_{a\in\partial i} m_{a\to i}\left(\sigma_i\right) \tag{2.36}$$

$$Z_{ia} = \sum_{\sigma_i} m_{a\to i}\left(\sigma_i\right)\nu_{i\to a}\left(\sigma_i\right) \tag{2.37}$$

It can be rigorously proven [99] that the set of BP fixed point equations provides exact values of the marginal distributions when the underlying graph is a tree. In this case, the probability distribution of the original model coincides with (2.21) and the beliefs (2.30)-(2.31) coincide with the true marginals [1]. Moreover, BP fixed points are in one-to-one correspondence with the stationary points of the Bethe free energy [143].

On loopy graphs, the true probability distribution does not factorize as in (2.21), so that the ansatz is not well defined; however, despite it is not possible to define a probability measure over the full set of degrees of freedom like (2.21), we can use the BP scheme to get approximate values of its marginal probabilities: in this way, the Bethe approximation provides a set of beliefs that satisfy *local* consistency conditions of the type (2.27), but they cannot be expressed as marginals of a single joint measure over the full set of variables. As a consequence, the corresponding approximation for the free energy cannot be used as a upper bound to the true free energy. On graphs with loops, this approach is commonly referred to as "Loopy Belief Propagation", even if the message-passing equations are exactly the same as (2.39)-(2.40). The convergence properties of (loopy) BP depend on the particular instance considered (i.e. the graph structure, the distribution of interactions, and the external temperature), so that convergence cannot be guaranteed in general. In particular, the presence of a phase transition in the model is typically connected to the existence to more than one BP fixed points, that in turn correspond to local minima of the Bethe free energy. The connection between BP fixed points and minima of the Bethe free energy allows to exploit other iterative schemes with different convergent properties than the message-passing scheme just discussed [94, 147].

Belief Propagation has been employed for several applications both in statistical physics and inference problems: for instance, in Low-Density-Parity-Check (LDPC) [21], learning in neural networks with discrete synapses [20], and more recently in our group in the context of inference problems in epidemic spreading processes: further details on the latter topic will be addressed in Chapter 7.

Relation to the cavity method In statistical physics, the BP equations can be derived following an equivalent approach known as the *cavity method*. The cavity method allows to obtain recursive equations for the marginal probabilities in a factor graph: these equations are obtained by computing free energy shifts when one node (either function or variable) in the graph is removed, thus creating a *cavity*. The key assumption is that the correlations between the remaining

[1] In this case, the BP equations can efficiently run by choising a proper schedule for the message updates: in particular, by choosing one node as a *root*, the messages can be propagated inwards starting from the leaves towards the root node, and then back out from the root towards the leaves; such procecdure gives the exact values for the marginals in just two steps.

variables are easier to treat in the cavity graph. In this perspective, the Bethe approximation corresponds to a factorized assumption on the neighbours of each factor node a, when the latter is removed from the graph:

$$g^{\setminus a}\left(\boldsymbol{\sigma}_{\partial a}\right) \overset{\text{Bethe}}{\sim} \prod_{i \in \partial a} \nu_{i \to a}\left(\sigma_i\right) \tag{2.38}$$

where $g^{\setminus a}$ denotes the cavity distribution, obtained by removing node a and all its links. This assumption is exact in the absence of loops, since removing one factor node breaks the graph into $|\partial a|$ *disconnected* components. It can be shown that the recursive relations obtained by using the cavity method coincide with the BP update equations. The cavity method allows to interpret the messages as marginal probabilities in a modified graphical model: in particular, the message $\nu_{i \to a}\left(\sigma_i\right)$ represents the marginal of node i in the *cavity* graph where factor node a has been removed; conversely, the message $m_{a \to i}\left(\sigma_i\right)$ corresponds to the marginal of node i in a modified factor graph in which all the factor nodes connected to i, except a, have been removed.
It is important to remark that the name “cavity method” is typically referred to its *ensemble* version, used to analyze the typical[2] properties of tree-like graphical models defined on random graphs [89]. In the next section, we briefly discuss how to apply the Bethe approximation to the Ising model.

2.2.2 Bethe approximation for the Ising model

On a generic pairwise graphical model, the message-passing equations can be easily expressed by choosing just one of the two sets of messages. In this case, all factor nodes have degree two, and in the following we will use the notation $a \equiv (ij)$ to specify the two nodes connected by the edge. Depending on which set is chosen, we get two alternative (sets of) fixed points equations:

$$m_{(ij) \to i}\left(\sigma_i\right) \propto \sum_{\sigma_j} \psi_{ij}\left(\sigma_i, \sigma_j\right) \prod_{k \in \partial j \setminus i} m_{(jk) \to j}\left(\sigma_j\right) \tag{2.39}$$

$$\nu_{i \to (ij)}\left(\sigma_i\right) \propto \prod_{k \in \partial i \setminus j} \sum_{\sigma_k} \psi_{ik}\left(\sigma_i, \sigma_k\right) \nu_{k \to (ik)}\left(\sigma_k\right) \tag{2.40}$$

If now we restrict to the Ising model where $\sigma_i \in \{-1{,}1\}$, each message can be parametrized by a single real number, physically interpreted as a local field. In particular, choosing the outgoing messages $\left\{\nu_{i \to (ij)}\right\}$ for the update rule, we rewrite them as follows:

$$\nu_{i \to (ij)}\left(\sigma_i\right) \hat{=} \frac{e^{\sigma_i \omega_{i \to (ij)}}}{2\cosh\omega_{i \to (ij)}}; \qquad \omega_{i \to (ij)} = \frac{1}{2} \log \frac{\nu_{i \to (ij)}\left(\sigma_i = +1\right)}{\nu_{i \to (ij)}\left(\sigma_i = -1\right)} \tag{2.41}$$

Plugging in (2.41) into (2.40) leads to the following update equation for the fields $\omega_{i \to (ij)}$:

$$\omega_{i \to (ij)} = \beta h_i + \sum_{k \in \partial i \setminus j} \operatorname{atanh}\left[\tanh\beta J_{ik} \tanh\omega_{k \to (ik)}\right] \tag{2.42}$$

where the first term comes from the message $m_{\psi_i \to i}$, $\psi_i = e^{\beta h_i \sigma_i}$, if an external field is present. Analogously, the magnetization of the single-node belief is given by:

$$m_i = \langle \sigma_i \rangle_{q^{(i)}} = \tanh\left[\beta h_i + \sum_{k \in \partial i} \operatorname{atanh}\left[\tanh\beta J_{ik} \tanh\omega_{k \to (ik)}\right]\right] \tag{2.43}$$

[2] with respect to the disorder distribution

As a final remark, we briefly discuss the Bethe solution for the ferromagnetic Ising model defined on a hypercubic lattice in d dimension: in this case, Eqs. (2.42)-(2.43) can be further be simplified by exploiting the translational invariance of the model, so that all the local fields will be identical to a unique quantity $\omega_{i\to(ij)}\hat{=}\omega$, leading to:

$$\omega = (2d-1)\,\text{atanh}\left[\tanh\beta J \tanh\omega\right] \tag{2.44}$$

$$m = \tanh\left[2d\,\text{atanh}\left[\tanh\beta J \tanh\omega\right]\right] \tag{2.45}$$

where we set to zero the external field. The above equations always admit a paramagnetic solution with $\omega, m = 0$: however, this solution becomes unstable at a critical value of the temperature $\beta_c J = \text{atanh}\left[(2d-1)^{-1}\right]$, that can be computed by analyzing the stability of the fixed point equation (2.44). At $\beta > \beta_c$ the Bethe approximation predicts a ferromagnetic phase with $\omega, m\,(\omega) \neq 0$. The critical exponents are, however, the same as in the mean field theory: for this reason, the Bethe approximation is referred to as a *mean-field* like approach.

2.3 Beyond the Bethe Approximation

The Bethe approximation is constructed in such a way that the entropic term in the free energy takes into account exactly the contribution coming from each factor node together with its neighbours. It is possible to generalize such a construction to include larger regions of the graphs exactly: this procedure defines a class of approximation techniques known as Cluster Variational Methods, briefly discussed in the next subsection.

On the other hand, several attempts to improve the BP algorithm and its convergence properties on loopy graphs have been carried out, see for instance [62, 94, 147, 148]

Another way to go beyond the Bethe approximation is to include loop corrections. For instance, Rizzo and Montanari in [91] presented a loop corrected version of the Bethe approximation, that is exploited to compute a refined value to the critical temperature of the ferromagnetic Ising model, as well as on spin glass models on random graphs. A similar but more general approach on generic factor graphs was developed in [92] by Mooji and Kappen, known as Loop Corrected Belief Propagation (LCBP): the latter works by applying standard BP on the cavity graphs for each variable node, and then combining all the cavities together into another message-passing approach to estimate self-consistently the single-node marginals. This method turns out too be exact if the graph contains only one loop. We will come back to these two methods in Chapter 4 where a comparison with Density Consistency will be discussed.

Another series of works by Chertkov and Chernyak [30, 31] show how to express the partition function in terms of a infinite series, each term representing a loop contribution: in this perspective, the Bethe approximation is recovered as the leading term in their expansion.

Finally, a common issue with BP is that only nearest-neighbours correlations can be easily estimated: although one could include additional constraints in the original factor graph to estimate long-range correlations, it is often cumberstone to do so. A common approach to determine long-range correlation relies on Linear Response theory: in this perspective, a powerful approach was developed by Welling and Teh in [138, 139], known as Susceptibility Propagation (SP). We will come back to that in Chapter 5 for the Inverse Ising Problem.

2.3.1 Cluster Variational Method

The cluster variational method (CVM) is class of approximation schemes that generalizes the Bethe approximation by taking into account exactly the effect of short loops. Historically, this approach was first introduced by Kikuchi in [75] in the context of lattice ferromagnetic models; a more general formulation was developed by Yedidia and collaborators in [143, 146]. The key idea behind the CVM is that an approximate variational free energy can be constructed by summing a series of local contributions, each one corresponding to a different *region* of the graph: this procedure is typically referred to as *region-based* free energy construction. A region R is defined by the set of function nodes A_R and the variable nodes V_R such that, for each factor node $a \in A_R$, all its neighbours belong to V_R, namely $V_R = \{i \mid i \in \partial a, \forall a \in A_R\}$. To each region R we associate a (region) belief, denoted with q_R and defined over the set of degrees of freedom living inside region R, namely $\boldsymbol{\sigma}_R = \{\sigma_i, i \in V_R\}$. With these definitions, we can define the (functional) region energy U_R and entropy S_R, and free energy F_R, respectively:

$$U_R\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] = -\frac{1}{\beta}\sum_{\boldsymbol{\sigma}_R} q_R\left(\boldsymbol{\sigma}_R\right)\sum_{a\in A_R}\log\psi_a\left(\boldsymbol{\sigma}_a\right) \tag{2.46}$$

$$S_R\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] = -\sum_{\boldsymbol{\sigma}_R} q_R\left(\boldsymbol{\sigma}_R\right)\log q_R\left(\boldsymbol{\sigma}_R\right) \tag{2.47}$$

$$F_R\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] = U\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] - TS_R\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] \tag{2.48}$$

A factor graph can be covered by a set of regions, denoted with $\mathcal{R}_0$, such that each variable and factor node is included in at least one region $R \in \mathcal{R}_0$: moreover, regions must be chosen in such a way that no element in $\mathcal{R}_0$ is a subregion of any other one in the same set. The regions in $\mathcal{R}_0$ are called *maximal regions*, and they represent the largest contribution taken into account exactly into the Gibbs variational free energy. The intuitive idea is that, choosing larger sizes for the elements in $\mathcal{R}_0$ will give better free energy approximations. In order to define a proper region-graph free energy on $\mathcal{R}_0$, we should pay attention to the intersections between their elements. With this in mind, we define $\mathcal{R}_1$ as the set of all possible intersections between two regions in $\mathcal{R}_0$: the procedure can clearly be iterated, since two elements in $\mathcal{R}_1$ might have intersections as well. By iterating this procedure, one can define a set of regions $\mathcal{R}^k$ that are intersections of the elements in $\mathcal{R}^{k-1}$ for some $k > 1$. The CVM is then defined by the union of these regions, namely $\mathcal{P} = \cup_{k=0}^{n}\mathcal{R}^k$: in this notation, $\mathcal{R}^n$ is a set of disjoint regions, so that there is no intersection between any pair of its elements. The reason why it is necessary to take into account these intersection is that, in order to have a valid approximation to the free energy, each factor/variable node must be included exacly once, without overcounting. For each $R \in \mathcal{P}$, its contribution to the free energy (2.48) must be multiplied by an integer coefficient, known as *counting number* and denoted with c_R. The counting numbers must satisfy the following relations:

$$\sum_{R\in\mathcal{P}} c_R \mathbb{I}\left[a \in A_R\right] = 1 \qquad \forall a \in F \tag{2.49}$$

$$\sum_{R\in\mathcal{P}} c_R \mathbb{I}\left[i \in V_R\right] = 1 \qquad \forall i \in V \tag{2.50}$$

where $\mathbb{I}$ denotes the identity function of the condition given by its argument. The above procedure allows to define the free energy approximation on the region set $\mathcal{P}$, known as the *Kikuchi variational free energy*:

$$\mathcal{F}^{\text{Kikuchi}} = \sum_{R\in\mathcal{P}} c_R U_R\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] - T\sum_{R\in\mathcal{P}} c_R S_R\left[q_R\left(\boldsymbol{\sigma}_R\right)\right] \tag{2.51}$$

The counting numbers can be easily computed recursively. By construction, the maximal regions $R \in \mathcal{R}_0$ will have $c_R = 1$. Then, the counting numbers of their intersections (and so on) can be recursively computed using the Moebius formula [110]:

$$c_R = 1 - \sum_{R'\supset R} c_{R'} \tag{2.52}$$

where the summation runs over all the regions R' that *include* R (so that $R \in \mathcal{R}^k$ and $R' \in \mathcal{R}^{k-1}$). A simple example of two different constructions is shown in Figure (2.3) for a pairwise model with 6 nodes[3]: in particular, by choosing a set of maximal regions where each element includes only one factor node a, the Kikuchi free energy coincides with the Bethe Free energy (2.24). Conversely, by choosing squared plaquettes as maximal regions leads to a more refined approximation since the short loops in each plaquette are corretly taken into account.

The above construction can be in principle applied by choosing arbitrary large regions, but the computational cost to evaluate (2.51) increases exponentially with the size of maximal regions. The extreme limit corresponds to consider only one region $R^* \in \mathcal{R}_0$ that coincides with the full factor graph: in this case, $V_{R^*} = V$ and $A_{R^*} = F$, and the corresponding belief coincides with the true joint measure (2.1), so that the Gibbs variational free energy (2.3) is recovered.

The minimization of (2.51) can be carried out following a similar approach used in the previous section to derive the BP message-passing equations. In general, one can define a constrained free energy starting from (2.51) by adding a set of normalization constraints for each region's belief, plus a set of consistency conditions between beliefs of two regions $R \in \mathcal{R}^k$ and $R' \in \mathcal{R}^{k-1}$ such that R' is the smallest superset including R (in the literature, R' defined in such a way is typically called a *parent* of region R, and viceversa R is a children of R'):

$$q\left(\boldsymbol{\sigma}_R\right) = \sum_{\boldsymbol{\sigma}_{V'_R \setminus V_R}} q\left(\boldsymbol{\sigma}_{R'}\right) \qquad \forall R' \supset R \tag{2.53}$$

The constraints can be added to (2.51) by using a suitable set of Lagrange multipliers: however, there are in general different ways to enforce these constraints, leading to different message-passing schemes which are equivalent only at fixed point, but differ in the dynamical update rules and/or convergence properties. It is out of the scope of this thesis to review them, and we refer to [143] for a more detailed discussion about this issue.

It is straightforward to show that, when the regions are chosen in such a way to include at most one factor node a, (2.51) reduces to the Bethe free energy, and the corresponding message-passing equations are the same as BP. For this reason, the message-passing equations derived from the minimization of the constrained Kikuchi free energy are typically referred to as Generalized Belief Propagation (GBP) equations , being equivalent to BP whenever the maximal regions include single factor nodes. GBP equations can be proven to give an exact estimation of the region beliefs whenever the graph topology contains loops only inside the maximal regions: for instance, for a pairwise model defined on a ladder system of two coupled linear chains with open boundary conditions, the resulting CVM with plaquette maximal regions (as in Figure 2.3) is exact [110]. In a similar but more general spirit, Cantwell and Newman [26, 76] proposed a generalized message-passing scheme to deal with loopy networks, by defining a series of approximations in which the correlations induced by loops of length $r + 2$ (or lower, for a certain positive integer r) is exactly taken into account: in this perspective, the $r = 0$ case corresponds to standard BP, and the computational cost of the iterative scheme grows exponentially with r. With respect to CVM, the main advantage of this method is that it does not need an explicit construction of region graph covering and it easily adapts to arbitrary topologies. On the other hand, analogously to the CVM, contributions coming from loops outside the maximal regions (or longer than $r + 2$) are not taken into account.

[3] The topology of Fig. 2.3 topology is called *ladder*, and it is composed by two linear chains with transverse interaction between adjacent sites in the two chains

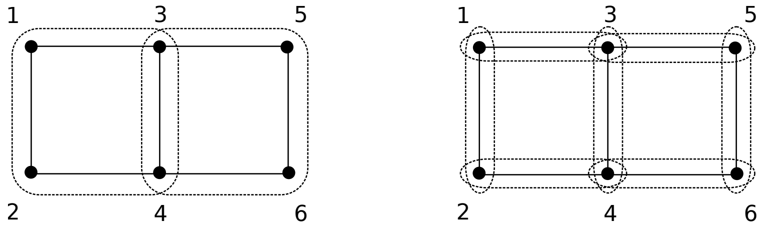

Figure 2.3: Construction of the maximal region graphs onto a ladder of 6 spins. Left: maximal regions correspond to square plaquettes. In this case, $\mathcal{R}_0 = \{\square_1, \square_2\}$ where $V_{\square_1} = \{1{,}2{,}3{,}4\}$,$V_{\square_2} = \{3{,}4{,}5{,}6\}$. Their intersection set is $\mathcal{R}_1 = \{(3{,}4)\}$, containing only 1 link. The corresponding counting numbers are $c_{\square_1} = c_{\square_2} = 1$ and $c_{(34)} = -1$. Right: maximal regions correspond to edges. In this case $\mathcal{R}_0 = E$ (i.e. the set of edges) and $\mathcal{R}_1 = V$ (set of nodes). The counting numbers are $c_l = 1$ for $l \in E$ and $c_i = 1 - d_i$ for $i \in V$, where d_i is node i's degree. The latter construction corresponds to the Bethe approximation.

2.4 Expectation Propagation

Expectation Propagation (EP) denotes a family of approximation schemes introduced by Minka in [90] to perform approximate inference in high-dimensional systems; at the same time, a very similar approximation was discovered by Opper and Winther in the statistical physics community [104] known as Expectation Consistency, based on the Adaptative-TAP method [103]. The main idea behind EP is to approximate an *intractable* - i.e. non integrable - multivariate probability distribution with a tractable family, whose parameters are fixed in such a way to satisfty local moment matching conditions. Expectation Propagation can be defined over generic probabilistic graphical models expressed in terms of an exponential family. However, in the following we will first focus on the simplest setup where the tractable family of approximation is Gaussian. In this way, the connection to Density Consistency discussed in Chapter 3 will be more evident. In this context, EP can be used to approximate marginals of a probability distribution written as the product of a multivariate Gaussian, denoted with $g(\boldsymbol{x})$, times a set of single-site functions ψ_i:

$$p(\boldsymbol{x}) = \frac{1}{Z} g(\boldsymbol{x}) \prod_{i=1}^{N} \psi_i(x_i) \tag{2.54}$$

In this setting, the variables x_i are continuous, i.e. $\boldsymbol{x} \in \mathbb{R}^N$. Graphical models of this type arise in many contexts of both statistical physics and inference. As a simple example, notice that the equilibrium distribution of an Ising model can always be written in the form (2.54), by including the Boltzmann weight into $g(\boldsymbol{x})$ and defining each function $\psi_i(x_i)$ as a combination of Dirac's deltas, in order to enforce the constraints that variables have to be defined over $\{-1{,}1\}$:

$$g(\boldsymbol{x}) = \exp\left[\beta \sum_{(ij)\in E} J_{ij} x_i x_j + \beta \sum_i h_i x_i\right] \tag{2.55}$$

$$\psi_i(x_i) = \frac{1}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right] \tag{2.56}$$

With the above identifications, from (2.54) one recovers exactly the Boltzmann law (1.14). On the other hand, the above parametrization can be used to describe a wide class of Bayesian inference problems known as Linear Estimation Problems.

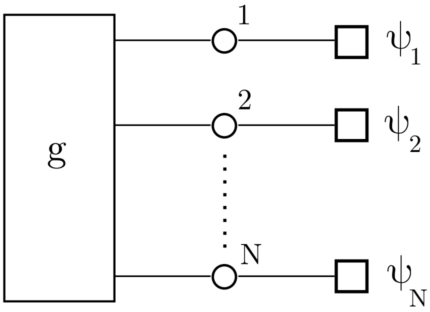

Figure 2.4: Factor graph representation of the distribution (2.54)

Linear estimation problems

Linear estimation problems (LEPs) arise in many research fields, such as theoretical computer science, signal analysis and computational biology. In general, linear estimation problems attempt to solve an under-determined linear system of equations in the form $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$, with $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{F} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{x} \in \mathbb{R}^N$: the vector $\boldsymbol{x}$ is typically referred to as an unknown input (source) signal, and $\boldsymbol{y}$ is a known output vector. When $M < N$, i.e. when the number of measurements (encoded in the vector $\boldsymbol{y}$) is lower than the number of unknowns (encoded in $\boldsymbol{x}$), there is in principle an infinite number of solutions to the linear system. Therefore, one needs to impose additional constraints to find a particular set of solutions, depending on the problem under investigation. In a Bayesian framework, such constraints can be enforced by adding suitable prior distributions on the unknown source vector components. First notice that any linear system of equations can be represented by a constraint of the type $\delta\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\right)$, where $\delta\left(x\right)$ is the Dirac's delta function. By using a Gaussian representation of the delta function, we get:

$$\delta\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\right) = \lim_{\Delta\to 0} \exp\left[-\frac{1}{\Delta}\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\right)^t\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\right)\right] \tag{2.57}$$

The parameter Δ plays the role of a noise added to the measurements (it can also be considered as a fictious temperature): the scenario with finite Δ can therefore be used whenever the original linear system is affected by the presence a Gaussian noise, namely $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{\varepsilon}$ where all the components of the noise-vector $\boldsymbol{\varepsilon}$ are i.i.d. Gaussian variables, $\varepsilon_i \sim N\left(0, \Delta\right)$. For instance, in the Compressed Sensing Problem [44] one attempts in finding a solution $\boldsymbol{x}^*$ where a fraction of the source components are zero: this information can be enforced by adding a sparsity prior ℓ^0, so that the overall (posterior) probabilty distribution over the set of solutions can be written as:

$$p\left(\boldsymbol{x} \mid \boldsymbol{y}\right) \propto \exp\left[-\frac{1}{\Delta}\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\right)^t\left(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\right)\right]\prod_i \psi_i\left(x_i\right) \tag{2.58}$$

$$\psi_i\left(x_i\right) = \left(1-\rho\right)\delta\left(x_i\right) + \rho N\left(x_i; 0, \sigma\right) \tag{2.59}$$

where $N\left(x\right)$ is a short-hand notation for a Gaussian density. In this setting, (2.57) plays the role of a likelihood function, and the prior ψ_i enforces an a-priori knowledge such that a fraction $1-\rho$ of the source components must be 0. In the CS problem, computing marginal probabilities over 2.58 allows to select relevant solutions where a fraction of the $\boldsymbol{x}$ components is 0 given the observed signal $\boldsymbol{y}$. Whenever ψ_i differs from a Gaussian density, an exact integration of (2.58) cannot be carried out. In this perspective, Expectation Propagation can be implemented to compute marginal probabilities of (2.58), and several applications have been recently carried out in our group: for the already cited Compressed Sensing problem [23], inference in metabolic networks (where, in particular, the matrix $\boldsymbol{F}$ is related to the stochiometric matrix describing the chemical reactions inside a cell) [22] and in tomographic images [97]. In the next section, we discuss the EP scheme in the simple setting where the approximating family is chosen to be a normal distribution.

2.4.1 EP algorithm

Let us start from the intractable probability distribution defined in (2.54), whose factor graph representation is shown in Figure 2.4. Expectation Propagation works by replacing each term ψ_i with a univariate Gaussian density, denoted with ϕ_i, and parametrized as follows:

$$\phi_i\left(x_i\right) = \exp\left[-\frac{1}{2}\Gamma_i x_i^2 + \gamma_i x_i\right] \tag{2.60}$$

The set $\{(\lambda_i, \Gamma_i)\}_{i=1,\dots,N}$ defines the ensemble of parameters encoded by EP. By replacing each prior ψ_i with (2.60), it is possible to construct a multivariate Gaussian density, denoted with q, that approximates the starting probability distribution:

$$q\left(\boldsymbol{x}\right) = \frac{1}{Z_q} g\left(\boldsymbol{x}\right) \prod_i \phi_i\left(x_i\right) \tag{2.61}$$

The constant Z_q and the moments of 2.61 $\langle \boldsymbol{x} \rangle_q = \boldsymbol{\mu}$, $\langle \boldsymbol{x}\boldsymbol{x}^t \rangle_q = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t$ can be computed by standard Gaussian integration. The factor graph representation of 2.61 is shown in the left plot of Figure 2.5. For each variable we can also define a *tilted* (or *leave-one-out*) distribution that is obtained from (2.61) by replacing the approximate factor ϕ_i with the true prior ψ_i:

$$q^{(i)}\left(\boldsymbol{x}\right) = \frac{1}{Z_i} g\left(\boldsymbol{x}\right) \psi_i\left(x_i\right) \prod_{j \neq i} \phi_j\left(x_j\right) \tag{2.62}$$

$$\propto q\left(\boldsymbol{x}\right) \frac{\psi_i\left(x_i\right)}{\phi_i\left(x_i\right)} \tag{2.63}$$

The idea behind EP is that the tilted distribution $q^{(i)}$ can be used as a tractable estimator for the single-node marginal over i. Indeed, by construction, all the other variables appearing in (2.62) can be marginalized by using standard Gaussian integration. As a consequence, the single-site marginal $q^{(i)}\left(x_i\right)$ can be written as the product of the prior ψ_i times a univariate Gaussian distribution, denoted with $g^{\setminus i}\left(x_i\right)$:

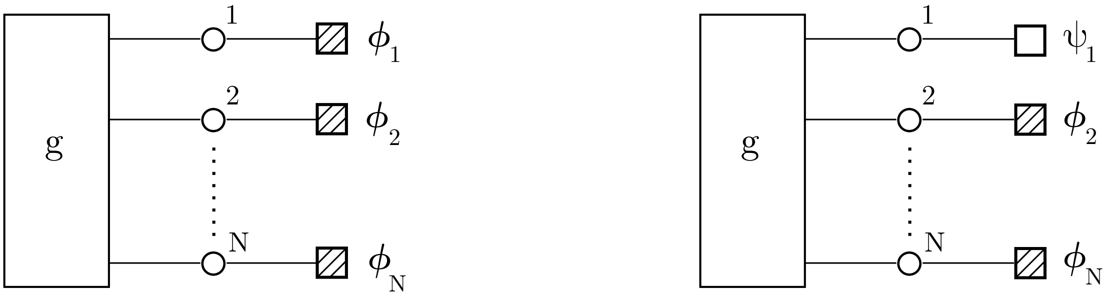

Figure 2.5: Factor graph representation of the "full" Gaussian measure (2.61) (left) and the tilted distribution (2.62) over node 1 (right).

$$q^{(i)}\left(x_i\right) = \int d\boldsymbol{x}_{\setminus i} q^{(i)}\left(\boldsymbol{x}\right) \propto \psi_i\left(x_i\right) \int d\boldsymbol{x}_{\setminus i} g\left(\boldsymbol{x}\right) \prod_{j \neq i} \phi_j\left(x_j\right) = \frac{1}{\tilde{z}_i} \psi_i\left(x_i\right) g^{\setminus i}\left(x_i\right) \tag{2.64}$$

The quantity $g^{\setminus i}(x_i)$ is typically referred to as a cavity distribution, and it is obtained by removing one single factor ϕ_i and marginalizing over all variables but i:

$$g^{\setminus i}(x_i) \propto \int d\boldsymbol{x}_{\setminus i} \frac{q(\boldsymbol{x})}{\phi_i(x_i)} \propto \frac{q(x_i)}{\phi_i(x_i)} \tag{2.65}$$

The above expression allows to write in an equivalent way the single-node marginal of the full Gaussian measure:

$$q(x_i) = \frac{1}{z_i}\phi_i(x_i)\, g^{\setminus i}(x_i) \tag{2.66}$$

Note that (2.64) and (2.66) only differ by the choice of the univariate factor, which is taken exactly (ψ_i) in the first equation and approximately (ϕ_i) in the second. It is therefore natural to choose a suitable set of conditions between these two marginals in order to determine the parameters of ϕ_i, namely (γ_i, Γ_i). In particular, within the EP scheme, these are fixed by imposing local *moment matching* conditions on the first two order statistics:

$$\langle x_i \rangle_{q^{(i)}} = \langle x_i \rangle_q = \mu_i \quad \forall i \tag{2.67}$$

$$\langle x_i^2 \rangle_{q^{(i)}} = \langle x_i^2 \rangle_q = \Sigma_{ii} + \mu_i^2 \quad \forall i \tag{2.68}$$

to be solved w.r.t. (γ_i, Γ_i). Alternatively, it is straightforward to show that the above moment matching conditions (2.67)-(2.68) can be obtained by a local minimization of the Kullback-Leiber divergence between the tilted distribution $q^{(i)}$ and the full Gaussian measure q:

$$\frac{\partial D_{KL}\left(q^{(i)} \,||\, q\right)}{\partial \gamma_i} = 0 \longrightarrow \langle x_i \rangle_q = \langle x_i \rangle_{q^{(i)}} \tag{2.69}$$

$$\frac{\partial D_{KL}\left(q^{(i)} \,||\, q\right)}{\partial \Gamma_i} = 0 \longrightarrow \langle x_i^2 \rangle_q = \langle x_i^2 \rangle_{q^{(i)}} \tag{2.70}$$

where D_{KL} is defined in (2.10), with a proper replacement of the sums with integrals over $x_i \in \mathbb{R}$. The set of equations (2.67)-(2.68) give a closed system to be solved iteratively w.r.t. $\{(\lambda_i, \Gamma_i)\}_{i=1,\ldots,N}$. In the simplest update scheme where all the parameters are updated simultaneously at each iteration (parallel update) the computational cost scales as $O\left(N^3\right)$, dominated by the inversion $\boldsymbol{\Sigma}$, i.e. the covariance matrix of (2.61). In the next section, we discuss how to derive the EP free energy, whose derivation is similar to the preliminary result presented in Section 3.4.2 for the Density Consistency free energy.

2.4.2 EP free energy

The following derivation of the EP free energy (discussed in [23]) comes from analogy with Belief Propagation. On a generic graphical model given by (2.1), the starting (intractable) density can be written using the Bethe factorization as:

$$p(\boldsymbol{\sigma}) = \frac{Z_{BA}}{Z}\hat{p}(\boldsymbol{\sigma}) \tag{2.71}$$

where $-\log Z_{BA}$ is the Bethe free energy expressed in terms of messages (2.34) and $\hat{p}$ is given by (2.21). Such relation holds on any loopy graph; however, if BP is exact (i.e. on a tree) the starting probability measure factorized exactly as in (2.21), so that $Z_{BA} = Z$. Following the same reasoning, we now derive the EP free energy by writing the starting probability density (2.54) in terms of the tilted distributions and the full Gaussian measure. Let us start from (2.63) and rewrite the tilted distirbution $q^{(i)}$ as follows:

$$\begin{aligned} q^{(i)}\left(\boldsymbol{x}\right) &= \frac{1}{Z_i} g\left(\boldsymbol{x}\right)\psi_i\left(x_i\right)\prod_{j\neq i}\phi_j\left(x_j\right) \\ &= \frac{Z_q}{Z_i} q\left(\boldsymbol{x}\right)\frac{\psi_i\left(x_i\right)}{\phi_i\left(x_i\right)} \end{aligned} \tag{2.72}$$

Then, by taking the product of all the tilted distributions, after some manipulations we get:

$$\begin{aligned} \prod_i q^{(i)}\left(\boldsymbol{x}\right) &= \prod_i \frac{Z_q}{Z_i} q\left(\boldsymbol{x}\right)\frac{\psi_i\left(x_i\right)}{\phi_i\left(x_i\right)} \\ &= \frac{Z_q^N}{\prod_i Z_i} q^N\left(\boldsymbol{x}\right)\frac{\prod_i \psi_i\left(x_i\right)}{\prod_i \phi_i\left(x_i\right)} \\ &= \frac{Z_q^N}{\prod_i Z_i} q^N\left(\boldsymbol{x}\right)\frac{Zp\left(\boldsymbol{x}\right)}{Z_q q\left(\boldsymbol{x}\right)} \\ &= \frac{Z_q^{N-1}}{\prod_i Z_i} q^{N-1}\left(\boldsymbol{x}\right) Zp\left(\boldsymbol{x}\right) \end{aligned} \tag{2.73}$$

where in the last line we explicity show the dependency over the true distribution p. Let us now rewrite the last expression as

$$p\left(\boldsymbol{x}\right) = \frac{Z_{EP}}{Z}\hat{p}\left(\boldsymbol{x}\right) \tag{2.74}$$

where

$$\hat{p}\left(\boldsymbol{x}\right) = \frac{\prod_i q^{(i)}\left(\boldsymbol{x}\right)}{q^{N-1}\left(\boldsymbol{x}\right)}; \qquad Z_{EP} = \frac{\prod_i Z_i}{Z_q^{N-1}} \tag{2.75}$$

By analogy with the BP case discussed at the beginning, whenever EP is exact, $Z_{EP} = Z$ and $p\left(\boldsymbol{x}\right) = \hat{p}\left(\boldsymbol{x}\right)$. In this setting - i.e. where the approximating family is Gaussian - EP becomes exact when the priors ψ_i are Gaussian distributed. In this case, the EP scheme is trivially solved by identifying $\psi_i \equiv \phi_i$; as a consequence, the starting density p, the full Gaussian measure q and all the tilted distributions will be identically equal (namely $p = q = q^{(i)}$ and $Z = Z_q = Z_i\ \forall i$) so that $\hat{p}\left(\boldsymbol{x}\right) = p\left(\boldsymbol{x}\right)$ and $Z_{EP} = Z$. We can now define the EP free energy as:

$$F_{EP} = -\log Z_{EP} = \left(N-1\right)\log Z_q - \sum_i \log Z_i \tag{2.76}$$

Its stationary points can be obtained by deriving (2.76) w.r.t. the gaussian parameters λ_i, Γ_i, leading to:

$$\frac{\partial F_{EP}}{\partial \gamma_i} = \left(N-1\right)\langle x_i\rangle_q - \sum_{j\neq i}\langle x_i\rangle_{q^{(j)}} \tag{2.77}$$

$$\frac{\partial F_{EP}}{\partial \Gamma_i} = \left(N-1\right)\langle x_i^2\rangle_q - \sum_{j\neq i}\langle x_i^2\rangle_{q^{(j)}} \tag{2.78}$$

The right hand sides of (2.77)-(2.78) depend on the moments of a spin i computed w.r.t. to a tilted distribution defined on another variable $j \neq i$. These “mixed” tilted moments are not the ones used by the algorithm to update the parameters through the moment matching conditions. However, it still possible to explicitly compute them by exploiting Gaussian integration properties, as discussed in Appendix A: in this way, it is easy to prove that EP fixed point equations (2.67)-(2.68) satisfy the above relations. The EP free energy can be used also to on-line learn the parameters encoded in the priors ψ_i, by using an expectation-maximization (EM) procedure [23].

2.4.3 Relation between EP and BP

On a discrete graphical models (2.1), it has been shown that Belief Propagation corresponds to a specific instance of Expectation Propagation, when the approximating family q is chosen to be a fully-factorized distribution over single nodes [90]. In order to prove it, let us start from the probability distribution (2.1), and replace each function node ψ_a with a factorized *discrete* distribution over single-node functions, denoted with ϕ_a:

$$\phi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \propto \prod_{i\in\partial a} m_{ai}\left(\sigma_i\right) \tag{2.79}$$

where m_{ai} is the marginal of ϕ_a over σ_i, by construction. For simplicity, we will assume that the degrees of freedom are binary spins, namely $\sigma_i \in \{-1{,}1\}$, even if the same reasoning can be extended to arbitrary discrete variables (e.g. Potts-like). The tractable joint distribution encoded by EP can be constructed by taking the product of all the approximate factors ϕ_a, similarly to 2.61:

$$q\left(\boldsymbol{\sigma}\right) \propto \prod_a \phi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \propto \prod_i \left[\prod_{a\in\partial i} m_{ai}\left(\sigma_i\right)\right] \tag{2.80}$$

The last equality (apart from a normalization factor) states that $q\left(\boldsymbol{\sigma}\right)$ is factorized over single nodes, so that single-node marginals can be computed straightforwardly. The tilted distributions can now be defined by removing one factor node ϕ_a from (2.80) and replacing it with the true factor ψ_a:

$$q^{(a)}\left(\boldsymbol{\sigma}\right) \propto q\left(\boldsymbol{\sigma}\right)\frac{\psi_a\left(\boldsymbol{\sigma}_{\partial a}\right)}{\phi_a\left(\boldsymbol{\sigma}_{\partial a}\right)} \propto \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right)\prod_{b\neq a}\prod_{i\in\partial b} m_{bi}\left(\sigma_i\right) \tag{2.81}$$

Eqs (2.80)-(2.81) can be easily marginalized over the neighbours of a since all the other spins' contributions are factorized:

$$q^{(a)}\left(\boldsymbol{\sigma}_{\partial a}\right) = \sum_{\boldsymbol{\sigma}_{\setminus\partial a}} q^{(a)}\left(\boldsymbol{\sigma}\right) \propto \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right)\prod_{j\in\partial a}\prod_{b\in\partial j\setminus a} m_{bj}\left(\sigma_j\right) \tag{2.82}$$

$$q\left(\boldsymbol{\sigma}_{\partial a}\right) = \sum_{\boldsymbol{\sigma}_{\setminus\partial a}} q\left(\boldsymbol{\sigma}\right) \propto \prod_{j\in\partial a}\prod_{b\in\partial j} m_{bj}\left(\sigma_j\right) \tag{2.83}$$

As discussed in the previous section, the EP scheme works by imposing moment matching condition between the two above marginals. First notice that, since each of them is a discrete distribution, imposing moment matching on single-node marginals is equivalent to impose that the two marginals are proportional, apart from a normalization factor:

$$\langle\sigma_i\rangle_{q^{(a)}} = \langle\sigma_i\rangle_q \qquad \Longleftrightarrow q^{(a)}\left(\sigma_i\right) \propto q\left(\sigma_i\right) \tag{2.84}$$

where $q^{(a)}\left(\sigma_i\right)$ is the tilted marginal over σ_i (and the same holds for $q\left(\sigma_i\right)$). Further notice that there is no need of imposing the matching of second-order moments as in the Gaussian EP scheme previously discussed: the reason is that a single-node distribution of a binary variable is uniquely defined by a single parameter (on the contrary, an univariate Gaussian has two sufficient statistics, namely its mean and variance). We first rewrite explicitly the single-node marginals of $q^{(a)}$and q

using their definitions (2.83)-(2.82):

$$\begin{aligned} q^{(a)}\left(\sigma_i\right) &\propto \sum_{\boldsymbol{\sigma}_{\partial a\setminus i}} \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \prod_{j\in\partial a}\prod_{b\in\partial j\setminus a} m_{bj}\left(\sigma_j\right) \\ &\propto \left[\prod_{b\in\partial i\setminus a} m_{ai}\left(\sigma_i\right)\right] \sum_{\boldsymbol{\sigma}_{\partial a\setminus i}} \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \prod_{j\in\partial a\setminus i}\prod_{b\in\partial j\setminus a} m_{bj}\left(\sigma_j\right) \\ q\left(\sigma_i\right) &\propto \prod_{a\in\partial i} m_{ai}\left(\sigma_i\right) \end{aligned}$$

Finally, using the above formulas and imposing (2.84), we get:

$$m_{ai}\left(\sigma_i\right) \propto \sum_{\boldsymbol{\sigma}_{\partial a\setminus i}} \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \prod_{j\in\partial a\setminus i}\prod_{b\in\partial j\setminus a} m_{bj}\left(\sigma_j\right)$$

The above expression coincides with the BP update equations for the messages, obtained by inserting (3.43) into (3.42), where the quantity m_{ai} is recognized as the factor-to-node BP message $m_{a\to i}\left(\sigma_i\right)$. As a final remark, we recall the definition of the EP cavity (2.65) distribution, that in the present framework is given by:

$$g^{\setminus a}\left(\boldsymbol{\sigma}\right) \propto \frac{q\left(\boldsymbol{\sigma}\right)}{\phi_a\left(\boldsymbol{\sigma}_{\partial a}\right)} \propto \prod_{b\neq a}\prod_{i\in\partial b} m_{bi}\left(\sigma_i\right) \tag{2.85}$$

The above expression is factorized over single nodes. In particular, its marginal over node a's neighbours can be written as:

$$g^{\setminus a}\left(\boldsymbol{\sigma}_{\partial a}\right) = \sum_{\boldsymbol{\sigma}_{\setminus\partial a}} g^{\setminus a}\left(\boldsymbol{\sigma}\right) \propto \prod_{i\in\partial a}\prod_{b\in\partial i\setminus a} m_{bi}\left(\sigma_i\right)$$

which corresponds exacly to the Bethe ansatz for the cavity distribution (2.38). Therefore, the above results provide a mapping between the Belief Propagation approach and Expectation Propagation. Note also that, by applying the same reasoning used in Section 2.4.2 to derive the EP free energy in this context, one recovers exactly the Bethe free energy. This relation will be further highlighted in the next Chapter, where the connection between BP, EP and Density Consistency will be discussed.

Chapter 3

Density Consistency

This chapter represents the main core of the manuscript, where we derive the Density Consistency scheme and analyze its properties. The method is constructed as a generalization of both the Belief Propagation and Expectation Propagation algorithms described in the previous Chapter, with the peculiar property to be exact on acyclic graphs. In particular, Section 3.1 presents the derivation for generic probabilistic graphical models of binary degrees of freedom. Section 3.2 discusses the main properties of Density Consistency, namely its exactness on trees and its relation to Belief Propagation fixed points, as well as the connection to Expectation Propagation. Section 3.3 discusses some algorithmic details and a pseudocode implementation. Finally, in Section 3.4 we present a possible generalization to non-binary degrees of freedom and a preliminar variational formulation.

3.1 Derivation

In this section, we are going to derive the Density Consistency approximation for generic probabilistic graphical models, using the factor graph representation introduced in section 1.2.1. We start by stating the problem addressed in the whole chapter, i.e. the computation of marginal distributions from a probabilistic graphical model defined by a density $p(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma} = \{\sigma_i\}_{i \in V}$ is the vector of degrees of freedom. The model is defined on a factor graph $G = (V, F, E)$ of $N = |V|$ variable nodes and $M = |F|$ factor nodes: each variable node represents a degree of freedom, and we will restrict for the rest of the chapter to binary (or Ising) spins, i.e. $\sigma_i \in \{-1,1\}$; further generalization to arbitray binary supports $\{a, b\}$ and non-binary variables will be discussed in Sec. 3.4.1. Each edge $(i, a) \in E$ connects a variable node to a factor node, so that the overall graph is bipartite w.r.t. $V \cup F$. Each factor node $a \in F$ is associated to a certain non-negative function $\psi_a(\boldsymbol{\sigma}_{\partial a})$ that depends on all the variable nodes i in the neighborhood of a. The probability distribution of such a model, denoted with $p(\boldsymbol{\sigma})$, is given by:

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\boldsymbol{\sigma}_a). \tag{3.1}$$

where the vector $\boldsymbol{\sigma}_a$ is a short-hand notation of $\boldsymbol{\sigma}_{\partial a} = \{\sigma_i, i \in \partial a\}$, and it will be kept for the rest of the discussion. The prefactor in Eq. (3.1) is the inverse of the partition function Z:

$$Z = \sum_{\boldsymbol{\sigma}} \prod_{a \in F} \psi_a(\boldsymbol{\sigma}_a) \tag{3.2}$$

where the summation runs over all the $\{-1,1\}^N$ spin configurations. A toy example is shown in Figure 3.1, where black dots identify the discrete variables σ_i (one for each node), and each white square represents a factor node $a \in F$. As already discussed, the computation of marginal distributions (or equivalently the partition function Z) has an exponential computational cost w.r.t. to the system size N. For convenience with the following discussion, we rewrite Eq. (3.1)

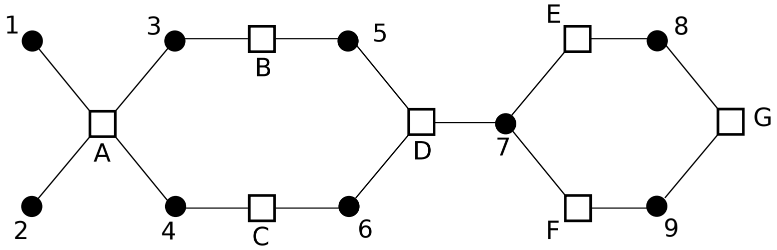

Figure 3.1: Factor graph representation of a toy model of the type (3.1) with 9 variable nodes and 7 factor nodes.

as a distribution of *real* variables, denoted with $x_i \in \mathbb{R}$:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\boldsymbol{x}_a) \prod_i \Delta_i(x_i), \tag{3.3}$$

where again $\boldsymbol{x}_a = \boldsymbol{x}_{\partial a}$ and

$$\Delta_i(x_i) = \frac{1}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right], \tag{3.4}$$

and $\delta(x)$ is the Dirac's Delta Function. The functions $\Delta_i(x_i)$ ensure that (3.3) is correctly defined on the support $\{-1{,}1\}^N$, so that there is no difference between Eq. (3.3) and (3.1). In the following, we will use the short notation $\Delta_i(x_i) \equiv \Delta_i$.

In the same spirit as for the Expectation Propagation algorithm described in Section 2.4, we will approximate the intractable distribution $p(\boldsymbol{x})$ with a family of Gaussian densities, whose parameters will be determined iteratively by imposing local consistency condition on the marginals. To this aim, we approximate each compatibility function $\psi_a(\boldsymbol{x}_a)$ in (3.3) with a multivariate Gaussian distribution, denoted by $\phi_a(\boldsymbol{x}_a)$ and parametrized as follows:

$$\phi_a(\boldsymbol{x}_a) = \exp\left[-\frac{1}{2}\boldsymbol{x}_a^t \boldsymbol{\Gamma}^{(a)} \boldsymbol{x}_a + \boldsymbol{x}_a^t \boldsymbol{\gamma}^{(a)}\right] \qquad \forall a \in F. \tag{3.5}$$

where the superscript $\cdot^t$ denotes the transpose vector. In this notation, each Gaussian density $\phi_a(\boldsymbol{x}_a)$ is parametrized by a vector $\boldsymbol{\gamma}^{(a)} \in \mathbb{R}^{|\partial a|}$, and by a symmetric matrix $\boldsymbol{\Gamma}^{(a)} \in \mathbb{R}^{|\partial a| \times |\partial a|}$. In a statistical mechanics jargon, the quantities $\left\{\gamma_i^{(a)}\right\}_{i \in \partial a}$ act as local fields on each variable and we refer to them as a *Gaussian fields*. The matrix $\boldsymbol{\Gamma}^{(a)}$ is the precision matrix of the Gaussian measure (3.5), and it is the inverse of its covariance matrix: it encodes a set of self-couplings $\Gamma_{ii}^{(a)}$as well as approximate pairwise (quadratic) interactions $\left\{\Gamma_{ij}^{(a)}\right\}_{i \neq j}$ between all the pair of nodes in the neighborhood of a. In this perspective, an equivalent parametrization can be constructed by defining each factor ϕ_a in terms of its first and second moments, respectively denoted with $\boldsymbol{\mu}^{(a)}$, $\boldsymbol{\Sigma}^{(a)}$; the mapping between the two parametrizations is given by:

$$\boldsymbol{\Sigma}^{(a)} = \left[\boldsymbol{\Gamma}^{(a)}\right]^{-1}, \qquad \boldsymbol{\mu}^{(a)} = \boldsymbol{\Sigma}^{(a)} \cdot \boldsymbol{\gamma}^{(a)} \tag{3.6}$$

where the superscript $^{-1}$ denotes the matrix inversion and the dot symbol $\cdot$ denotes the matrix-vector product. However, in the rest of the discussion we will use the parametrization of ϕ_a in terms of linear and quadratic terms as in (3.5). Therefore, Density Consistency will be defined by the set of approximate Gaussian factors $\{\phi_a\}_{a\in F}$. Since the product of Gaussian densities results in another normal distribution (i.e. the Gaussian family is closed under the product of their densities), by taking the product of all the $\{\phi_a\}_{a\in F}$ it is possible to construct a multivariate Gaussian distribution over the full set of N variables:

$$q\left(\boldsymbol{x}\right) \propto \prod_{a\in F} \phi_a\left(\boldsymbol{x}_a\right) = \frac{1}{Z_q} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right]. \tag{3.7}$$

In the last equality, the quantities $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ denote the mean vector and the covariance matrix over the measure (3.7), respectively:

$$\boldsymbol{\mu} = \langle \boldsymbol{x} \rangle_q, \tag{3.8}$$

$$\boldsymbol{\Sigma} = \langle \boldsymbol{x}\boldsymbol{x}^t \rangle_q - \langle \boldsymbol{x} \rangle_q \langle \boldsymbol{x}^t \rangle_q. \tag{3.9}$$

The covariance matrix $\boldsymbol{\Sigma}$ is symmetric by construction and it has to be positive definite in order for (3.7) to be defined. The quantity Z_q in (3.7) denotes the normalization factor of the distribution:

$$Z_q = \sqrt{\left(2\pi\right)^N \log\left|\det \boldsymbol{\Sigma}\right|} \tag{3.10}$$

The Gaussian moments (3.8)-(3.9) can be easily computed starting from the set of parameters $\left\{\boldsymbol{\gamma}^{(a)}, \boldsymbol{\Gamma}^{(a)}\right\}_{a\in F}$ encoded in each factor ϕ_a. By construction of $q\left(\boldsymbol{x}\right)$, the following relations hold:

$$\left(\boldsymbol{\Sigma}^{-1}\right)_{ij} = \begin{cases} \sum_{\substack{a\in F\\ i,j\in\partial a}} \Gamma_{ij}^{(a)} & i \neq j \\ \sum_{\substack{a\in F\\ i\in\partial a}} \Gamma_{ii}^{(a)} & i = j \end{cases} \qquad \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)_i = \sum_{\substack{a\in F\\ i\in\partial a}} \gamma_i^{(a)}. \tag{3.11}$$

In this way, the Gaussian moments $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ can be constructed by inverting (3.11) at given parameters $\left\{\boldsymbol{\gamma}^{(a)}, \boldsymbol{\Gamma}^{(a)}\right\}_{a\in F}$. Notice that (3.7) encodes the same factorization of the original probability distribution (3.1), and indeed the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ has the same structure of the starting factor graph: in particular, an off-diagonal element $\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}$ will be non-zero if nodes i and j are both connected to (at least) one factor node. As an example, Figure 3.2 shows the factor graph representation of (3.7) for the toy model in Figure 3.1. Since Gaussian densities are defined

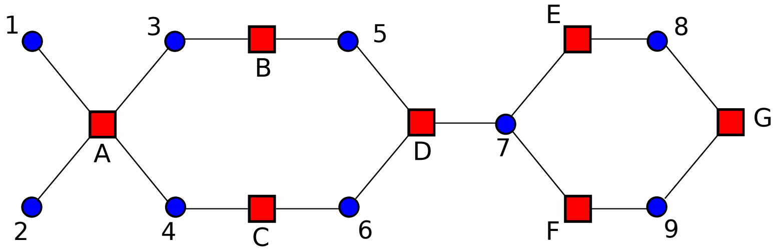

Figure 3.2: Factor graph representation the multivariate Gaussian distribution 3.7 referred to the toy model in Figure 3.1. Blue dots refer to nodes with continuous (real) variable x_i and the red squares represent approximate Gaussian factors ϕ_a.

over continuous degrees of freedom, in Figure 3.2 variable nodes are represented by blue dots, to avoid confusion with previously defined discrete variables; on the other hand, the approximate factors ϕ_a are shown as red squares, a notation that will be useful to understand the following definitions. It is important to remark that that the Gaussian parametrization takes into account only pairwise effective interactions: neverthless, we will keep using the factor representation of the Gaussian measure as in Figure (3.2) to avoid confusion. This simple equivalence is shown in Figure 3.3 for the 4−body factor in the toy model of Figure 3.1.

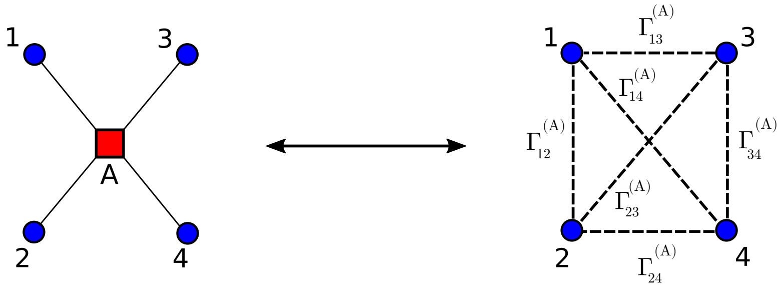

Figure 3.3: Representation of the approximate factor A of Figure 3.2 in terms of pairwise interactions. For simplicity, the diagonal terms of the precision matrix $\Gamma_{ii}^{(A)}$ and linear terms $\gamma_i^{(A)}$ $(i \in \partial A)$ are not shown.

The set $\left\{\boldsymbol{\gamma}^{(a)}, \boldsymbol{\Gamma}^{(a)}\right\}_{a\in F}$ is the ensemble of parameters encoded by Density Consistency, that need to be determined by an appropriate iterative scheme proposed in the following. The total number of parameters is $\sum_{a\in F} |\partial a| \left(|\partial a| + 3\right)/2$: indeed, for each Gaussian factor (3.5), there are $|\partial a|$ linear terms (Gaussian fields), $|\partial a|$ diagonal entries for the precision matrix $\boldsymbol{\Gamma}^{(a)}$ (also called self-couplings), and $\binom{|\partial a|}{2} = |\partial a| \left(|\partial a| - 1\right)/2$ non diagonal terms (couplings).

3.1.1 Tilted distributions

Since we are interested in computing marginal distributions over factor nodes, a more refined approximation can be obtained by replacing all the functions ψ_a with their Gaussian counterparts, except for one. This is the same procedure used within the EP scheme discussed in the previous Chapter. We thus define another set of probability measures called *tilted* distributions, one for each factor node a, denoted with $q^{(a)}(\boldsymbol{x})$ and defined as follows:

$$q^{(a)}(\boldsymbol{x}) = \frac{1}{Z_a} \prod_{\substack{b\in F \\ b\neq a}} \phi_b(\boldsymbol{x}_b) \times \Psi_a(\boldsymbol{x}_a) \quad \forall a \in F \tag{3.12}$$

where

$$\Psi_a(\boldsymbol{x}_a) = \psi_a(\boldsymbol{x}_a) \prod_{i\in\partial a} \Delta_i(x_i), \tag{3.13}$$

and ψ_a is the *true* factor associated to node a. In practice, the distributions (3.12) are constructed by removing the corresponding Gaussian density ϕ_a and replacing it with the true factor ψ_a, with the addition of the set of constraints $\{\Delta_i\}_{i\in\partial a}$, defined in (3.4): in this way, the discrete nature of the variables in the neighborood of a is correctly taken into account. The main idea behind Density Consistency is that the distribution (3.12) will be a tractable estimator of the marginal distribution of (3.1) on the variables $i \in \partial a$, as the factor ψ_a is correctly included in (3.12): conversely, all the other degrees of freedom are encoded into a normal distribution, and therefore they can be marginalized out analytically. Figure 3.4 shows the factor graph representation of a tilted distribution in the toy model of Figure 3.1 over factor node D: in order to keep in mind the

discrete support of $\{i \in \partial a\}$, each of them is represented as a black dot, analogously to Figure 3.1. Equivalently, a discrete degree of freedom is can be graphically represented as a continuous variable x_i with the additional 1-body function Δ_i, as shown in the right part of Figure 3.4.
It is instructive to rewrite both the Gaussian distribution (3.7) and the tilted (3.12) by isolating

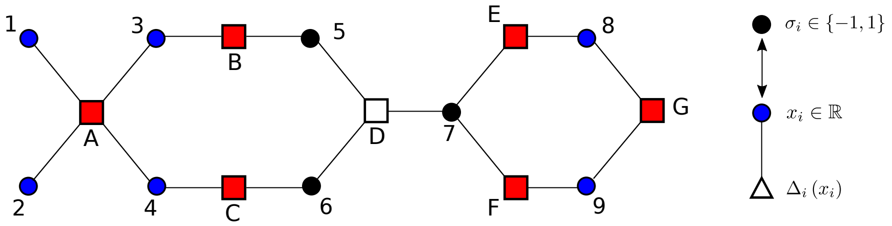

Figure 3.4: Factor graph representation of a tilted distribution over factor node D defined by Eq. (3.12), referred to the toy model in Figure 3.1. In this case, the neighbour nodes of D (variables 5,6,7) are represented as discrete variables (black dots), as a result of the constraints defined by (3.4).

the terms associated to node a:

$$q^{(a)}(\boldsymbol{x}) \propto q(\boldsymbol{x}) \frac{\Psi_a(\boldsymbol{x}_a)}{\phi_a(\boldsymbol{x}_a)} \propto g^{\backslash a}(\boldsymbol{x}) \Psi_a(\boldsymbol{x}_a), \tag{3.14}$$

$$q(\boldsymbol{x}) \propto q(\boldsymbol{x}) \frac{\phi_a(\boldsymbol{x}_a)}{\phi_a(\boldsymbol{x}_a)} \propto g^{\backslash a}(\boldsymbol{x}) \phi_a(\boldsymbol{x}_a), \tag{3.15}$$

where

$$g^{\backslash a}(\boldsymbol{x}) \propto \frac{q(\boldsymbol{x})}{\phi_a(\boldsymbol{x}_a)} \propto \prod_{b \neq a} \phi_b(\boldsymbol{x}_b). \tag{3.16}$$

Indeed, a part of a normalization constant, the tilted distribution can be rewritten as the product of the factor $\Psi_a(\boldsymbol{x}_a)$ and a Gaussian density $g^{\backslash a}(\boldsymbol{x})$, that will be denoted in the rest of the thesis as *Gaussian cavity* distribution. For each factor node a, its corresponding cavity $g^{\backslash a}(\boldsymbol{x})$ can be constructed from (3.7) by simply removing the Gaussian factor ϕ_a, as shown by Eq. (3.16). From the definition of the tilted distribution we can now compute its marginal over the variables $\{i, i \notin \partial a\}$, by means of a Gaussian integral.

$$q^{(a)}(\boldsymbol{x}_a) = \int d\boldsymbol{x}_{\backslash a} q^{(a)}(\boldsymbol{x}) = \Psi_a(\boldsymbol{x}_a) \int d\boldsymbol{x}_{\backslash a} g^{\backslash a}(\boldsymbol{x}) = \frac{1}{z_a} g^{\backslash a}(\boldsymbol{x}_a) \Psi_a(\boldsymbol{x}_a) \tag{3.17}$$

From now on, we use the notation $\rho(\boldsymbol{x}_a)$ to represent the marginal of ρ over the subset of variables $\boldsymbol{x}_a$, with ρ being an arbitrary probability distribution. In Eq. (3.17) the notation $\int d\boldsymbol{x}_{\backslash a}$ indicates the integral over all the variables not connected to factor node a, namely $\boldsymbol{x}_{\backslash a} = \{x_k, k \notin \partial a\}$. The quantity $g^{\backslash a}(\boldsymbol{x}_a)$ in (3.17) is the marginal cavity distribution, that we write explicitly as the exponential of a quadratic form:

$$g^{\backslash a}(\boldsymbol{x}_a) \propto \int d\boldsymbol{x}_{\backslash a} g^{\backslash a}(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}_a^t \boldsymbol{S}^{(a)} \boldsymbol{x}_a + \boldsymbol{x}_a^t \boldsymbol{w}^{(a)}\right] \tag{3.18}$$

where $\boldsymbol{w}^{(a)} \in \mathbb{R}^{|\partial a|}$ is a vector of cavity fields and $\boldsymbol{S}^{(a)} \in \mathbb{R}^{|\partial a| \times |\partial a|}$is a (symmetric by construction) square matrix. In principle, these quantities should be computed starting from the definition (3.16) and then marginalizing over the neighbours of a. However, a simple procedure allows to compute them from the marginal distribution of the full Gaussian measure. Notice first the following equivalence:

$$g^{\backslash a}\left(\boldsymbol{x}_a\right) \propto \int d\boldsymbol{x}_{\backslash a} g^{\backslash a}\left(\boldsymbol{x}\right) \propto \frac{1}{\phi_a\left(\boldsymbol{x}_a\right)} \int d\boldsymbol{x}_{\backslash a} q\left(\boldsymbol{x}\right) \propto \frac{q\left(\boldsymbol{x}_a\right)}{\phi_a\left(\boldsymbol{x}_a\right)} \tag{3.19}$$

Using the above formula, and expressing $q\left(\boldsymbol{x}_a\right)$ as an exponential form, the cavity parameters can be computed much easier through the following relations:

$$\boldsymbol{S}^{(a)} = \left(\boldsymbol{\Sigma}_{[\partial a,\partial a]}\right)^{-1} - \boldsymbol{\Gamma}^{(a)} \tag{3.20}$$

$$\boldsymbol{w}^{(a)} = \left(\boldsymbol{\Sigma}_{[\partial a,\partial a]}\right)^{-1} \cdot \boldsymbol{\mu}_{[\partial a]} - \boldsymbol{\gamma}^{(a)} \tag{3.21}$$

where $\boldsymbol{\mu}_{[\partial a]}$ (resp. $\boldsymbol{\Sigma}_{[\partial a,\partial a]}$) denotes the sub-block of $\boldsymbol{\mu}$ (resp. $\boldsymbol{\Sigma}$) on the ∂a indices. By construction, the contribution to the cavity couplings (i.e. the off-diagonal elements of $\boldsymbol{S}^{(a)}$) comes from all the walks that connect two nodes in the cavity graph, due to the matrix inversion. For instance, in the toy model of Figure (3.1), removing node D splits the factor graph in two disconnected components: however, variables 5 and 6 are still connected by a walk, so that their correlation will be approximately taken into account by the Gaussian cavity distribution, in particular by the off-diagonal term of the matrix $\boldsymbol{S}^{(D)}$ associated to the edge (5,6).

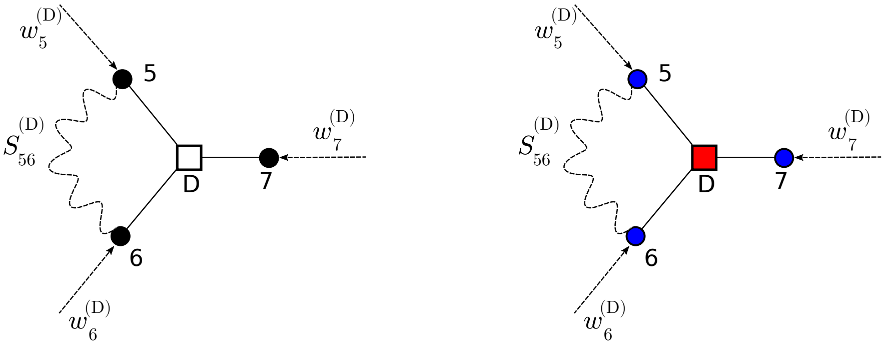

Figure 3.5: Representation of the marginal tilted distribution over factor D of the Toy model of Figure 3.1, highliting the effect of the cavity distribution. For simplicity, the diagonal terms of the cavity coupling matrix $\boldsymbol{S}^{(D)}$are not shown.

For convenience, we report also the expression of the marginal Gaussian measure $q\left(\boldsymbol{x}_a\right)$:

$$\begin{aligned} q\left(\boldsymbol{x}_a\right) &\propto \int d\boldsymbol{x}_{\backslash a} q\left(\boldsymbol{x}\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{x}_a - \boldsymbol{\mu}_{[\partial a]}\right)^t \left(\boldsymbol{\Sigma}_{[\partial a,\partial a]}\right)^{-1}\left(\boldsymbol{x}_a - \boldsymbol{\mu}_{[\partial a]}\right)\right] \\ &\propto g^{\backslash a}\left(\boldsymbol{x}_a\right)\phi_a\left(\boldsymbol{x}_a\right) \end{aligned} \tag{3.22}$$

where the last line follows from the definition of the cavity. The marginal tilted and gaussian distribution are graphically shown in Figure 3.5 for the toy model, when computed over factor node D: in particular, the waved line represents the effective cavity coupling coming from the walk between spin 5 and 6, as previously discussed. On a generic loopy factor graph, each pair of nodes (i,j) in the cavity graph obtained by removing factor node a might still be connected by some walks: if this happens, their correlation will be taken into account approximately by the corresponding element $S_{ij}^{(a)}$ of the cavity coupling matrix.

The moments of the marginal tilted distribution (3.17) can now be easily computed by performing a finite summation over $\{-1{,}1\}^{|\partial a|}$:

$$\langle \boldsymbol{x}_a \rangle_{q^{(a)}} = \frac{1}{z_a} \int d\boldsymbol{x}_a \boldsymbol{x}_a \Psi_a \left(\boldsymbol{x}_a\right) g^{\setminus a} \left(\boldsymbol{x}_a\right) = \frac{1}{z_a} \sum_{\sigma_i, i \in \partial a} \boldsymbol{\sigma}_a \psi_a \left(\boldsymbol{\sigma}_a\right) g^{\setminus a} \left(\boldsymbol{\sigma}_a\right) \tag{3.23}$$

$$\langle \boldsymbol{x}_a \boldsymbol{x}_a^t \rangle_{q^{(a)}} = \frac{1}{z_a} \int d\boldsymbol{x}_a \boldsymbol{x}_a \boldsymbol{x}_a^t \Psi_a \left(\boldsymbol{x}_a\right) g^{\setminus a} \left(\boldsymbol{x}_a\right) = \frac{1}{z_a} \sum_{\sigma_i, i \in \partial a} \boldsymbol{\sigma}_a \boldsymbol{\sigma}_a^t \psi_a \left(\boldsymbol{\sigma}_a\right) g^{\setminus a} \left(\boldsymbol{\sigma}_a\right) \tag{3.24}$$

where $z_a = \sum_{\sigma_i, i \in \partial a} \psi_a \left(\boldsymbol{\sigma}_a\right) g^{\setminus a} \left(\boldsymbol{\sigma}_a\right)$ is the partition function of the marginal tilted distribution, and the variables $\sigma_i \in \{-1{,}1\}$ are re-introduced to stress that the marginal tilted is in practice a discrete probability. Notice that moments of the tilted distribution are affected by the presence of non-diagonal elements in the cavity coupling matrix $\boldsymbol{S}^{(a)}$; on the contrary, its diagonal entries are not effective since $\sigma_i^2 = \text{const}$ for $\sigma_i = \pm 1$, so that their contribution can be included in the normalization constant z_a. It is important to remark that in order to evaluate (3.23)-(3.24), it is not necessary to know all the elements of $\boldsymbol{\Sigma}$: only the elements of the sub-block $[\partial a, \partial a]$ enter into the computation of the marginal tilted moments. In principle, one could think about iterative schemes to compute only the elements of $\boldsymbol{\Sigma}$ needed, instead of relying on standard matrix inversion techniques. Neverthless, the other entries of $\boldsymbol{\Sigma}$ (i.e. Σ_{ij} for $i \in \partial a, j \in \partial b$, $b \neq a$) can be still used to estimate long-range pairwise correlations.

3.1.2 DC condition

As stated at the beginning, the goal is to define a family of approximation schemes such that computation of marginals is exact in the case of acyclic graphs. To do so, for each node a, we impose a matching of the *density* values between the single-node marginals of the tilted distribution $q^{(a)}$, defined by Eq. (3.12)-(3.17) and the full Gaussian q (3.7) for each node $i \in \partial a$ on the support $\{-1{,}1\}$. This condition can be rephrased as $q^{(a)}\left(x_i\right) \propto q\left(x_i\right)$, where the dependency on x_i on the two distributions implies that we are considering their marginal distributions. We first rewrite the single-node marginals of $q^{(a)}$ and q for a variable $i \in \partial a$:

$$q^{(a)}\left(x_i\right) = \int d\boldsymbol{x}_{a \setminus i} q^{(a)}\left(\boldsymbol{x}_a\right) = \frac{1 + x_i \left\langle x_i \right\rangle_{q^{(a)}}}{2} \Delta_i\left(x_i\right) \tag{3.25}$$

where in the last equality the single-node marginal is written in terms of its first moment $\left\langle x_i \right\rangle_{q^{(a)}}$ (namely, the magnetization), without loss of generality. In the same spirit, the marginal of the full Gaussian distribution $q\left(\boldsymbol{x}\right)$ over node i can be simply written as:

$$q\left(x_i\right) = \int d\boldsymbol{x}_{\setminus i} q\left(\boldsymbol{x}\right) = \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \exp\left[-\frac{\left(x_i - \mu_i\right)^2}{2\Sigma_{ii}}\right] \tag{3.26}$$

Therefore, the matching of the density values of (3.25) and (3.26) on $x_i = \{-1{,}1\}$ can be rephrased as:

$$q^{(a)}\left(x_i\right) \propto q\left(x_i\right) \iff \frac{q\left(x_i = +1\right)}{q\left(x_i = -1\right)} = \frac{q^{(a)}\left(x_i = +1\right)}{q^{(a)}\left(x_i = -1\right)} \tag{3.27}$$

After some straightforward algebra, the following condition is obtained:

$$\frac{\mu_i}{\Sigma_{ii}} = \operatorname{atanh} \left\langle x_i \right\rangle_{q^{(a)}} \qquad \forall i \in \partial a, a \in F. \tag{3.28}$$

Eq. (3.28) is called *DC condition* and it is chosen because it ensures exactness on acyclic graphs: a rigorous proof will be discussed in the next section. Qualitatively, this condition imposes that the single-node marginal $q\left(x_i\right)$ has the same behaviour of $q^{(a)}\left(x_i\right)$ when evaluated on the discrete support $\{-1{,}1\}$: this is possible because a univariate gaussian can always be fitted on two values, provided that their sufficient statistics μ_i and Σ_{ii} satisfy (3.28). An equivalent way of deriving (3.28) is by imposing a *moment* matching condition between the single-node marginals of the tilted distribution $q^{(a)}$ and a modified Gaussian distribution $\hat{q}^{(i)}\left(\boldsymbol{x}\right)$, obtained from (3.7) by adding a discrete constraint Δ_i on variable i:

$$\hat{q}^{(i)}\left(\boldsymbol{x}\right)=\frac{1}{\hat{Z}_i}q\left(\boldsymbol{x}\right)\Delta_i\left(x_i\right) \tag{3.29}$$

Notice that both $\hat{q}^{(i)}$ and $q^{(a)}$are *discrete* distributions over node i. It is straightforward to show that $\langle x_i\rangle_{\hat{q}^{(i)}}=\tanh\frac{\mu_i}{\Sigma_{ii}}$, so that the DC condition just described can also be written as:

$$\left\langle x_i\right\rangle_{q^{(a)}}=\left\langle x_i\right\rangle_{\hat{q}^{(i)}} \qquad \forall i\in\partial a, a\in F$$

This alternative derivation will be used in Section 3.4.2 where preliminar calculations to derive a variational DC free energy will be carried out.
The total number of equations obtained so far by imposing (3.28) is $\sum_a\left|\partial a\right|$. From now on, we call *Density Consistency* (DC) any scheme that enforces Eq. (3.28). As a final remark, notice that Eq. 3.28 guarantees that single node marginals for a certain spin i are independent on the tilted distribution used to compute them (at least at fixed point):

$$\left\langle x_i\right\rangle_{q^{(a)}}=\tanh\frac{\mu_i}{\Sigma_{ii}} \qquad \forall i\in\partial a \tag{3.30}$$

where the right-hand side does not depend on a.

3.1.3 DC closure

DC condition (3.28) is not enough to fix all the parameters encoded in the set of Gaussian factors $\left\{\phi_a\right\}_{a\in F}$. In principle, there are infinite choices to fix the remaining $\sum_a\left|\partial a\right|\left(\left|\partial a\right|+1\right)/2$ parameters. We propose to complement (3.28) with two further conditions, i.e. the matching of first moments and Pearson correlation coefficients, between the (marginal) tilted distributions and the (marginal) Gaussian measure, for each $a\in F$:

$$\mu_i=\left\langle x_i\right\rangle_{g^{(a)}} \tag{3.31}$$

$$\mathrm{corr}_q\left(x_i,x_j\right)=\mathrm{corr}_{q^{(a)}}\left(x_i,x_j\right) \tag{3.32}$$

where $\mu_i=\left\langle x_i\right\rangle_q$ and

$$\mathrm{corr}_\rho\left(x_i,x_j\right)\hat{=}\frac{\langle x_ix_j\rangle_\rho-\langle x_i\rangle_\rho\langle x_j\rangle_\rho}{\sqrt{\left(1-\langle x_i\rangle_\rho^2\right)\left(1-\langle x_j\rangle_\rho^2\right)}} \tag{3.33}$$

is the Pearson correlation coefficient between to variables i and j of a distribution ρ (in this case $\rho\in\left\{q,q^{(a)}\right\}$). Finally, by putting together the DC condition (3.28) and the above matching equations (3.31)-(3.32) we get the following system of *closure* equations, for each factor node

$a \in F$:

$$\mu_i = \langle x_i \rangle_{q^{(a)}} \qquad \forall i \in \partial a \tag{3.34a}$$

$$\Sigma_{ii} = \frac{\langle x_i \rangle_{q^{(a)}}}{\text{atanh}\langle x_i \rangle_{q^{(a)}}} \qquad \forall i \in \partial a \tag{3.34b}$$

$$\Sigma_{ij} = \eta \left(\langle x_i x_j \rangle_{q^{(a)}} - \langle x_i \rangle_{q^{(a)}} \langle x_j \rangle_{q^{(a)}} \right) \sqrt{\frac{\Sigma_{ii}\Sigma_{jj}}{\left(1 - \langle x_i \rangle^2_{q^{(a)}}\right)\left(1 - \langle x_j \rangle^2_{q^{(a)}}\right)}} \qquad \forall i, j \in \partial a, i \neq j \tag{3.34c}$$

where the Gaussian variances Σ_{ii} are derived from (3.28) by using (3.34a). Other possible closures are discussed in Section 3.2.2. Note that despite (3.34b) is not well-defined for $m_i = 0$, it has a finite limit: in particular, $\lim_{\langle x_i \rangle_{q^{(a)}} \to 0} \Sigma_{ii} = 1$ (see also 3.6). Notice also that in (3.34c) a further parameter η has been added. The quantity η plays the role of an interpolation between a full DC solution, obtained by matching the Pearson coefficient, and the BP fixed points: indeed, as it will be discussed in the next section, setting $\eta = 0$ is equivalent to neglect cavity correlations and it turns out to give BP fixed points on *any* graph topology. However, for the time being, let us put $\eta = 1$ for simplicity. The set (3.34) is a system of an equal number of equations and unknowns, that can be iteratively solved w.r.t. to the Gaussian parameters $\left\{\boldsymbol{\gamma}^{(a)}, \boldsymbol{\Gamma}^{(a)}\right\}_{a \in F}$ to provide an estimation of the original distribution's moments onto each factor node.

The choice behind the matching of the Pearson correlation coefficient is justified by the following argument. Suppose to have already applied a set of closure conditions to fix both the first moments and the variances of the marginal Gaussian distribution $q\left(\boldsymbol{x}_{\partial a}\right)$ (respectively, $\left\{\mu_i, \Sigma_{ii}\right\}_{i \in \partial a}$). The remaining parameters to fix are the off-diagonal covariances, whose number is $\binom{|\partial a|}{2}$. Matching the Pearson coefficient is equivalent to apply the same trasformation used to map the tilted variances to the Gaussian's ones, also on the nearest neighbours' covariances. For simplicity, we now restrict to factor nodes of degree 2, even if the following argument holds for arbitrary connectivity. Let us start from the covariance matrix of the marginal tilted distribution, defined as:

$$\langle \boldsymbol{x}_{\partial a} \boldsymbol{x}^T_{\partial a} \rangle_{q^{(a)}} - \langle \boldsymbol{x}_{\partial a} \rangle_{q^{(a)}} \langle \boldsymbol{x}^T_{\partial a} \rangle_{q^{(a)}} = \begin{pmatrix} 1 - m_i^2 & c_{ij} \\ c_{ij} & 1 - m_j^2 \end{pmatrix} \tag{3.35}$$

where $m_i = \langle x_i \rangle_{q^{(a)}}$ and c_{ij} is the connected covariance between i and j. First notice that, thanks to the Cauchy-Schwartz inequality, $c_{ij}^2 \leq \left(1 - m_i^2\right)\left(1 - m_j^2\right)$, so that the determinant of (3.35) is always non-negative. DC works by imposing local consistency conditions between $q^{(a)}\left(\boldsymbol{x}_{\partial a}\right)$ and $q\left(\boldsymbol{x}_{\partial a}\right)$, so that the two covariance matrix will be connected by a certain transformation

$$\begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ij} & \Sigma_{jj} \end{pmatrix} = \underline{F}\left[\begin{pmatrix} 1 - m_i^2 & c_{ij} \\ c_{ij} & 1 - m_j^2 \end{pmatrix}\right] \tag{3.36}$$

where $\underline{F} : \mathbb{R}^{|\partial a|(|\partial a|+3)/2} \to \mathbb{R}^{|\partial a|(|\partial a|+3)/2}$. The diagonal elements of the covariance matrix in (3.22) are already fixed thanks to (3.34a)-(3.34b), so that $\Sigma_{ii} = F_{ii}\left(m_i\right) = m_i/\text{atanh}m_i$. The idea is to transform also the non-diagonal covariances in such a way to mimic the same trasformation from the variances of the tilted distributions onto the Gaussian's ones. By defining the following quantity:

$$A_i\left(m_i\right) = \sqrt{\frac{m_i}{\left(1 - m_i^2\right)\text{atanh}m_i}} \tag{3.37}$$

the resulting covariance matrix of the (marginal) Gaussian distribution will be given by:

$$\langle \boldsymbol{x}_{\partial a} \boldsymbol{x}^T_{\partial a} \rangle_q - \langle \boldsymbol{x}_{\partial a} \rangle_q \langle \boldsymbol{x}^T_{\partial a} \rangle_q = \begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ij} & \Sigma_{jj} \end{pmatrix} = \begin{pmatrix} \left(1 - m_i^2\right) A_i^2 & c_{ij} A_i A_j \\ c_{ij} A_i A_j & \left(1 - m_j^2\right) A_j A_j \end{pmatrix} \tag{3.38}$$

With the above mapping, the covariances Σ_{ij} are fixed by $\Sigma_{ij} = F_{ij}\left(c_{ij}, m_i, m_j\right) = c_{ij} A_i A_j$. It is easy to verify that such a condition is equivalent to (3.34c). Indeed, by rewriting A_i in terms of Σ_{ii} (and the same for A_j), we get:

$$\Sigma_{ij} = c_{ij} A_i A_j = c_{ij} \sqrt{\frac{\Sigma_{ii}}{\left(1 - m_i^2\right)}} \sqrt{\frac{\Sigma_{jj}}{\left(1 - m_j^2\right)}} \tag{3.39}$$

Finally, by dividing both sides by $\sqrt{\Sigma_{ii}\Sigma_{jj}}$ we recover the Pearson correlation matching as in (3.34c). This matching for the off-diagonal covariances seems to be a natural choice to ensure that, if the covariance matrix of the tilted distribution (3.35) is invertible, so it is the Gaussian's one (3.36). Indeed, the determinant of (3.38) is equal to $A_i^2 A_j^2 \left[\left(1 - m_i^2\right)\left(1 - m_j^2\right) - c_{ij}^2\right]$, that is always non-negative by construction: in particular, the product $A_i^2 A_j^2$ is positive by definition of (3.37) (i.e. $m_i/\text{atanh} m_i \geq 0$), and the term in square brakets is non-negative thanks to the Cauchy-Schwartz inequality. Finally, by using this mapping it is necessary to fix an arbitrary set of closures on the first moments $\mu_i = \alpha^{(1)}\left(m_i\right)$ and on the diagonal entries $\Sigma_{ii} = \alpha^{(2)}\left(m_i\right)$ in such a way to satisfy DC condition (3.28), i.e. $\alpha^{(1)}\left(m_i\right)/\alpha^{(2)}\left(m_i\right) = \text{atanh} m_i$; then, by defining $A_i\left(m\right) = \sqrt{\alpha^{(2)}\left(m\right)/\left(1 - m_i^2\right)}$, the non-diagonal elements can be computed by using (3.39).

3.1.4 Parameters' update

The above set of closure equations provides a way to fix the Gaussian moments $\boldsymbol{\mu}_{[\partial a]}$, $\boldsymbol{\Sigma}_{[\partial a, \partial a]}$ in terms of the tilted ones. Their effect is then incorporated into the full Gaussian density q by updating the parameters encoded into ϕ_a, i.e. the approximating factor over node a. The update rule of its parameters follows from (3.20)-(3.21):

$$\left(\boldsymbol{\Gamma}^{(a)}\right)^{\tau+1} = \left[\boldsymbol{\Sigma}^{\tau}_{[\partial a, \partial a]}\right]^{-1} - \left(\boldsymbol{S}^{(a)}\right)^{\tau} \tag{3.40}$$

$$\left(\boldsymbol{\gamma}^{(a)}\right)^{\tau+1} = \left[\boldsymbol{\Sigma}^{\tau}_{[\partial a, \partial a]}\right]^{-1} \cdot \boldsymbol{\mu}^{\tau}_{[\partial a]} - \left(\boldsymbol{w}^{(a)}\right)^{\tau} \tag{3.41}$$

where τ is an integer corresponding to the iteration number and $\boldsymbol{w}^{(a)}$, $\boldsymbol{S}^{(a)}$ are the cavity parameters.

3.2 Properties

In this section, we summarize the main properties of Density Consistency, namely its exactness on trees and the relation to Belief Propagation on generic loopy graphs, as well as the relation to Expectation Propagation.

3.2.1 Exactness on trees and relation to Belief Propagation

On acyclic graphs, DC condition (3.28) is sufficient to guarantee exact computation of marginals, independently on the other closure equations used to fix the remaining paramters of $\{\phi_a\}_{a\in F}$. Moreover, by neglecting cavity covariances DC marginals coincide with the Belief Propagation fixed points on any graph topology. These two properties can be rigorously proven in the following theorem. For convenience, we first recall the Belief Propagation fixed point equations, already discussed in Sec (2.2):

$$m_{a\to i}(\sigma_i) \propto \sum_{\boldsymbol{\sigma}_{\partial a\setminus i}} \psi_a(\boldsymbol{\sigma}_a) \prod_{j\in\partial a\setminus i} \nu_{j\to a}(\sigma_j) \tag{3.42}$$

$$\nu_{i\to a}(\sigma_i) \propto \prod_{b\in\partial i\setminus a} m_{b\to i}(\sigma_i) \tag{3.43}$$

$$b_i(\sigma_i) \propto \prod_{b\in\partial i} m_{b\to i}(\sigma_i) \tag{3.44}$$

where $\sigma_i \in \{-1,1\}$. The first two are recognized as the BP message-passing equations (resp. (2.32)-(3.43) in Section 2.2), and the third one is the single-node belief (2.31).

Theorem 1. *If (H1) the factor graph is acyclic or (H2) DC scheme applies null covariances ($\eta = 0$), the quantity $g^{\setminus a}(x_i) \propto \nu_{i\to a}(x_i)$ satisfies (3.43), and the single node beliefs $q^{(a)}(x_i) \propto b_i(x_i)$ satisfy (3.44)(3.42) when $x_i \in \{-1,1\}$.*

Proof. Under either hypothesis (H1 or H2), the marginal cavity distribution $g^{\setminus a}(\boldsymbol{x}_a)$ (3.18) is factorized: respectively, under H1 this is true by construction since the graph is a tree, while under H2 it holds because in the full Gaussian measure $q(\boldsymbol{x})$ we neglect all the connected correlations Σ_{ij} (i.e. the matrix $\boldsymbol{\Sigma}$ is diagonal). In both cases, we rewrite the marginal cavity distribution as the product over single-node distributions:

$$g^{\setminus a}(\boldsymbol{x}_a) \propto \prod_{i\in\partial a} g^{\setminus a}(x_i) \propto \prod_{i\in\partial a} \nu_{i\to a}(x_i) \tag{3.45}$$

Let us define the quantity $m_{a\to i}(x_i) \propto \frac{q(x_i)}{\nu_{i\to a}(x_i)}$. Thanks to DC condition (3.28), $q(x_i) \propto q^{(a)}(x_i)$ when $x_i \in \{-1,1\}$. Therefore:

$$\begin{aligned}
m_{a\to i}(x_i) &\propto \frac{q(x_i)}{\nu_{i\to a}(x_i)} && (3.46)\\
&\propto \frac{q^{(a)}(x_i)}{\nu_{i\to a}(x_i)} && (3.47)\\
&\propto \frac{1}{\nu_{i\to a}(x_i)} \int d\boldsymbol{x}_{\partial a\setminus i} q^{(a)}(\boldsymbol{x}_a) && (3.48)\\
&\propto \frac{1}{\nu_{i\to a}(x_i)} \int d\boldsymbol{x}_{\partial a\setminus i} g^{\setminus a}(\boldsymbol{x}_a)\,\Psi_a(\boldsymbol{x}_a) && (3.49)\\
&\propto \int d\boldsymbol{x}_{\partial a\setminus i} \prod_{j\in\partial a\setminus i} \nu_{j\to a}(x_j)\,\Psi_a(\boldsymbol{x}_a) && (3.50)
\end{aligned}$$

where in (3.47) we used DC condition, in (3.48)(3.49) the definition of the marginal tilted distribution in terms of cavities (3.17) and (3.45) to derive the last line. Eq. (3.50) is identical to (3.42): indeed, by virtue of the constraints included in Ψ_a, both right sides of (3.50) have measure only on $x_i \in \{-1,1\}$. We conclude that the set of messages $\{m_{a\to i}, \nu_{i\to a}\}$ satisfy (3.42) under either hypothesis H1 and H2. Notice also the following relation:

$$\begin{aligned}
m_{a\to i}\left(x_i\right) \propto & \frac{q\left(x_i\right)}{\nu_{i\to a}\left(x_i\right)} \\
& \frac{1}{\nu_{i\to a}\left(x_i\right)}\int d\boldsymbol{x}_{\partial a\setminus i} q\left(\boldsymbol{x}_a\right) \qquad (3.51)\\
& \propto \frac{1}{\nu_{i\to a}\left(x_i\right)}\int d\boldsymbol{x}_{\partial a\setminus i} g^{\setminus a}\left(\boldsymbol{x}_a\right)\phi_a\left(\boldsymbol{x}_a\right) \\
& \propto \int d\boldsymbol{x}_{\partial a\setminus i}\prod_{j\in\partial a\setminus i}\nu_{j\to a}\left(x_j\right)\phi_a\left(\boldsymbol{x}_a\right) \qquad (3.52)
\end{aligned}$$

Under hypothesis (H2), the full Gaussian measure is factorized by assumption, $q\left(\boldsymbol{x}\right)\propto\prod_{i\in V}q\left(x_i\right)$. By construction, also the approximate Gaussian factors will be factorized:

$$\begin{aligned}
\phi_a\left(\boldsymbol{x}_a\right) & \propto \frac{q\left(\boldsymbol{x}_a\right)}{g^{\setminus a}\left(\boldsymbol{x}_a\right)} \\
& \propto \prod_{i\in\partial a}\frac{q\left(x_i\right)}{\nu_{i\to a}\left(x_i\right)} \\
& \propto \prod_{i\in\partial a} m_{a\to i}\left(x_i\right) \qquad (3.53)
\end{aligned}$$

an therefore $\phi_a\left(\boldsymbol{x}_a\right)\propto\prod_{i\in\partial a}\phi_a\left(x_i\right)$ with $\phi_a\left(x_i\right)\propto m_{a\to i}\left(x_i\right)$. Therefore we get:

$$\begin{aligned}
\nu_{i\to a}\left(x_i\right) & \propto g^{\setminus a}\left(x_i\right) \\
& \propto \int d\boldsymbol{x}_{\setminus i}\prod_{c\neq a}\phi_c\left(\boldsymbol{x}_c\right) \\
& \propto \int d\boldsymbol{x}_{\setminus i}\prod_{c\neq a}\prod_{j\in\partial c} m_{c\to j}\left(x_j\right) \\
& \propto \prod_{c\in\partial i\setminus a} m_{c\to i}\left(x_i\right)\int d\boldsymbol{x}_{\setminus i}\prod_{c\neq a}\prod_{j\in\partial c\setminus i} m_{c\to j}\left(x_j\right) \\
& \propto \prod_{c\in\partial i\setminus a} m_{c\to i}\left(x_i\right)
\end{aligned}$$

where the integral is just a constant that can be included into the normalization factor. To derive (3.43) under (H1), we define the quantity T_b as the set of factors in the connected component of

b once i is removed. We get:

$$\begin{aligned}
\nu_{i\to a}\left(x_i\right) &\propto g^{\setminus a}\left(x_i\right)\\
&\propto \int d\boldsymbol{x}_{\setminus i}\prod_{c\neq a}\phi_c\left(\boldsymbol{x}_c\right)\\
&\propto \int d\boldsymbol{x}_{\setminus i}\prod_{b\in\partial i\setminus a}\phi_b\left(\boldsymbol{x}_b\right)\prod_{c\in T_b\setminus b}\phi_c\left(\boldsymbol{x}_c\right)\\
&\propto \prod_{b\in\partial i\setminus a}\left[\int d\boldsymbol{x}_{b\setminus i}\phi_b\left(\boldsymbol{x}_b\right)\prod_{j\in\partial b\setminus i}g^{\setminus b}\left(x_j\right)\right]\\
&\propto \prod_{b\in\partial i\setminus a}\left[\int d\boldsymbol{x}_{b\setminus i}\phi_b\left(\boldsymbol{x}_b\right)\prod_{j\in\partial b\setminus i}m_{b\to j}\left(x_j\right)\right]\\
&\propto \prod_{b\in\partial i\setminus a}m_{b\to i}\left(x_i\right)
\end{aligned}$$

where in the last line we used (3.52). Under either hypothesis H1, H2, the quantity $\nu_{i\to a}\left(x_i\right)$ satisfies (3.43). Finally, by construction $q\left(x_i\right)\propto v_{i\to a}\left(x_i\right)m_{a\to i}\left(x_i\right)$, so that also the equation for the single node belief (3.44) is satisfied. □

Therefore, we proved that Density Consistency is exact on acyclic graphs and gives the same BP fixed points on any graph topology, provided that cavity correlations are neglected. Moreover, a simple inspection of the update equations in (1) shows that DC update rule equivalent to an ordinary BP update. In this sense, Density Consistency can be considered as a generalization of Belief Propagation in which the factorization assumption in the cavity distribution is related, in such a way to include some effective interactions between the nodes in the cavity. These interactions are encoded by a Gaussian distribution, that easily allows to perform analytic marginalization over any set of variables. If one neglects cavity covariances, the full Gaussian measures is factorized over nodes, so that there is no contribution coming from cavity couplings to the moments of the tilted distribution defined by (3.23)-(3.24). The result of Theorem 1 under Hypothesis H2 is equivalent to what derived by Minka and discussed at the end of Section 2.4: the only difference is that in the latter case the approximating family is defined as a joint factorized distribution over *discrete* degrees of freedom. Here instead we deal with continuous distributions: however, DC condition takes care of that, by imposing a consistency on the density values so that the marginals $q\left(x_i\right)$ are fitted on the discrete support of the binary spin, so to be equivalent to the moment matching condition (2.84). In this sense, Density Consistency can be considered as a generalization of the method presented in Section 2.4.3 where the approximating family is not factorized, but rather encoded by a family of multivariate (continuous) Gaussian densities.

The interpolation parameter η

Theorem 1 proves that setting $\eta = 0$ in (3.34c) is equivalent to assume that the cavity distribution is factorized, and therefore DC fixed points coincide with BP fixed points. The role of the interpolation parameter η is well defined for the two limit values of $\eta = 0{,}1$: the first one corresponds to BP fixed points, the second to DC fixed points obtained by "fully" matching the Pearson correlation coefficient. However, in principle one could also set an intermediate value of the interpolation parameter in the interval $\eta \in (0{,}1)$: in this case, cavity correlations are damped w.r.t. to the full DC solution. In this scenario, the meaning of DC approximation in this regime has not a clear interpretation, but using a value of $\eta < 1$ can help convergence in some regimes where a full DC solution (obtained with $\eta = 1$) cannot be found. In this sense, η can be considered as a hyperparameter of this approximation and its effect will be discussed more in details in Chapter 4.

3.2.2 Closure equations and relation to EP

Theorem 1 states that Density Consistency allows to compute exact marginal distributions on acyclic graphs, provided that DC condition (3.28) is satisfied. In principle, the remaining set of closure conditions could be chosen arbitrarily: for instance, one could either match the full set of second moments, namely $\langle \boldsymbol{x}_a \boldsymbol{x}_a^t \rangle_{q^{(a)}} = \langle \boldsymbol{x}_a \boldsymbol{x}_a^t \rangle_q$ and use DC conditions (3.28) to fix the first moments μ_i of the full Gaussian distribution $q(\boldsymbol{x})$, namely:

$$\mu_i = \left(1 - \langle x_i \rangle^2_{q^{(a)}}\right) \operatorname{atanh}\langle x_i \rangle_{q^{(a)}} \tag{3.54a}$$

$$\Sigma_{ii} = 1 - \langle x_i \rangle^2_{q^{(a)}} \tag{3.54b}$$

$$\Sigma_{ij} = \langle x_i x_j \rangle_{q^{(a)}} - \langle x_i \rangle_{q^{(a)}} \langle x_j \rangle_{q^{(a)}} \tag{3.54c}$$

In principle, any scheme satisfying (3.28) is exact on trees independently on the other closure equations used to fix the remaining parameters. In the following, we present another closure update scheme directly inspired by standard implementations of Expectation Propagation, obtained by imposing the matching of the first two moments between the full Gaussian distribution an the tilted distributions:

$$\mu_i = \langle x_i \rangle_{q^{(a)}} \tag{3.55a}$$

$$\Sigma_{ii} = 1 - \langle x_i \rangle^2_{q^{(a)}} \tag{3.55b}$$

$$\Sigma_{ij} = \langle x_i x_j \rangle_{q^{(a)}} - \langle x_i \rangle_{q^{(a)}} \langle x_j \rangle_{q^{(a)}} \tag{3.55c}$$

for $\forall a \in F$. In the rest of the manuscript, we will refer to the set of equations (3.55) as “EP” closure. This set of equations is not exact on trees in general, as (3.28) is not satisfied. In this sense, Density Consistency can be considered as an extension to the Gaussian Expectation Propagation algorithm to multivariate factors, but with a different consistency condition designed in such a way that the computation of marginals is exact on trees when the degrees of freedom are binary spins. To highlight the difference between the two sets of closures, we show in Figure 3.6 the behaviour of μ_i/Σ_{ii} as function of the tilted magnetization $\langle x_i \rangle_{q^{(a)}}$. In particular, for the DC closure the quantity μ_i/Σ_{ii} is directly obtained from DC condition (3.28), while in the EP closure it can be derived by using (3.55a)-(3.55b), that leads to $\mu_i^{EP}/\Sigma_{ii}^{EP} = \langle x_i \rangle_{q^{(a)}} / \left(1 - \langle x_i \rangle^2_{q^{(a)}}\right)$. Qualitatively, the behaviour is similar between the two methods, but only DC condition is exact on acyclic graphs, independently on the magnetizations. It is interesting to notice that in the simple case of zero magnetizations also the EP closure becomes exact on trees: however, despite they give the same fixed points, the two closures differ in the dynamical update of the parameters: indeed, we found that EP closure has typically poor convergence performances if compared to DC. Moreover, EP closure is exact also in the limit of extreme polarized variables, i.e. when the magnetization of spin i tends to 1 (or -1), that eventually occurs only in a zero temperature limit. The EP closure will be extensively used in Chapter 5 for the inverse Ising problem.

3.2.3 Weight Gauge

Another interesting property concerns the possibility to move freely Gaussian-like densities in and out the exact factors $\psi_a(x_a)$. We first notice that the derivation presented so far can be

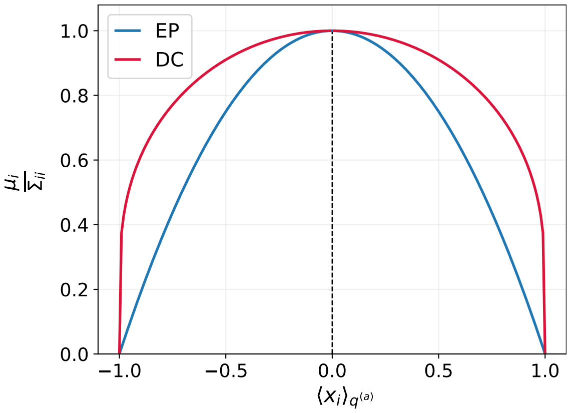

Figure 3.6: Plot of μ_i/Σ_{ii} as function of the tilted magnetization $\langle x_i \rangle_{q^{(a)}} = m_i$. Comparison between DC condition (3.28) and the equivalent "EP" results obtained by using (3.55a) and (3.55b).

carried out in the same way if the starting probability distribution contains a multivariate normal distribution $\Phi(\boldsymbol{x})$:

$$p(\boldsymbol{x}) \propto \Phi(\boldsymbol{x}) \prod_a \psi_a(\boldsymbol{x}_a) \prod_i \Delta_i(x_i) \tag{3.56}$$

With the above parametrization, the full Gaussian approximation and the tilted distributions respectively read:

$$q(\boldsymbol{x}) \propto \Phi(\boldsymbol{x}) \prod_{a \in F} \phi_a(\boldsymbol{x}_a) \tag{3.57}$$

$$q^{(a)}(\boldsymbol{x}) \propto \Phi(\boldsymbol{x}) \prod_{b \neq a} \phi_b(\boldsymbol{x}_a) \times \Psi_a(\boldsymbol{x}_a) \quad a \in F \tag{3.58}$$

Let us now define a set of Gaussian densities $\{\rho_a(\boldsymbol{x}_a)\}_{a \in F}$. We now rewrite Eqs.(3.56)-(3.57)-(3.58) by inserting the set $\{\rho_a(\boldsymbol{x}_a)\}$:

$$p(\boldsymbol{x}) \propto \left[\Phi(\boldsymbol{x}) \prod_a \rho_a(\boldsymbol{x}_a)\right] \prod_a \frac{\psi_a(\boldsymbol{x}_a)}{\rho_a(\boldsymbol{x}_a)} \prod_i \Delta_i(x_i) \tag{3.59}$$

$$\begin{aligned} q(\boldsymbol{x}_a) &\propto \int d\boldsymbol{x}_{\setminus a} \Phi(\boldsymbol{x}) \prod_b \phi_b(\boldsymbol{x}_b) \\ &\propto \int d\boldsymbol{x}_{\setminus a} \Phi(\boldsymbol{x}) \prod_b \phi_b(\boldsymbol{x}_b) \times \frac{\prod_b \rho_b(\boldsymbol{x}_b)}{\prod_b \rho_b(\boldsymbol{x}_b)} \\ &\propto \int d\boldsymbol{x}_{\setminus a} \left[\Phi(\boldsymbol{x}) \prod_b \rho_b(\boldsymbol{x}_b)\right] \times \prod_b \frac{\phi_b(\boldsymbol{x}_b)}{\rho_b(\boldsymbol{x}_b)} \end{aligned} \tag{3.60}$$

$$\begin{aligned} q^{(a)}(\boldsymbol{x}_a) &\propto \Psi_a(\boldsymbol{x}_a) \int d\boldsymbol{x}_{\setminus a} \frac{\Phi(\boldsymbol{x}) \prod_b \phi_b(\boldsymbol{x}_b)}{\phi_a(\boldsymbol{x}_a)} \\ &\propto \frac{\Psi_a(\boldsymbol{x}_a)}{\rho_a(\boldsymbol{x}_a)} \int d\boldsymbol{x}_{\setminus a} \frac{\Phi(\boldsymbol{x}) \prod_b \phi_b(\boldsymbol{x}_b)}{\phi_a(\boldsymbol{x}_a)/\rho_a(\boldsymbol{x}_a)} \times \frac{\prod_b \rho_b(\boldsymbol{x}_b)}{\prod_b \rho_b(\boldsymbol{x}_b)} \\ &\propto \frac{\Psi_a(\boldsymbol{x}_a)}{\rho_a(\boldsymbol{x}_a)} \int d\boldsymbol{x}_{\setminus a} \frac{\left[\Phi(\boldsymbol{x}) \prod_b \rho_b(\boldsymbol{x}_b)\right]}{\phi_a(\boldsymbol{x}_a)/\rho_a(\boldsymbol{x}_a)} \times \prod_b \frac{\phi_b(\boldsymbol{x}_b)}{\rho_b(\boldsymbol{x}_b)} \end{aligned} \tag{3.61}$$

where $\Psi_a(\boldsymbol{x}_a) = \psi_a(\boldsymbol{x}_a)\prod_{i\in\partial a}\Delta_i(x_i)$. Define now the following distributions:

$$\Phi'(\boldsymbol{x}) = \Phi(\boldsymbol{x})\prod_b \rho_b(\boldsymbol{x}_b) \qquad \phi'_a(\boldsymbol{x}_a) = \frac{\phi_a(\boldsymbol{x}_a)}{\rho_a(\boldsymbol{x}_a)} \qquad \psi'_a(\boldsymbol{x}_a) = \frac{\psi_a(\boldsymbol{x}_a)}{\rho_a(\boldsymbol{x}_a)} \tag{3.62}$$

Inserting these expression in (3.61)-(3.60), we get the same parametrization of Eqs. (3.56)-(3.57)-(3.58):

$$q^{(a)}(\boldsymbol{x}_a) \propto \Psi'_a(\boldsymbol{x}_a)\int d\boldsymbol{x}_{\setminus a}\frac{\Phi'(\boldsymbol{x})\prod_b \phi'_b(\boldsymbol{x}_b)}{\phi'_a(\boldsymbol{x}_a)} \tag{3.63}$$

$$q(\boldsymbol{x}_a) \propto \int d\boldsymbol{x}_{\setminus a}\Phi'(\boldsymbol{x})\prod_b \phi'_b(\boldsymbol{x}_b) \tag{3.64}$$

Since DC scheme imposes local constraints between the marginals $q^{(a)}(\boldsymbol{x}_a)$ and $q(\boldsymbol{x}_a)$, a certain approximating family defined by the set of Gaussian factors $\{\phi_a\}_{a\in F}$ for the distribution (3.56) and identified by $\left(\Phi, \{\psi_a\}_{a\in F}\right)$ leads to an equivalent family $\{\phi'_a = \phi_a/\rho_a\}_{a\in F}$ for the distribution (3.59) identified by $\left(\Phi' = \Phi\prod_b \rho_b, \{\psi'_a = \psi_a/\rho_a\}_{a\in F}\right)$ for arbitrary Gaussian densities $\{\rho_a\}_{a\in F}$. Notice that the distributions (3.56) and (3.59) are exactly the same, and therefore the estimate of marginal distributions must coincide. The latter reasoning holds independently on the set of closure equations used. A first consequence is that adding a diagonal matrix into the starting pdf does not modify at all fixed points: however, it may be use to prevent numerical issues arising in the inversion of the covariance matrix . In particular, this property can be useful in the case of Ising-like models, where the Hamiltonian contains only linear and quadratic terms.

3.3 Algorithmic details and implementation

In this section, we discuss some additional details about the implementation of Density Consistency. Given a set of closure equations like (3.34), the Gaussian parameters can be updated iteratively by virtue of Eqs. (3.40)-(3.41) until some convergence criterion is reached. In principle, there are several possible update strategies: we discuss below the two simplest ones, providing some details about the computational cost for each case.

Parallel update scheme In a parallel update scheme (PU), parameters $\left\{\boldsymbol{\gamma}^{(a)}, \boldsymbol{\Gamma}^{(a)}\right\}_{a\in F}$ are updated simultaneously at each iteration. In this scenario, at each iteration the full covariance matrix of the Gaussian distribution $\boldsymbol{\Sigma}$ needs to be inverted just once, and then all the cavity parameters $\left\{\boldsymbol{y}^{(a)}, \boldsymbol{S}^{(a)}\right\}_{a\in F}$ are computed by using Eqs. (3.20)-(3.21). The computational cost of one iteration scales like $O\left(N^3 + \sum_a 2^{|\partial a|}\right)$, where the cubic term comes from the matrix inversion's cost using standard Gaussian elimination [1]; conversely, $\sum_a 2^{|\partial a|}$ is the number of operations required to compute moments of marginal tilted distribution defined in Eqs. (3.23)-(3.24). However, on sparse topologies, the connectivity (i.e. the number of neighbours) of each factor node is finite w.r.t. N, so that the computational time required to compute marginal tilted moments is negligible w.r.t. the one required for the matrix inversion, and therefore the computational complexity of DC scales like $O\left(N^3\right)$ (for comparison, on sparse graphs BP's computational cost

[1] Other algorithms have been designed to perform matrix operations with a better asymptotic computational cost w.r.t. Gauss-Jordan decomposition, i.e. $O\left(N^\delta\right)$ where $2 < \delta < 3$ [37, 39]. However, the large prefactors in the running time make their implementation not feasible in practice.

is $O(N^{\alpha})$ with $\alpha < 2$). On the other hand, on fully connected models $|\partial a| \sim O(N)$ and the computational cost becomes exponentially with N, making the proposed scheme unfeasible for large sizes. An empirical estimation of the computational cost per iteration is shown in Figure 3.7.

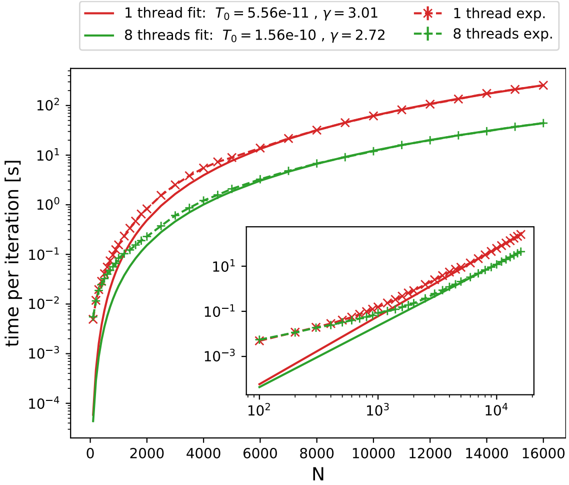

Figure 3.7: Computation time per iteration vs N where DC is used to estimate marginals on an Ising model defined on a Random Regular graph with degree $k = 5$. The results are averaged over 10 instances, shown in log-scale. The inset shows the same plot in log-log scale. The red points are obtained by using a single CPU thread, the green ones with 8 threads (i.e. the typical default setting of LinearAlgebra libraries when running on multi-thread). In both cases, the full lines represent the best fit curve using $T(N) = T_0 N^{\gamma}$ as fitting function, whose parameters are shown in the caption.

Random sequential update scheme In a random sequential update scheme (RSU), at each iteration we select a random permutation of the factor nodes: each time the parameters of one factor node a are updated, the full Gaussian density (3.7) is re-constructed. In this way, the update of parameters of factor ϕ_a is immediately encoded into the approximation, used for the next factor node in the (random) sequence. In this case, the computational cost becomes $O\left(N^3 M\right)$ on sparse graphs, where $M = |F|$ is the number of factor nodes. Therefore, RSU is slower in terms of computation time per iteration, but it typically requires a smaller number of iterations to converge, and it can be exploited in some regimes where the PU fails to do so.

As a final remark on the computational complexity, it should be possible in principle to exploit faster iterative methods designed to invert sparse matrices, in such a way to reduce the overall computational cost of the algorithm. These methods typically allow to retrieve a subset of the inverse matrix elements (e.g. the diagonal terms). Since the only elements needed by DC to compute tilted moments are the diagonal entries of $\mathbf{\Sigma}$ and the off-diagonal entries [47, 67, 83]: we leave this issue for future investigations.

3.3.1 Pseudocode implementation

A simple pseudocode implementation is shown in Algorithm 3.1 for the parallel update scheme. A Julia implementation of the code is available at [27].

Algorithm 3.1 Density Consistency

```
Input: set of compatibility functions $\{\psi_a\}_{a\in F}$, maximum tolerance $\varepsilon$, maximum number of iterations $\tau_{max}$, set of closure equations.
Initialize $\left\{\left(\boldsymbol{\gamma}^{(a)}\right)^{\tau=0}, \left(\boldsymbol{\Gamma}^{(a)}\right)^{\tau=0}\right\}_{a\in F}$
repeat for $\tau < \tau_{max}$
    compute $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ by using (3.11)
    for $a \in F$ do
        compute cavity fields and couplings $\boldsymbol{w}^{(a)}, \boldsymbol{S}^{(a)}$ from Eq. (3.20)-(3.21)
        compute tilted moments $\langle \boldsymbol{x}_a \rangle_{q^{(a)}}, \langle \boldsymbol{x}_a \boldsymbol{x}_a^t \rangle_{q^{(a)}}$
        compute the Gaussian moments $\boldsymbol{\mu}_{[\partial a]}$, $\boldsymbol{\Sigma}^{\tau}_{[\partial a, \partial a]}$ using (3.34) (or any other set of closures)
        update Gaussian parameters $\left(\boldsymbol{\gamma}^{(a)}\right)^{\tau+1}, \left(\boldsymbol{\Gamma}^{(a)}\right)^{\tau+1}$ using (3.40)-(3.41)
until convergence $\delta_* < \varepsilon$ where $\delta_* \in \{(3.65), (3.66)\}$
Output: tilted moments $\left\{\langle \boldsymbol{x}_a \rangle_{q^{(a)}}, \langle \boldsymbol{x}_a \boldsymbol{x}_a^t \rangle_{q^{(a)}}\right\}_{a\in F}$
```

With regard to the initial condition, the simplest choice is to initialize the Gaussian fields $\left\{\boldsymbol{\gamma}^{(a)}\right\}_{a\in F}$ random, and the precision matrices $\left\{\boldsymbol{\Gamma}^{(a)}\right\}_{a\in F}$ as identity matrices. The convergence criterion can be defined both w.r.t. Gaussian parameters and tilted moments, respectively defined as:

$$\delta_p = \max_{a\in F}\left\{\left|\left(\boldsymbol{\gamma}^{(a)}\right)^{\tau+1} - \left(\boldsymbol{\gamma}^{(a)}\right)^{\tau}\right| + \left|\left(\boldsymbol{\Gamma}^{(a)}\right)^{\tau+1} - \left(\boldsymbol{\Gamma}^{(a)}\right)^{\tau}\right|\right\}, \tag{3.65}$$

$$\delta_m = \max_{a\in F}\left\{\left|\langle \boldsymbol{x}_a\rangle^{\tau+1}_{q^{(a)}} - \langle \boldsymbol{x}_a\rangle^{\tau}_{q^{(a)}}\right| + \left|\langle \boldsymbol{x}_a\boldsymbol{x}_a^t\rangle^{\tau+1}_{q^{(a)}} - \langle \boldsymbol{x}_a\boldsymbol{x}_a^t\rangle^{\tau}_{q^{(a)}}\right|\right\}, \tag{3.66}$$

where $\langle\cdot\rangle^{\tau}$ is the expectation value computed with parameters at iteration τ. In particular, the converge condition over parameters given by (3.65) is typically stronger w.r.t. the moments' one (3.66), and it generally requires more iterations to be reached. To overcome eventual stability issues during the iteration, a damping $\rho \in [0,1]$ can be added. In this case, the update rule is modified as follows:

$$\lambda^{\tau+1} = \rho\lambda^{\tau} + (1-\rho)\,\lambda^{new} \qquad \lambda \in \left\{\boldsymbol{\gamma}^{(a)}, \boldsymbol{\Gamma}^{(a)}\right\}_{a\in F} \tag{3.67}$$

where λ^{new} is the proposed update given by (3.40)-(3.41). Notice that the same code can be used to compute BP fixed points on any graph topology, simply by using a closure that satisfies DC condition (3.28) and setting $\eta = 0$ to neglect cavity covariances.

3.4 Future directions

In the previous sections, we have presented a detailed description of the Density Consistency scheme, highlighting its main properties. An evaluation of its performances, compared to other approximation methods, will be discussed in the next chapter. In the following section, we present some preliminar discussion about two possible future directions: first, a generalization to non-binary variables is discussed in Section 3.4.1; secondly, we provide a preliminar variational formulation for the DC method, discussed in Section 3.4.2.

3.4.1 Generalization to non-binary variables

Density Consistency is based on the assumption that the probability values of a single-node discrete binary distribution can be fitted by a univariate Gaussian distribution, as described in Sec. 3.1.2. Even if the derivation presented so far has been carried out on probabilistic graphical models of binary spins, i.e. variables a symmetric support on $\{-1,1\}$, it is straightfoward to apply the same method any model defined on variables s_i having an arbitrary binary support over $\{a,b\}$, where $a,b\in\mathbb{R}$. Indeed, the following linear transformation can always be applied:

$$s_i \in \{a,b\} \longrightarrow s_i = \frac{b-a}{2}\sigma_i + \frac{b+a}{2} \qquad \sigma_i \in \{-1,1\}\,. \tag{3.68}$$

It is therefore sufficient to map the starting distribution of $p\left(\boldsymbol{s}\right)$ onto a binary distribution $p\left(\boldsymbol{\sigma}\right)$ by using Eq. (3.68) and apply the same machinery discussed before. On the other hand, when the degrees of freedom take more than 2 values there is no general way to fit single-node marginals with a univariate Gaussian distribution and DC condition (3.28) cannot be applied as such.
Many statistical physics models are defined by *multi-state* degrees of freedom, i.e. when the number of values that each variable s_i can take is > 2: for instance, the Blume-Emery-Griffiths model, introduced in [16] as a classical model to describe He$^3-$He^4 mixtures, is defined over $3-$states variables, namely $s_i \in \{-1,0,1\}$. The BEG belongs to the more general class of Potts-like models [113, 142], where the degrees of freedom are assumed to take k different values (often

named *colors*): the Potts model has recently found many interesting applications in computational biology, in particular for the structural inference in protein domains from coevolutionary sequences [36]. In addition, several combinatorial optimization problems are defined by graphical models of multistate variables, like graph-coloring [53, 96]. In the following, we discuss a possible way to generalize Density Consistency to multi-state variable models. Let us define a probabilistic graphical model on a factor graph $G = (V, E, F)$ in the same spirit as in the previous sections. This time, each node $i \in V$ represents a variable s_i taking values in a finite alfabet $\mathcal{X}$ with $k > 2$ states, namely $s_i \in \mathcal{X} = \{\theta_1, \dots, \theta_k\}$. The probability distribution of such a model, denoted with $p(\boldsymbol{s})$, can be written as:

$$p(\boldsymbol{s}) = \frac{1}{Z} \prod_{a \in F} \psi_a(\boldsymbol{s}_a) \tag{3.69}$$

Notice that the total number of configurations of (3.69) is k^N. A possible way to apply Density Consistency is to replace each $k-$valued variable s_i with a $k-$component vector of binary variables, denoted with $\boldsymbol{\sigma^i} \in \mathbb{R}^k$, where $\sigma^i_\alpha \in \{-1,1\}$ $\forall \alpha = 1, .., k$. This procedure increases the number of independent variables encoded by DC approximation to kN. However, in order to avoid including non-physical configurations, the following constraint needs to be imposed on each vector $\boldsymbol{\sigma^i}$:

$$\sum_{\alpha=1}^{k} \sigma^i_\alpha = 2 - k \qquad \forall i \in V \tag{3.70}$$

In this way, for each node i, only configurations of the type $\boldsymbol{\sigma}^i = \{1, -1, ..., -1\}$ (and its permutations) are allowed, in order to select just one among the k states. This procedure is also known as *one-hot encoding* in the machine learning community and it is illustrated in Figure 3.8 in the simplest case of $k = 3$. The name “one-hot” is justified by noticing that only configuration in which one the spins is up (*hot*) and the others are down (*cold*) are allowed; all the others - for instance, configurations where two spins are $+1$ - would imply that the corresponding variable s_i should be at the same time in two different states, and therefore they do not correspond to physical configurations of the starting model defined by (3.69). For each degree of freedom, the

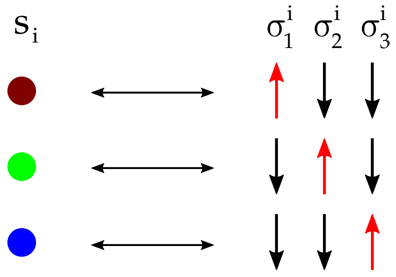

Figure 3.8: Illustration of the One-Hot encoding for a 3-colored variable.

constraints (3.70) can be implemented by adding a set of delta functions in the original probability distribution, which is now a function of $\{\boldsymbol{\sigma}^i\}_{i \in V}$. By rewriting Eq. (3.69) in terms of the new binary variables and taking into account for the constraints (3.70) we get:

$$p\left(\boldsymbol{\sigma}^1, \dots, \boldsymbol{\sigma}^N\right) = \frac{1}{Z} \prod_{a \in F} \psi_a\left(\{\boldsymbol{\sigma}^i\}_{i \in \partial a}\right) \prod_{i \in V} \delta\left(\sum_{\alpha=1}^{k} \sigma^\alpha_i - (2 - k)\right) \tag{3.71}$$

which is mathematically equivalent to (3.69). The correlations induced by these constraints on the spin components of each $\boldsymbol{\sigma^i}$ introduce short loops even when the original graph is a tree.

Neverthless, it is still possible to write a set of matching equation similar to (3.28). Density Consistency can be now applied in the same way as described in Section 3.1: to use the same notation as before, we introduce a set of kN continuous variables $\left\{\boldsymbol{x^i}\right\}_{i\in V} = \left\{x_\alpha^i\right\}_{i\in V}^{\alpha=1,\dots,k}$. In particular, the Gaussian measure $q\left(\left\{\boldsymbol{x^i}\right\}_{i\in V}\right)$ and the tilted distributions $q^{(a)}\left(\left\{\boldsymbol{x^i}\right\}_{i\in V}\right)$ can be defined as:

$$q\left(\left\{\boldsymbol{x^i}\right\}_{i\in V}\right) \propto \prod_a \phi_a\left(\left\{\boldsymbol{x}^i\right\}_{i\in\partial a}\right) \tag{3.72}$$

$$q^{(a)}\left(\left\{\boldsymbol{x^i}\right\}_{i\in V}\right) \propto \left(\prod_{b\neq a}\phi_b\right)\Psi_a\left(\left\{\boldsymbol{x}^i\right\}_{i\in\partial a}\right) \tag{3.73}$$

where again the functions ϕ_a are Gaussian densities defined as in (3.5), with the only difference that each of them is now a multivariate normal distribution of $k \times |\partial a|$ variables. The tilted distributions $q^{(a)}$ will encode the linear constraints (3.70) for each $i \in \partial a$:

$$\Psi_a\left(\left\{\boldsymbol{x}^i\right\}_{i\in\partial a}\right) = \psi_a\left(\left\{\boldsymbol{x}^i\right\}_{i\in\partial a}\right)\prod_{i\in\partial a}\delta\left(\sum_{\alpha=1}^k x_i^\alpha - (2-k)\right) \tag{3.74}$$

DC condition

DC condition can now be generalized to multistate variables by imposing a matching of the density values between the distributions (3.72) Eq. (3.73) marginalized over all the variables except the spin-vector $\boldsymbol{x}^i$, for $i \in \partial a$:

$$q^{(a)}\left(\boldsymbol{x}^i\right) \propto q\left(\boldsymbol{x}^i\right) \qquad \forall i \in \partial a, a \in F$$

The density matching holds on the support defined by $\boldsymbol{x}^i = \{1, -1, ..., -1\}$ and its permutations, which gives the following k equations:

$$\begin{gathered} q^{(a)}\left(\boldsymbol{x}^i = \{1,-1,\dots,-1\}\right) \propto q\left(\boldsymbol{x}^i = \{1,-1,\dots,-1\}\right) \\ q^{(a)}\left(\boldsymbol{x}^i = \{-1,1,\dots,-1\}\right) \propto q\left(\boldsymbol{x}^i = \{-1,1,\dots,-1\}\right) \\ \vdots \\ q^{(a)}\left(\boldsymbol{x}^i = \{-1,-1,\dots,1\}\right) \propto q\left(\boldsymbol{x}^i = \{-1,-1,\dots,1\}\right) \end{gathered} \tag{3.75}$$

Notice also that the above generalized DC condition can be used to prove Theorem (1) under the same hypothesis in the case of multistate variables. However, it is not clear yet how to generalized the set of closure equations (3.34) to this scenario, where both single-site correlation between different "colors" (i.e. terms like $\Sigma_{ii}^{\alpha\beta}$) and nearest-neighbours correlation (i.e. $\Sigma_{ij}^{\alpha\beta}$) need to be fixed. Notice also that in the above system (3.75) only $k-1$ equations need to be explicitly solved, and the last one will be automatically satisfied by normalization. In this perspective, it is probably necessary to define the one hot encoding only w.r.t. $k-1$ variables, otherwise the Gaussian covariance matrix of (3.72) would have a null determinant since a number N of its rows would be linearly dependent on the remaining $N(k-1)$ (this should be equivalent to fix the so-called *lattice gas gauge* in Potts-like models [36]). We leave these issues for a future investigation.

3.4.2 Towards a variational formulation

In this final section, we discuss a possible way to derive a variational approach to Density Consistency. The idea is to derive a free energy as function of the Gaussian parameters so that its stationary points coincide with the DC fixed points, in the same way as carried out in Section 2.4.2 for Gaussian EP. Let us start by rewriting the tilted distribution (3.12) in terms of the full Gaussian measure and the single-node constrained distributions $\hat{q}^{(i)}$ defined in Eq. (3.29):

$$
\begin{aligned}
q^{(a)}\left(\boldsymbol{x}\right) &= \frac{1}{Z_a}\left[\prod_{b\neq a}\phi_b\right]\psi_a\prod_{i\in\partial a}\Delta_i \times \frac{\phi_a\prod_{i\in\partial a}\hat{Z}_i\left(\prod_c\phi_c\right)^{|\partial a|-1}}{\phi_a\prod_{i\in\partial a}\hat{Z}_i\left(\prod_c\phi_c\right)^{|\partial a|-1}}\\
&= \frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_a}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(\boldsymbol{x}\right)}{\left(\prod_c\phi_c\right)^{|\partial a|-1}}\frac{\psi_a}{\phi_a}\\
&= \frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_a}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(\boldsymbol{x}\right)}{\left(\prod_c\phi_c\right)^{|\partial a|-1}}\frac{\psi_a}{\phi_a}\times\frac{Z_q^{|\partial a|-1}}{Z_q^{|\partial a|-1}}\\
&= \frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_aZ_q^{|\partial a|-1}}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(\boldsymbol{x}\right)}{q\left(x\right)^{|\partial a|-1}}\frac{\psi_a}{\phi_a}
\end{aligned}
\tag{3.76}
$$

where $\hat{Z}_i$ (resp. Z_q) is the partition function of $\hat{q}^{(i)}$ (resp. q). Now, by taking the product of all the tilted distributions written as in Eq. (3.76), we get:

$$
\begin{aligned}
\prod_a q^{(a)}\left(\boldsymbol{x}\right) &= \prod_a\left[\frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_aZ_q^{|\partial a|-1}}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(\boldsymbol{x}\right)}{q\left(\boldsymbol{x}\right)^{|\partial a|-1}}\frac{\psi_a}{\phi_a}\right] =\\
&= \prod_a\left[\frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_aZ_q^{|\partial a|-1}}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(\boldsymbol{x}\right)}{q\left(\boldsymbol{x}\right)^{|\partial a|-1}}\right]\times\prod_a\frac{\psi_a}{\phi_a}\times\frac{Z\prod_i\Delta_i}{Z\prod_i\Delta_i}\\
&= \prod_a\left[\frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_aZ_q^{|\partial a|-1}}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(x\right)}{q\left(\boldsymbol{x}\right)^{|\partial a|-1}}\right]\times\frac{p\left(\boldsymbol{x}\right)Z}{\prod_a\phi_a\prod_i\Delta_i}\times\frac{\left(\prod_a\phi_a\right)^{N-1}}{\left(\prod_a\phi_a\right)^{N-1}}\\
&= \prod_a\left[\frac{\prod_{i\in\partial a}\hat{Z}_i}{Z_aZ_q^{|\partial a|-1}}\frac{\prod_{i\in\partial a}\hat{q}^{(i)}\left(x\right)}{q\left(\boldsymbol{x}\right)^{|\partial a|-1}}\right]\times\frac{p\left(\boldsymbol{x}\right)Z}{\prod_i\hat{Z}_i\hat{q}^{(i)}\left(\boldsymbol{x}\right)}\times\left[q\left(\boldsymbol{x}\right)Z_q\right]^{N-1}\\
&= p\left(\boldsymbol{x}\right)Z\frac{\prod_i\left[\hat{Z}_i\hat{q}^{(i)}\right]^{d_i-1}}{\left(\prod_a Z_a\right)\left[Z_q q\left(\boldsymbol{x}\right)\right]^{(1-N)+\sum_a(|\partial a|-1)}}
\end{aligned}
\tag{3.77}
$$

Using the above expression and rewriting it to isolate the distribution of the original model $p\left(\boldsymbol{x}\right)$, we get:

$$
\begin{aligned}
p\left(\boldsymbol{x}\right) &= \frac{\prod_a Z_a q^{(a)}\left(\boldsymbol{x}\right)\left[Z_q q\left(\boldsymbol{x}\right)\right]^{(1-N)+\sum_a(|\partial a|-1)}}{Z\prod_i\left[\hat{Z}_i\hat{q}^{(i)}\left(\boldsymbol{x}\right)\right]^{d_i-1}}\\
&= \frac{\prod_a Z_a\left[Z_q\right]^{(1-N)+\sum_a(|\partial a|-1)}}{Z\prod_i\hat{Z}_i^{d_i-1}}\frac{\prod_a q^{(a)}\left(x\right)\left[q\left(\boldsymbol{x}\right)\right]^{(1-N)+\sum_a(|\partial a|-1)}}{\prod_i\left[\hat{q}^{(i)}\left(\boldsymbol{x}\right)\right]^{d_i-1}} && (3.78)\\
&= \tilde{p}\left(\boldsymbol{x}\right)\frac{Z_{DC}}{Z} && (3.79)
\end{aligned}
$$

where

$$
\tilde{p}\left(\boldsymbol{x}\right) = \frac{\prod_a q^{(a)}\left(x\right)\left[q\left(\boldsymbol{x}\right)\right]^{(1-N)+\sum_a(|\partial a|-1)}}{\prod_i\left[\hat{q}^{(i)}\left(\boldsymbol{x}\right)\right]^{d_i-1}} \tag{3.80}
$$

$$
Z_{DC} = \frac{\prod_a Z_a\left[Z_q\right]^{(1-N)+\sum_a(|\partial a|-1)}}{\prod_i\hat{Z}_i^{d_i-1}} \tag{3.81}
$$

The ratio $\alpha = Z_{DC}/Z$ will be equal to 1 when DC is exact, which occurs on acyclic graphs, where Density Consistency coindice with the Bethe Approximation. This allows to define the DC free energy as follows:

$$F_{DC} = -\log Z_{DC} = -\sum_a \log Z_a + \sum_i (d_i - 1)\log \hat{Z}_i - \left[(1-N) + \sum_a (|\partial a| - 1)\right] \log Z_q \tag{3.82}$$

Notice that, on a (undirected) factor graph, the sum of the degrees of variable nodes is equal to the sum of the degrees of all factor nodes, namely $\sum_a |\partial a| = \sum_i d_i$. This allows to simplify the exponent of Z_q as:

$$(1-N) + \sum_a (|\partial a| - 1) = \sum_i d_i - M - N + 1 \hat{=} \Xi \tag{3.83}$$

where $N = |V|$ and $M = |F|$.

Stationary points

We now compute the stationary condition of F_{DC} by setting to 0 its derivatives w.r.t. the Gaussian parameters of each factor node a:

$$\frac{\partial F_{DC}}{\partial \gamma_i^{(a)}} = -\sum_{b \neq a} \langle x_i \rangle_{q^{(b)}} + \sum_j (d_j - 1) \langle x_i \rangle_{\hat{q}^{(j)}} = 0 - \Xi \langle x_i \rangle_q \quad \forall i \in \partial a \tag{3.84}$$

$$\frac{\partial F_{DC}}{\partial \Gamma_{ij}^{(a)}} = -\sum_{b \neq a} \langle x_i x_j \rangle_{q^{(b)}} + \sum_k (d_j - 1) \langle x_i x_j \rangle_{\hat{q}^{(k)}} - \Xi \langle x_i x_j \rangle_q = 0 \quad \forall i, j \in \partial a \tag{3.85}$$

In the above expression, most of the terms in the two summations refer to expectation values of variables computed w.r.t. a tilted distribution defined on other factor nodes not connected to them. For simplicity, we rewrite the first stationary condition to highlight their contribution:

$$\frac{\partial F_{DC}}{\partial \gamma_i^{(a)}} = -\sum_{\substack{b \neq a \\ i \in \partial b}} \langle x_i \rangle_{q^{(b)}} - \sum_{\substack{b \neq a \\ i \notin \partial b}} \langle x_i \rangle_{q^{(b)}} + (d_i - 1) \langle x_i \rangle_{\hat{q}^{(i)}} + \sum_{j \neq i} (d_j - 1) \langle x_i \rangle_{\hat{q}^{(j)}} - \Xi \langle x_i \rangle_q \tag{3.86}$$

In particular, the second term represents the sum of expectation values of variable x_i computed over tilted distributions defined on factor node not connected to i. Neverthless, all these moments can be analytically evaluated by exploting properties of Gaussian integration and we refer to Appendix A for a more detailed discussion.

Stationarity over fields

In the following, we will show that DC closure equations satisfy Eq. (3.84). The computation of these "mixed" tilted moments is discussed in Appendix A, we report here the final results for the first-order moments:

$$\langle x_i \rangle_{q^{(b)}} = \mu_i - \sum_{k \in \partial b} \frac{\left[\mathcal{S}^{(\partial b \cup i)}\right]_{ik}}{\left[\mathcal{S}^{(\partial b \cup i)}\right]_{ii}} \left[\langle x_k \rangle_{q^{(b)}} - \mu_k\right] \quad \forall b : i \notin \partial b \tag{3.87}$$

$$\langle x_i \rangle_{\hat{q}^{(j)}} = \mu_i - \frac{\left[\mathcal{S}^{(j \cup i)}\right]_{ij}}{\left[\mathcal{S}^{(j \cup i)}\right]_{ii}} \left(\langle x_j \rangle_{q^{(j)}} - \mu_j\right) \quad \forall j \neq i \tag{3.88}$$

where the matrix $\mathcal{S}^{(\partial b \cup i)}$ is the inverse of the sub-block of the covariance matrix of the full Gaussian measure, namely $\mathcal{S}^{(\partial b \cup i)} = \left(\boldsymbol{\Sigma}_{[\partial b \cup i, \partial b \cup i]}\right)^{-1}$ (the same holds for $\mathcal{S}^{(j \cup i)}$). Inserting Eqs

(3.87)-(3.88) into (3.86), and using (3.28)-(3.34a) (respectively, the set of DC condition and the first moment matching equations) and after some straighforward algebra, we get the following expression for the stationarity condition (3.86):

$$\begin{aligned}\frac{\partial F_{DC}}{\partial \gamma_i^{(a)}} = & -\sum_{b\neq a, i\in\partial b} \langle x_i\rangle_{q^{(b)}} + \sum_{b\neq a, i\notin\partial b}\sum_{k\in\partial b} \frac{\left[\mathcal{S}^{(\partial b\cup i)}\right]_{ik}}{\left[\mathcal{S}^{(\partial b\cup i)}\right]_{ii}} \left[\langle x_k\rangle_{q^{(b)}} - \mu_k\right] + \\ & (d_i - 1)\langle x_i\rangle_{q^{(i)}} - \sum_{j\neq i}(d_j - 1)\frac{\left[\mathcal{S}^{(j\cup i)}\right]_{ij}}{\left[\mathcal{S}^{(j\cup i)}\right]_{ii}}\left(\langle x_j\rangle_{q^{(j)}} - \mu_j\right) + \\ & -\left[\Xi + (M - d_i) - \left(\sum_j d_j - d_i - (N-1)\right)\right]\mu_i = 0\end{aligned}$$

where we used $\langle x_i\rangle_q = \mu_i$ as in Eq. (3.8). Notice now that, using the definition of $\Xi = \sum_i d_i - M - N + 1$, the last term disappears. Finally, using both DC condition and the first moment matching (resp. Eqs (3.28)-(3.34a)), the above expression becomes identically satisfied. Therefore, given any set of closure equations satisfying DC condition and the first moment matching, the free energy (3.82) will be stationary with respect to variations of the linear terms $\left\{\gamma^{(a)}\right\}_{a\in F}$. The set of stationarity conditions w.r.t. to quadratic terms (3.85) involve similar (but way longer) calculations, that depend also on the closure equation used to fix 2−points correlations. We verified that the matching of Pearson correlation coefficient (3.34c) does *not* imply stationarity of (3.85). In principle, we expect that the free energy will be stationary by choosing another suitable consistency condition over 2-point correlations: however, this issue is still an open problem and we leave it for future investigations.

Chapter 4

Results : forward problem

This chapter presents a series of numerical and analytic results obtained through Density Consistency, in comparison to other approximation methods. The whole chapter focuses on the *forward* (or *direct*) problem, i.e. the estimation of marginal probabilities from a known model. In particular, Section (4.1) presents some results on finite size systems, with a particular focus on Ising-like models: in this case, Density Consistency is discussed in the simplified setup of pairwise graphical models. In addition, some preliminar results about combinatorial optimization problems are discussed in Sec (4.2). Finally, Section 4.3 presents an analytic solution based on Density Consistency for the Ferromagnetic Ising model on hypercubic lattices in the thermodynamic limit.

4.1 Ising Model

4.1.1 Density Consistency approach

A generic pairwise graphical model of binary spins can always be written in terms of an Ising Hamiltonian. In this case, it is possible to simplify the derivation of Density Consistency because the factor graph representation reduces to a simple graph. Therefore, we are going to recall the main steps of the derivation presented in Chapter 3, that will be extensively used in Sec. 4.3 as well as in Chapter 5 in the context of the Inverse Ising Problem.
As discussed in Section 1.2, the Ising model can be defined on an arbitrary graph $G = (V, E)$ with $N = |V|$ nodes and a set of edge links E. On each node a discrete variable is defined, taking values in $\{-1,1\}$. On each edge $(ij) \in E$ a real quantity J_{ij} identifies the pairwise interaction between spin i and j; edges are assumed to be undirected, so that the matrix defined by the whole set of couplings $\boldsymbol{J} = \{J_{ij}\}_{i,j\in V}$ is symmetric by construction[1]. Moreover, a local external field, denoted with h_i, acts on each spin i. At a certain inverse temperature β, the equilibrium probability distribution is expressed by:

$$p\left(\boldsymbol{x}\right) = \frac{1}{Z} \exp\left[\beta \sum_{\langle i,j\rangle} J_{ij} x_i x_j + \beta \sum_i h_i x_i\right] \prod_i \Delta_i\left(x_i\right) \tag{4.1}$$

Notice that where we already defined (4.1) w.r.t. continuous variables $x_i \in \mathbb{R}$ as in (3.3) and we introduced the constraints $\Delta_i\left(x_i\right)$, defined by (3.4), such that the probability distribution has a

[1] In a matrix notation, a coupling J_{ij} is assumed to be zero for a non-existing edge in the graph, i.e $J_{ij} = 0 \, \forall \, (i,j) \notin E$

non zero-measure only over the binary support $\{-1,1\}^N$. We rewrite (4.1) in a factorized form

$$p(\boldsymbol{x}) \propto \prod_{\langle i,j\rangle} \psi_{ij}(x_i, x_j) \prod_i \Delta_i(x_i) \tag{4.2}$$

where

$$\psi_{ij}(x_i, x_j) = \exp\left[\beta J_{ij} x_i x_j + \beta h_i^{(ij)} x_i + \beta h_j^{(ij)} x_j\right] \tag{4.3}$$

In this notation, the quantity $h_i^{(ij)}$ (resp. $h_j^{(ij)}$) denotes a certain fraction of the local field h_i (resp. h_j) contained into the factor ψ_{ij}. In this way, external fields can be distributed among all the factors corresponding to n.n. pairs, with the following constraints:

$$\sum_{j\in\partial i} h_i^{(ij)} = h_i,\ \forall i \tag{4.4}$$

The simplest choice is to uniformly distribute the local field h_i among its neighbours, i.e. $h_i^{(ij)} = h_i/|\partial i|$. Although this choice may seem heuristic, notice that, thanks to the weight gauge property described in Sec. (3.2.3), the way in which local fields are distributed among factors is irrelevant for DC scheme, as it is always possible to move in and out Gaussian densities from the factors ψ_{ij}. This notation will be also be useful in the next chapter.
Density Consistency's derivation can be carried out in the same way as in Chapter 3, by identifying each factor node a with the link (ij). In particular, the approximating Gaussian densities are now defined for each edge (ij) in the graph:

$$\phi_{ij}(x_i, x_j) = \exp\left[-\frac{1}{2}(x_i, x_j)\boldsymbol{\Gamma}^{(ij)}(x_i, x_j)^t + (x_i, x_j)\boldsymbol{\gamma}^{(ij)}\right] \quad \forall (ij) \in E \tag{4.5}$$

Here $\boldsymbol{\Gamma}^{(ij)}$ is a 2×2 matrix, $\boldsymbol{\gamma}^{(ij)}$ is a 2−components column vector defined by:

$$\boldsymbol{\Gamma}^{(ij)} = \begin{pmatrix} \Gamma_{ii}^{(ij)} & \Gamma_{ij}^{(ij)} \\ \Gamma_{ij}^{(ij)} & \Gamma_{jj}^{(ij)} \end{pmatrix}, \qquad \boldsymbol{\gamma}^{(ij)} = \begin{pmatrix} \gamma_i^{(ij)} \\ \gamma_j^{(ij)} \end{pmatrix} \qquad \forall (i,j) \in E \tag{4.6}$$

For simplicity, we recall the definition of the full Gaussian measure $q(\boldsymbol{x})$,obtained by taking the product of all the 4.5, and the set of tilted distributions, each one defined on a particular edge (ij), and denoted with $q^{(ij)}(\boldsymbol{x})$:

$$q(\boldsymbol{x}) \propto \prod_{i<j} \phi_{ij}(x_i, x_j) \propto \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \tag{4.7}$$

$$q^{(ij)}(\boldsymbol{x}) \propto \prod_{\substack{(k,l)\in E\\(k,l)\neq(ij)}} \phi_{kl}\Psi_{ij} \propto q(\boldsymbol{x})\frac{\Psi_{ij}}{\phi_{ij}}, \quad \forall (i,j) \in E \tag{4.8}$$

where $\Psi_{ij} = \psi_{ij}\Delta_i\Delta_j$. As described in the previous chapter, each tilted distribution can be used to estimate marginal densities over the variables (ij), where a correction term arises from the presence of the Gaussian cavity distribution $g^{\setminus(ij)} \propto q/\phi_{ij}$. In particular, Eq. 4.8 can be marginalized over all the set of variables but (x_i, x_j) by a standard Gaussian integration, leading to the following marginal density:

$$q^{(ij)}(x_i, x_j) = \int dx_{\setminus i,j} q^{(ij)}(\boldsymbol{x}) \propto g^{\setminus(ij)}(x_i, x_j)\Psi_{ij}(x_i, x_j), \tag{4.9}$$

where $g^{\setminus ij}(x_i, x_j)$ is the marginal cavity distribution, defined as follows:

$$g^{\setminus(ij)}(x_i, x_j) \propto \int d\boldsymbol{x}_{\setminus ij}\frac{q(\boldsymbol{x})}{\phi_{ij}(x_i, x_j)} \propto \exp\left[-\frac{1}{2}(x_i, x_j)\boldsymbol{S}^{(ij)}(x_i, x_j)^t + (x_i, x_j)\boldsymbol{w}^{(ij)}\right] \tag{4.10}$$

with $\boldsymbol{S}^{(ij)}$ being a 2×2 matrix, $\boldsymbol{w}^{(ij)}$ being a 2−components vector. Their expression follows directly from the general expression (3.20)-(3.21):

$$w_i^{(ij)} = \frac{\Sigma_{jj}\mu_i - \Sigma_{ij}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \gamma_i^{(ij)} \tag{4.11}$$

$$w_j^{(ij)} = \frac{-\Sigma_{ij}\mu_i + \Sigma_{ii}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \gamma_j^{(ij)} \tag{4.12}$$

$$S_{ij}^{(ij)} = \frac{-\Sigma_{ij}}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \Gamma_{ij}^{(ij)} \tag{4.13}$$

For simplicity, the diagonal elements of $\boldsymbol{S}^{(ij)}$ are not shown, since they do not contribute to the tilted marginals. The moments of (4.9) can be easily computed by performing a summation over $\{-1,1\}^2$:

$$\langle x_i \rangle_{q^{(ij)}} = \tanh\left[a_i^{(ij)} + \operatorname{atanh}\left(\tanh b^{(ij)} \tanh a_j^{(ij)}\right)\right] \tag{4.14a}$$

$$\langle x_j \rangle_{q^{(ij)}} = \tanh\left[a_j^{(ij)} + \operatorname{atanh}\left(\tanh b^{(ij)} \tanh a_i^{(ij)}\right)\right] \tag{4.14b}$$

$$\langle x_i x_j \rangle_{q^{(ij)}} = \tanh\left[b^{(ij)} + \operatorname{atanh}\left(\tanh a_i^{(ij)} \tanh a_j^{(ij)}\right)\right] \tag{4.14c}$$

with

$$a_i^{(ij)} = \beta h_i^{(ij)} + w_i^{(ij)} \tag{4.15a}$$

$$a_j^{(ij)} = \beta h_j^{(ij)} + w_j^{(ij)} \tag{4.15b}$$

$$b^{(ij)} = \beta J_{ij} - S_{ij}^{(ij)} \tag{4.15c}$$

It is even more clear how the cavity distribution modifies the tilted moments by the addition of an effective coupling $-S_{ij}^{(ij)}$(i.e. the off-diagonal term of the coupling cavity matrix $\boldsymbol{S}^{(ij)}$), as well as cavity fields $w_i^{(ij)}, w_j^{(ij)}$. When neglecting cavity correlations, DC scheme provides BP fixed points on any graph topology, as discussed in the previous chapter: in this case, the cavity field $w_i^{(ij)}$ coincides to the cavity message defined in (2.41) (and similarly for $w_j^{(ij)}$).

Weight gauge

Since the Ising Hamiltonian contains only linear and quadratic terms, the Boltzmann measure 4.1 can be written as a Gaussian density in terms of continuous variables. This allows to exploit the weight gauge property discussed in Sec. 3.2.3. Following that notation, we rewrite the original probability distribution as follows:

$$p(\boldsymbol{x}) = \frac{1}{Z}\Phi(\boldsymbol{x}) \prod_{\langle i,j\rangle} \psi'_{ij}(x_i,x_j) \prod_i \Delta_i(x_i) \tag{4.16}$$

$$\Phi(\boldsymbol{x}) = \exp\left[-\frac{1}{2}\boldsymbol{x}^t (-\beta \boldsymbol{J})^{-1} \boldsymbol{x} + \beta \boldsymbol{h}^t \boldsymbol{x}\right] \tag{4.17}$$

with $\psi'_{ij}(x_i,x_j) = 1$. The above parametrization allows to apply Density Consistency by moving all the terms in the Hamiltonian outside the factors ψ_{ij} and reparametrize the Gaussian and titled distributions as:

$$q(\boldsymbol{x}) \propto \Phi(\boldsymbol{x}) \prod_{\langle i,j\rangle} \phi'_{ij}(x_i,x_j) \tag{4.18}$$

$$q^{(ij)}(\boldsymbol{x}) \propto \Phi(\boldsymbol{x}) \prod_{\substack{(k,l)\in E\\(k,l)\neq(ij)}} \phi'_{ij}(x_i,x_j)\,\Psi'_{ij}(x_i,x_j) \tag{4.19}$$

where $\Psi_{ij}^{'} = \Delta_i \Delta_j$. We remark that running DC in this setup gives the same fixed points as before, on any graph topology and under any set of closure equations.

Equivalent parametrization by adding univariate factors

Another equivalent parametrization of DC scheme can be performed by moving the local external fields onto a separate set univariate factors. In this way, the probability distribution of the Ising model can be expressed in the following form:

$$p(\boldsymbol{x}) \propto \prod_{\langle i,j \rangle} \psi_{ij}^{0}(x_i, x_j) \prod_i \psi_i(x_i) \prod_i \Delta_i(x_i) \tag{4.20}$$

where now the factors ψ_{ij}^0 contain only the interaction term, namely $\psi_{ij}^0(x_i, x_i) \propto e^{\beta J_{ij} x_i x_j}$, and the local fields are included into $\psi_i(x_i) \propto e^{\beta h_i x_i}$. DC's derivation follows in the same way as previously discussed, with the difference that now an additional set of univariate Gaussian density ϕ_i is included:

$$\phi_i(x_i) \propto \exp\left[-\frac{1}{2}\Gamma^{(i)} x_i^2 + \gamma^{(i)} x_i\right] \qquad \forall i \in V \tag{4.21}$$

DC scheme is now defined by an equivalent family $\left(\left\{\phi_{ij}^0\right\}_{(ij)\in E}, \left\{\phi_i\right\}_{i \in V}\right)$. Also in this case, the fixed points are the same as in the previous parametrization: the reason is the set of DC closures is trivially solved in the case of univariate factors. Indeed, consider the DC condition (3.28) between the marginal tilted distribution defined w.r.t. the single-site factor i and denoted with $q^{(i)}$:

$$\begin{gathered} q^{(i)}(x_i) \propto q(x_i) \\ g^{\setminus i}(x_i) \times \psi_i(x_i) \propto g^{\setminus i}(x_i) \times \phi^{(i)}(x_i) \end{gathered}$$

which is trivially solved by $\gamma^{(i)} = \beta h_i$ and independently on the additional parameter $\Gamma^{(i)}$. The last expression follows directly from the definitions of the tilted distribution and the full Gaussian measure, both of them marginalized over all the variables but x_i.

4.1.2 Results

In this section we show some results for the Ising model on different regimes, by varying the graph topology and the distribution of couplings and external fields. In order to compare the results with a ground truth estimate, we performed long Monte Carlo (MC) simulations using the Gibbs sampling procedure [55] with a total number of sampled configurations $M = 10^6$. The dynamics is run for $2M$ steps, the first half needed to equilibrate the MC dynamics. We remark that each "step" here corresponds to N Gibbs-sampling sweeps, each time performed on a random permutation of the spins. By denoting the set of sampled configurations with $\{\boldsymbol{\sigma}^\mu\}_{\mu=1}^M$, the first and second moments experimental moments can be computed as:

$$m_i^{MC} = \langle \sigma_i \rangle_{MC} = \frac{1}{M} \sum_{\mu=1}^{M} \sigma_i^\mu \tag{4.22}$$

$$\chi_{ij}^{MC} = \langle \sigma_i \sigma_j \rangle_{MC} = \frac{1}{M} \sum_{\mu=1}^{M} \sigma_i^\mu \sigma_j^\mu \tag{4.23}$$

and the connected-correlations are defined as $C_{ij} = \chi_{ij} - m_i m_j$. All DC simulations have been performed by using a numerical precision $\varepsilon = 10^{-7}$ on the tilted moments, with a damping parameter $\rho = 0.95$ to improve convergence. A first subset of results is shown in Figure 4.1. In this case, all of the instances refer to ferromagnetic models with heterogeneous couplings, i.e. $J_{ij} > 0 \; \forall \, (i,j) \in E$, without external fields; therefore, magnetizations are null and not shown. In all the four panels we scatter plot the nearest neighbour correlations computed by DC and BP w.r.t. the Monte Carlo values. As expected, DC turns out to significantly improve BP's estimate of correlations in all the cases analyzed. A particular instructive case is the panel (a) of Figure (4.1), where the model is defined on a Random Regular Graph of fixed degree, with constant couplings (i.e. $J_{ij} = J$): since the connectivity of all nodes are the same, BP's estimate of correlations is equal for each pair of adjacent spins; neverthless, the graph contains a certain number of (long) loops, their non-negligible contribution (at finite size) is well captured by DC thanks to the presence of cavity couplings, as discussed before. Another set of results is shown in

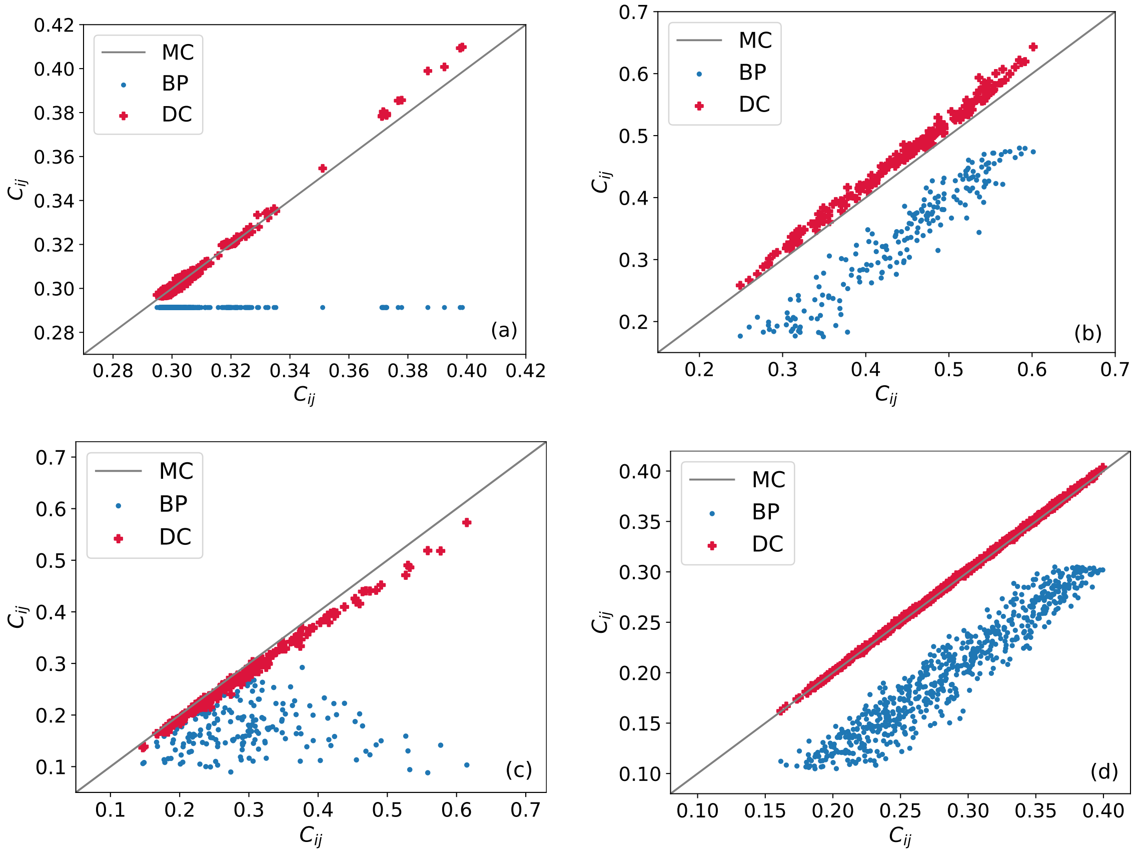

Figure 4.1: Comparison of DC and BP on single-instances ferromagnetic Ising models with null external fields. (a) Random Regular Graph (RRG) with $N = 300$, constant degree $k = 4$ and $\beta = 0.3$. (b) 2−dimensional square lattice with periodic boundary conditions (PBC), $N = 10^2$ and $\beta = 0.35$. (c) . Barabasi-Albert graph, $N = 100$, $n_0 = k = 2$ (the solution is found by using $\eta^* = 0.95$ and it is divergent for $\eta > \eta^*$). (d) 3−dimensional cubic lattice with PBC, $N = 6^3$and $\beta = 0.21$. Except for panel (a), in all the other instances couplings are drawn from a uniform distribution in (0.5,1.5). All the panels represent the scatter plot of nearest neighobурs correlations obtained through BP and DC, compared to Monte Carlo estimates.

Figure 4.2 for three spin glass models with binary interactions, so that on each edge the coupling J_{ij} is sampled in $\{-1,1\}$ with equal probability. Each of the panels in Figure 4.2 shows a typical instance for three different architectures: a scale-free graph generated through the Barabasi-Albert [8] model (panel a), a random regular graph (panel b) and a $3d$ lattice (panel c). Again, each panel shows the scatter plot of the nearest-neighbours correlations against MC estimates (in zero field). Since there are two values for the couplings, BP will estimate only two possible correlations between any pair of n.n. spins, equal apart from the sign. On the other hand, DC estimates take into account the presence of all the loops, giving reasonably good estimates for the correlations compared to MC. On locally tree-like graphs, we expect the loop contributions encoded by DC through the Gaussian cavities to be non-negligible only for a system with finite size: on the other hand, by increasing the number of nodes, their effect becomes negligible since the typical loop length increases as $\log N$, and BP becomes asymptotically exact in the limit $N \to \infty$. This behaviour is confirmed on Figure 4.3, where the cavity couplings $S_{ij}^{(ij)}$ predicted by DC are computed for a ferromagnetic Ising model defined on a random regular graph, with increasing size N. In this regime, independently on the temperature, the cavity couplings display a power-law decay. The same qualitative behaviour also holds for other tree-like topologies (e.g. Erdos-Reny), and by choosing different distributions of the couplings (e.g. spin-glass). We conclude the present section by showing a preliminar set of results where DC is compared against Linear Response (LR) techniques. In principle, correlations of arbitrary length can be estimated using LR onto any mean-field like approximation (even in the naif Mean field theory, where correlations are not taken into account within the trial probability distribution). In particular, we now evaluate LR correlations computed with respect to the Bethe Approximation: this approach was first derived by Welling and Teh [138], who designed an iterative message-passsing algorithm defined on cavity susceptibilities, known as Susceptibility Propagation (SP). The approach we use in the following is based on an analytic solution to compute the full covariance matrix of a known Ising model w.r.t. a BP fixed point, first developed in [116]. This approach allows to retrieve LR correlations without running an iterative algorithm, thus being much faster (it only requires a single matrix inversion) and free from numerical convergence issues. Moreover, it can be easily extended to infer couplings and fields from a series of data: for this reason, we will discuss it more in details in Chapter 5 in the context of the inverse Ising problem. The advantage of SP is that the true covariance matrix is retrieved whenever the starting graph is acyclic, so that both n.n. and long range-correlations can be exactly computed. Figure 4.4 shows two example scatter plots by comparing n.n. correlations (in zero field) computed through Susceptibility Propagation. In particular, the two instances are the same as Figure 4.1 panel (a) and Figure 4.2 panel (c), respectively. SP is able to significantly improve the estimation of couplings w.r.t. standard BP, and it estimates an heterogeneous set

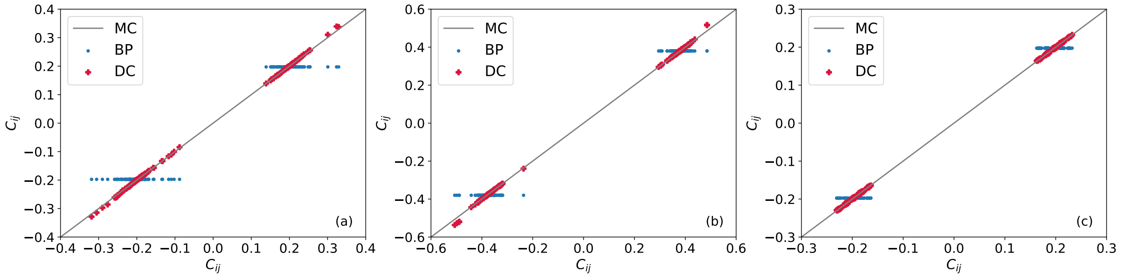

Figure 4.2: Comparison of DC and BP on single-instances spin glass models with null external fields. (a) Barabasi-Albert graph, $N = 100$, $n_0 = k = 2$, $\beta = 0$. (b) Random Regular Graph (RRG) with $N = 100$, constant degree $k = 4$ and $\beta = 0.4$. (c) 3−dimensional cubic lattice with PBC, $N = 6^3$and $\beta = 0.2$. In all the three models, couplings are binary, i.e. $J_{ij} \pm 1$ with equal probabilty. All the figures represent the scatter plot of nearest neighbours correlations obtained through BP and DC, compared to Monte Carlo estimates.

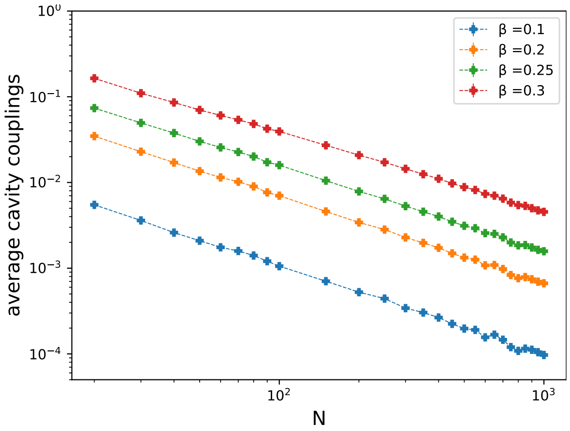

Figure 4.3: Cavity couplings as estimated by DC at fixed point (through (5.56)) w.r.t. system size N. The topology is a random regular graph with fixed connectivity $k = 4$ and constant couplings $J_{ij} = 1$ among n.n. spins. The plot shows the average cavity couplings $-S_{ij}^{(ij)}$ on all the n.n. pairs of spins, for 4 different values of β (shown in the legend), in log-log scale; mean and stardand errors are computed over 20 instances, by varying the graph topology.

of n.n. correlations due to the random structure of the graph in the left panel of Figure 4.4, and because of the random sign of the couplings in the right panel. On spin glass models, SP gives comparable performances w.r.t. DC (the qualitative behaviour of the right panel in Figure 4.4 is similar also on different graph topologies). On the other hand, on ferromagnetic models it typically overestimates the magnitute of such correlations, givin worse performances if compared to DC. On this regard, an additional set of results is shown in Section 4.3.5.

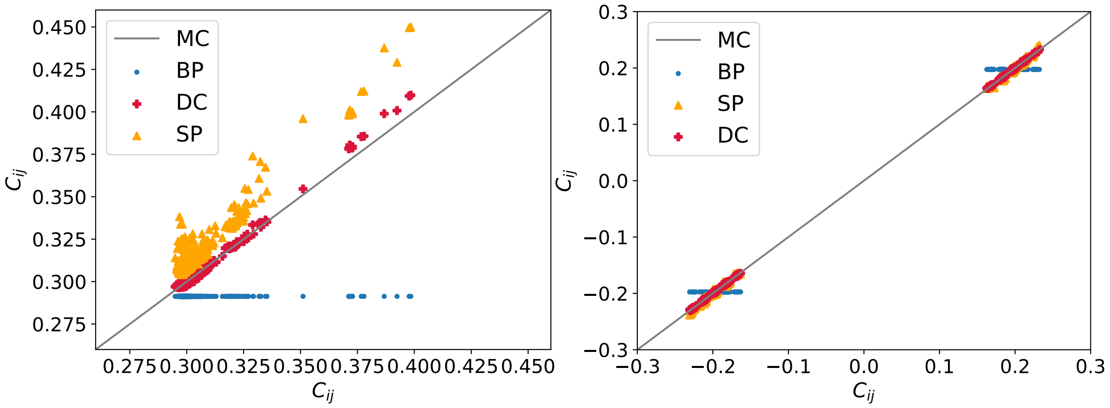

Figure 4.4: Comparison of DC, BP and Susceptibility Propagation (SP) on two single-instances Ising models with null external fields. Left panel: random Regular Graph (RRG) with $N = 300$, constant degree $k = 4$ and $\beta = 0.3$, constant coupling $J_{ij} = 1$ among n.n. nodes. Right panel: 3−dimensional cubic lattice with PBC, $N = 6^3$and $\beta = 0.2$, binary couplings $J_{ij} = \pm 1$ sampled with equal probability. All the figures represent the scatter plot of nearest neighoburs correlations obtained through BP, DC and SP, compared to Monte Carlo estimates.

Effect of the interpolation parameter η

As discussed in the previous chapter (section 3.2.2) the DC closures can be modified by introducing a real parameter $\eta \in [0,1]$ in the Pearson matching equation (Eq. (3.34c)): in particular, setting $\eta = 0$ is equivalent to neglect cavity correlations, so that DC fixed points coincide with Belief Propagation's ones, on any graph topology. In many regimes, e.g. at low temperatures for the Ising model, a full DC closure ($\eta = 1$) does not converge: the reason could be that the loop contributions coming from the cavity make the covariance matrix of the tilted distributions to be ill-defined. A possible way to solve this issue is to choose a value of $\eta < 1$ to help DC in reaching a fixed point. This issue arises especially in the case of ferromagnetic models close to the critical temperature, where the contribution of long-range correlations is larger: for instance, the result in Figure 4.1, panel (c) is obtained for $\eta^* = 0.95$, as for $\eta > \eta^*$ DC did not converge on this particular model. In general, decreasing η makes the correlations induced by the Gaussian cavity to be less dominant with respect to the direct link's contribution between two nearest neighbours spins.
The top panel of Figure 4.5 shows the behaviour of n.n. correlations in a $3d$ ferromagnetic Ising model with constant couplings as a function $\eta \in [0,1]$, evaluated at different temperatures. In particular the left plot displays the error between DC's tilted estimates and Monte Carlo, computed as

$$\Delta_C = \sqrt{\frac{\sum_{\langle i,j \rangle} \left(C_{ij}^{\text{est}} - C_{ij}^{*} \right)^2}{|E|}} \tag{4.24}$$

where C_{ij}^* corresponds to the ground-truth estimate given by Monte Carlo (eventually, it can be computed by evaluating the exact trace over all the configurations, when possible). Conversely, the upper-right panel of Figure 4.5 shows the average tilted correlations vs η, normalized to the value $\eta = 1$. It is evident in both cases how at high temperatures the loop contributions induced by the cavity are unrelevant to estimate correlations, and indeed the corresponding curves are almost constant w.r.t. η; as soon as the temperature decreases, the loop contribution become more and more important, and the error decreases by increasing η. Further notice how the tilted correlations show a high non-linear behaviour w.r.t. η. The lower-left panel shows the scatter plot of n.n. correlations against MC obtained with different values of η: the instance is the same of Figure 4.1 (c) for the Random Regular Graph. The lower-right panel shows again the error over correlations on a spin glass model with binary couplings, on a small $2d$ lattice of size $N = 4^2$: this time the error over correlations is computed w.r.t. to the exact trace, and averaged over 20 instances. In spin glass models, the effective coupling induced by the cavity on an edge (ij) might not have the same sign as the direct link's contribution, depending in general on the particular instance of the coupling matrix.
The error shows a similar behaviour w.r.t. the ferromagnetic case, i.e. decreasing with η. However, at small value of the temperature a minimum appears at $\eta = \eta^*$ and DC stops converging for $\eta > \eta^*$. In principle, one can heuristically choose η as the maximum value at which a DC fixed point can be found.

Comparison with LCBP

We now present a small set of results where DC is compared against Loop Corrected Belief Propagation (LCBP), already introduced in Chapter 2. In particular, we used the code provided in [93] to run LCBP. Figure 4.6, shows the equilibrium magnetizations of two Ising models in the presence of random external fields $h_i \sim \pm h_0$, again plotted against Monte Carlo's estimates. DC turns out to give comparable estimates w.r.t. Loop Corrected Belief Propagation (LCBP) [92] in several cases. We underline that, despite the computational cost per iteration of LCBP on bounded-degree graphs being $O\left(N^2\right)$, the prefactor depends strongly on the degree distribution,

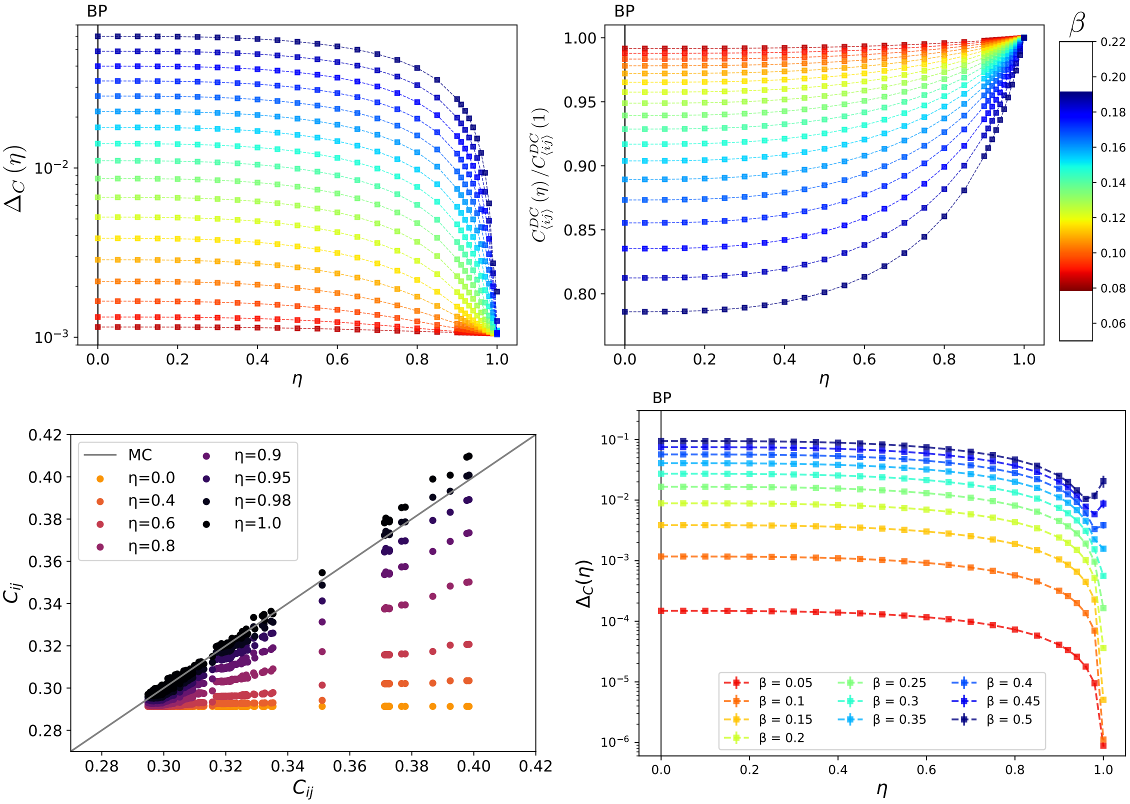

Figure 4.5: Effect of the interpolation parameter η on tilted moments. The top panel refers to a $3d$ lattice Ising model with constant ferromagnetic couplings, $N = 6^3$ and $J_{ij} = 1$. Upper-left panel: error over n.n. correlations with respect to MC estimates; upper-right panel: tilted correlations estimated through DC, normalized over the $\eta = 1$ value. Each curve shows the behaviour vs $\eta \in [0,1]$ ($\eta = 0$ corresponds to BP, as shown by the vertical line), for different values of β in the range [0.05,0.22]; the lower-extreme is close to the critical temperature in the thermodynamic limit. Lower-left panel: scatter plot of n.n. correlations for different values of η, on the same instance of Figure 4.1 (a). Lower-right panel: error between DC's estimates of n.n. correlations w.r.t. the exaustive trace, over a spin glass model with binary couplings $J_{ij} \pm 1$ defined over a square lattice of size $N = 4^2$, averaged over 20 instances.

also the number of iterations required to converge is normally much larger the one required by DC. In both the instances shown in Figure 4.6, LCBP did not to converge at smaller values of the temperature (shown in the caption): this could indicate that below a certain temperature the method is not able anymore to satisfy consistency conditions between the different BP fixed points.
On the other hand, correlations were not accessible through the code provided in [93]. To have a fair comparison, we also tried to estimate 2-points correlations with LCBP, by adding set of "pair" variable nodes x_{ij} for each link (i, j) in the original graph: then, imposing constraints on the probability distribution of x_{ij} of the type $\delta_{x_i x_j, x_{ij}}$ we can compute the 2-points correlation by evaluating $\langle x_{ij}\rangle = \langle x_i x_j\rangle$ at convergence, without modifying the marginal distributions of the starting model. However, the addition of such pair variables (whose number is simply equal to the number of edges in the graph, $|E|$) increases too much the computation time of the algorithm, which never converged in the cases we analyzed.

As discussed in the previous chapter, Loop Corrected Belief Propagation (LCBP) works by computing several BP fixed points (one for each cavity distribution in which one node and all the factors connected to it are removed) and then imposing consistency over single-node beliefs among them. Therefore, for each cavity distribution it computes fixed points by still assuming a tree-factorization, i.e. by neglecting correlations coming from other cycles in the (cavity) graph. So it computes a higher order approximation by relying on lower order ones, on a simplified interaction graph. In this sense, it can be considered as a first-order correction to BP and indeed it improves BP estimates of single-node marginals, as shown in 4.6. On the other hand, DC can be considered as a novel zero-th order approximation in which all 2-points cavity correlations are taken into account, in a single self-consistent set of equations. In this perspective, LCBP and DC correction methods are also in some sense orthogonal, and so it principle possible to design a sort of 'Loop Corrected Density Consistency' (LCDC) approximation in which each cavity distribution is computed by DC, and then single-node marginals would be determined in a self-consistent way.

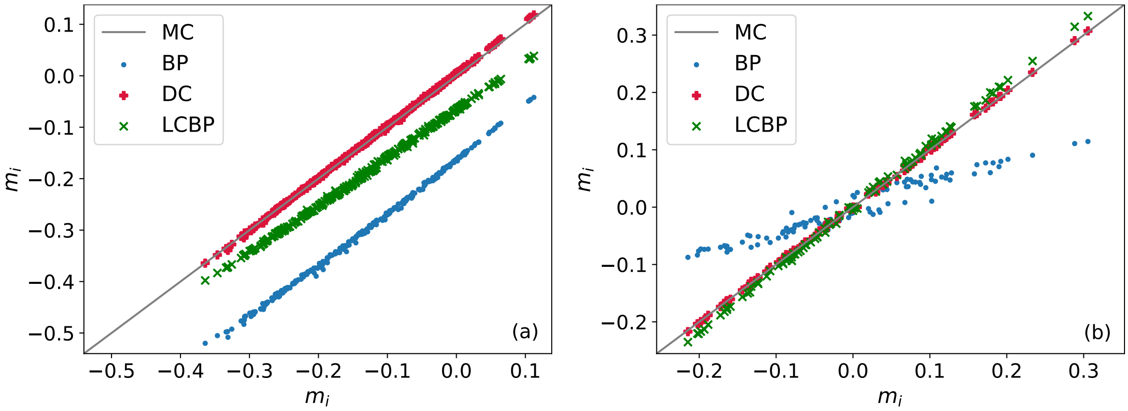

Figure 4.6: Comparison of DC, BP and LCBP on single-instances of disordered systems. Left: Magnetization of Ferromagnetic Ising Model on a Random Regular (RR) Graph with $N = 300$ and constant connectivity $k = 4$, $\beta = 0.35$, constant couplings $J_{ij} = 1$ and random binary fields of $h_i \sim \pm 0.3$. Right: Magnetizations of AntiFerromagnetic Ising Model on a triangular lattice with $N = 100, |E| = 6N$, constant couplings $J_{ij} = -1$, $\beta = 0.52$ and random binary fields $h_i \sim \pm 0.2$.

Chess-like plaquette-DC on hypercubic lattices and comparison with Cluster Variational Method

On regular structures, we discuss a possible way to generalize DC in such a way to take into account larger regions of the graphs explicitly, in a similar spirit to Cluster Variational Method. Let us consider a hypercubic lattice in d dimension with size L on each side: the case $d = 1$ corresponds to a linear chain, $d = 2$ to a square lattice, $d = 3$ to a cubic lattice, and so on. DC can be generalized by grouping together small plaquettes of spins, of size 2^d, into a single factor node (denoted with $\psi_\Box$), in such a way that we allow only for *site-overlaps* between adjacent plaquettes. Figure 4.7 shows this procedure in the case of a two-dimensional square lattice, where only the gray plaquettes are selected, so that the resulting factor graph has a chess-like structure. Such a procedure can be applied at any d-dimensional lattice with periodic boundary conditions, if L is even. In this way, each link in the original lattice appears in exactly 1 plaquette-node $\Box$. To make an equivalence with the notation used in Section (2.3.1), the set of maximal regions $\mathcal{R}_0$ must be chosen in such a way that the set of their intersections ($\mathcal{R}_1$) contains only single-nodes: in particular, with this chess-like covering, $\mathcal{R}_1 = V$.

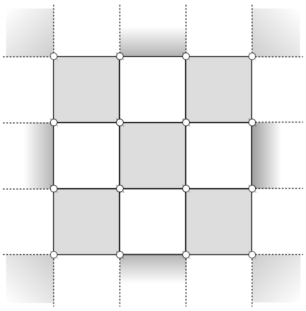

Figure 4.7: Plaquette distributions over a square lattice. The grey plaquettes with only single-site overlaps are the ones encoded by a generalized Density Consistency scheme.

The probability distribution of the Ising model can therefore be rewritten as:

$$p\left(\boldsymbol{x}\right) \propto \prod_{\square} \psi_{\square}\left(\boldsymbol{x}_{\square}\right) \tag{4.25}$$

where the product runs over distinct plaquettes, and each of the compatibility functions $\psi_{\square}$ wil contain all the interactions inside the plaquette. DC can be now applied by replacing each factor $\psi_{\square}$ with a certain multivariate Gaussian density of 2^d variables, denoted with $\phi_{\square}$, and repeating the procedure discussed in the previous chapter. We shall call this approximation *plaquette DC* (pDC). The reason why only site-overlaps are allowed is because DC condition (3.28) is constructed over single-node marginals. On the other hand, by including “all” the plaquettes in the original lattice, there would be intersections between adjacent plaquettes on links, rather than single nodes: in this case, DC cannot be applied as it is and DC condition should be generalized in order to mimic a density consistency over pairwise marginal distributions. This would allow to construct an extension of the Cluster Variational method in which all the interactions inside each plaquette are taken exaclty into account, and the rest is approximated by a Gaussian cavity distribution, encoding all the correlations in the cavity graph where one plaquette is removed. Another possible way to include exactly larger regions of the graph exploits the generalization to multistate variables presented in Section 3.4.1. In this case, the possible strategy would be to re-define the Ising model w.r.t. to *superspin* nodes, each one associated to a plaquette of size 2^d with a number of states $\mathcal{X} = 2^{2^d}$. We leave both issues for future investigations.

We now present some numerical comparisons against the Cluster Variational Method described in Chapter 2. In particular, we run CVM simulations by using the implemenation discussed in [42], specifically designed for $2-$dimensional square lattices where the maximal regions are chosen to be plaquettes of 4 spins. In this case, we run both DC and the plaquette version (pDC) just discussed. Figure (4.8) shows two sets of results on a square lattice with size $L = 10$ and no external fields. In both scenarios, we plotted the error over correlations w.r.t. to MC estimates. The left panel refers to a ferromagnetic model with heterogeneous couplings: DC and pDC show comparable performances with respect to CVM (black curve) in the high-temperature regime, and slight improvements towards the transition temperature. All the three methods significantly improve BP's estimates, whose poor performances are due to the high number of short loops. The right panel shows the behaviour on a spin glass with binary couplings, and DC/pDC results are shown for different values of η (more details in the caption): in this case, CVM seems always to outperform DC (and its plaquette extension). As one could expect, pDC always (altough slightly) improves the pairwise implementation in all the simulations we have run, and it typically has less convergence issues. We can conclude that on spin glass models on structured graphs the absence of long-range correlations makes the loop contributions induced by the Gaussian cavity less relevant to correctly estimate the n.n. correlations, and CVM has to be preferred on such low-dimensional systems.

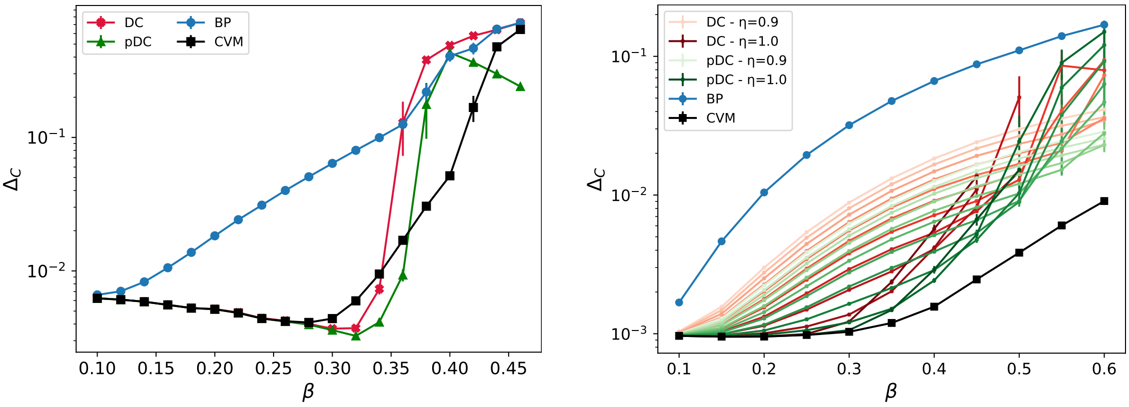

Figure 4.8: Comparison between BP, CVM, DC and pDC on Ising-like models defined over a square lattice of size $L = 10$ (with PBC). The two plots show the error over n.n. correlations w.r.t. MC estimates, at different values of β. In the left panel interactions are ferromagnetic, sampled in the range $\left[\frac{1}{2}, \frac{3}{2}\right]$, and the MC is carried out using the Wolf algorithm [141]. Both DC (red) and pDC (green) are run using $\eta = 0.99$. The right panel refers to a spin glass model where $J_{ij} \pm 1$ with equal probability: DC (red curves) and pDC (green curves) are shown for different values of $\eta \in [0.9{,}1.0]$, the color intensity being increasing with η. In both panels, results are averaged over 20 instances, by varying the random seed used to sample the coupling matrix. Each DC and pDC curve is plotted for all the temperatures where convergence is reached on at least half of the instances analyzed.

Breakdown scenario

In general, we expect that DC will give reasonably good estimates of marginals as long as the correlations in the true cavity distribution can be well described by the ones of a Gaussian density. To be more precise, we analyzed some non-Gaussianity measures of the cavity distribution. For simplicity, we considered a homogeneous Ising model on a square lattice of size 5×5 with PBC, in zero-field: we computed exact marginals on the cavity model in which one link (i, j) was removed (in such a homogeneous model, the choice of the particular link (ij) is unrelevant since it is translational invariant). Let us denote with $p^{\setminus(ij)}$ the true probability distribution of the cavity model, and with $g^{\setminus(ij)}\left(\boldsymbol{x}\right)$ the Gaussian distribution whose first two moments match the magnetization and correlation matrix of the trace just computed. Therefore, by construction, $\langle \boldsymbol{x} \rangle_{p^{\setminus(ij)}} = \langle \boldsymbol{x} \rangle_{g^{\setminus(ij)}}$ and $\langle \boldsymbol{x}\boldsymbol{x}^t \rangle_{p^{\setminus(ij)}} = \langle \boldsymbol{x}\boldsymbol{x}^t \rangle_{g^{\setminus(ij)}}$. The coupling matrix of a Gaussian can be computed from its covariance matrix as $\tilde{\boldsymbol{J}} = -\left(\boldsymbol{\Sigma}^{\setminus(ij)}\right)^{-1}/\beta$, where $\boldsymbol{\Sigma}^{\setminus(ij)} = \langle \boldsymbol{x}\boldsymbol{x}^t \rangle_{p^{\setminus(ij)}} - \langle \boldsymbol{x} \rangle_{p^{\setminus(ij)}} \langle \boldsymbol{x}^t \rangle_{p^{\setminus(ij)}}$. If the true cavity distribution behaves as a Gaussian, we expect the inferred coupling matrix to have the same structure of the original graph. Therefore, we can define a measure of non-Gaussian behaviour as the ratio between the sum of inferred couplings which do not belong to the original graph over the sum of true couplings:

$$w\left(\beta\right) = \frac{\sum_{(k,l)\notin E} \tilde{J}_{kl}}{\sum_{(k,l)\in E} J_{kl}} \tag{4.26}$$

Another possible measure of non-Gaussianity is the absolute difference between the 4th-order moment $x_i x_j x_k x_l$ computed under the distribution $g^{\setminus(ij)}$ and $p^{\setminus(ij)}$, respectively:

$$r_4(\beta) = \left| \langle x_i x_j x_k x_l \rangle_{g^{\setminus(ij)}} - \langle x_i x_j x_k x_l \rangle_{p^{\setminus(ij)}} \right| \tag{4.27}$$

where k, l are chosen in such a way that the tuple (i, j, k, l) forms a plaquette in the original lattice. We computed these two quantities in a (inverse) temperature range $\beta \in [0.1,0.45]$, together with the error over the 2-point correlation referred to the link (i, j), the latter being computed between the trace and DC's estimate on the full model:

$$\varepsilon(\beta) = \frac{\left| \boldsymbol{\Sigma}_{ij}^{DC} - \boldsymbol{\Sigma}_{ij} \right|}{\boldsymbol{\Sigma}_{ij}} \tag{4.28}$$

Figure 4.9 shows the behaviour of these three quantities vs β. As expected, at large temperatures the cavity distribution behaves well enough as a Gaussian distribution, and DC's estimate well predicts the true correlations. When the cavity distribution deviates from a Gaussian - in particular, when the fourth-order moment deviates from 0 - DC's prediction starts to deviate from the true one, up to a certain point at which the method stops to converge. A physical interpretation of the lack of convergence on square lattices in the thermodynamic limit (or eventually, other ferromagnetic models with low average connectivity, i.e. $\langle d_i \rangle < 4$) will be discussed in Sec. 4.3 for the high-dimensional limit.

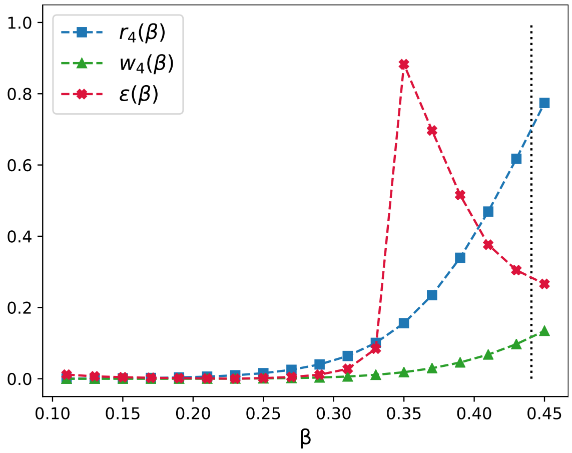

Figure 4.9: Plot of the two non-Gaussianity measures r_4 and w as a function of β, together with the normalized error over the two point correlation on the link (i, j). The model is a square lattice Ising with size $N = 5^2$ in zero field. The dashed black line denotes the critical temperature in the thermodynamic limit [102].

4.2 k-SAT

Satisfiability is a paradigmatic class of constraint satisfaction problems, formulated by a set of logical inputs whose state is asked to satisfy a number of constraints [95, 99]. An istance of a satisfiability problem is defined by a set of N Boolean variables $\{s_i \in \{0,1\}\}_{i=1:N}$ and a set of M constraints, also called *clauses*. Each clause is a logical `OR` between k *literals*, each of them corresponding to one variable s_i or its negation (represented by the `NOT` logical operator and denoted with $\bar{s}_i = 1 - s_i$). By virtue of the `OR` operator, each clause is satisfied by all the configurations where at least one literal is true, so that there is only one configuration where the clause is not satisfied, i.e. the one where all the literals are false. We shall define z_i as the literal of node i in a certain clause, so that $z_i = s_i$ or $z_i = \bar{s}_i$. In the k-SAT problem, each clause involves exactly k variables, and it can be written as $C_a\left(s_{i_1}, \dots s_{i_k}\right) = z_{i_1}\left(s_{i_1}\right) \vee \dots \vee z_{i_k}\left(s_{i_k}\right)$ where $\vee$ denotes the `OR` operator. Since in the SAT problem all the clauses need to be satisfied simultaneously, a $k-$SAT formula is given by the logical `AND` ($\wedge$) operation over the full set of clauses:

$$\mathcal{C}\left(s_1, \dots, s_N\right) = \bigwedge_{a=1}^{M} C_a\left(\boldsymbol{s}_{\partial a}\right) \tag{4.29}$$

where $\boldsymbol{s}_{\partial a} = \{s_{i_\alpha}, \alpha = 1 \dots k\}$. In particular, $\mathcal{C} = \mathtt{T}$ (true) if the configuration $\boldsymbol{s} = (s_1, \dots, s_N)$ satisfies all the clauses, and $\mathcal{C} = \mathtt{F}$ (false) if at least one of them is false. The k-SAT problem corresponds to find one (or more) configuration(s) where the $\mathcal{C}$ is true.
The above notation allows to easily express any $k-$SAT instance by using a factor graph representation. In the following, we will work with binary spins rather than boolean variables, so to have a more similar notation w.r.t. the statistical physics language. We denote with V the set of nodes, each of them corresponding to a binary spin $\sigma_i \in \{-1,1\}$. The factor graph representation of a k-SAT formula is defined by $G = (V, E, F)$ where each factor node $a \in F$ is associated to one clause C_a. Notice that all factor nodes have an equal degree $|\partial a| = k$. In terms of spin variables, each literal z_i can be expressed by the product $\xi_i \sigma_i$, where $\xi_i = 1$ if $z_i = s_i$ and $\xi_i = -1$ if the variable is negated ($z_i = \bar{s}_i$). To each factor node a, its corresponding check function ψ_a is just an indicator function over clause a, provided the mapping between boolean variables and spin variables. With this parametrization, the (uniform) probability distribution of a random k-SAT formula can be written using the general expression (3.1) as:

$$p\left(\boldsymbol{\sigma}\right) = \frac{1}{Z} \prod_{a=1}^{M} \psi_a\left(\boldsymbol{\sigma}_{\partial a}\right) \tag{4.30}$$

$$\psi_a\left(\boldsymbol{\sigma}_{\partial a}\right) = \mathbb{I}\left[\sum_{\alpha=1}^{k} \xi_{i_\alpha} \sigma_{i_\alpha} > -k\right] = 1 - \mathbb{I}\left[\sum_{\alpha=1}^{k} \xi_{i_\alpha} \sigma_{i_\alpha} = -k\right] \tag{4.31}$$

where $\mathbb{I}$ is the identity function and $\boldsymbol{\sigma}_{\partial a} = \{\sigma_{i_\alpha}, i_\alpha \in \partial a, \alpha = 1, \dots, k\}$. Therefore, each clause defines a hard constraint where $\psi_a = 0$ only on one configuration of its neighbours, where all the literals $\xi_{i_\alpha} \sigma_{i_\alpha} = -1$. The partition function Z counts the number of solutions to the SAT formula. Each $k-$SAT instance can be drawn from an ensemble $\mathtt{SAT}_k\left(N, M\right)$, by selecting M clauses from the all $\binom{N}{k} 2^k$ possible choices. In the thermodynamic limit where both $N, M \to \infty$ with a finite ratio $\alpha = M/N$, a phase transition separates a SAT regime where the probability to have a configuration satisfying all the constraints is 1, from an UNSAT phase where such probability tends to 0. The transition occurs at a finite value $\alpha_c\left(k\right)$ (for instance, $\alpha_c\left(k=3\right) \approx 4.27$, $\alpha_c\left(k=4\right) \approx 9.9$ as estimated through the 1RSB cavity method [99]). In addition to the SAT-UNSAT threshold there exist several other critical points, separating phases in which standard message-passsing algorithms do not have locally stable fixed points, or condensation phenomena arise (and the configuration space breaks down into disconnected components). In this regime finding solutions is much harder and one has to employ message-passing algorithms defined on the replica-symmetric broken space (such as Survey Propagation [18, 98]).
In the present section, we just want to present a small set of preliminar results by comparing DC performanges against BP, in small-size k-SAT instances where the exact trace over configurations is possible. Results are shown in Figure 4.10 for $N = 20$, $k = 3,4$, averaged over 40 instances. We compare BP and DC's estimates of the first and second-order moments, by computing normalized errors w.r.t. to the exaustive trace. DC is run for different values of $\eta \in (0.85,0.99)$, and each curve is plotted up to the value of α at which convergence is reached on more than half the instances considered. Despite DC is able to better estimate the moments at small values of α,

its performances get worse than BP at higher values of α, especially on the first moments. At higher values of α DC seems not to converge on most of the instances considered. A possible explanation is that, in such regimes with small N, it might happen that two nodes appear in more than one clause, so that DC is not able to satisfy at the same time constraints induced by clauses overlapping on more than 1 node. In principle, it should be possible to generalize DC to take into account consistency on two body marginals as already discussed in the previous section for the plaquette extension. Furthermore, it would be interesting to see if DC could be generalized in such a way to take into account the disconnected structure of the configuration space arising close to the SAT-UNSAT transition, similarly to the 1RSB cavity method. We leave this issue for future investigations.

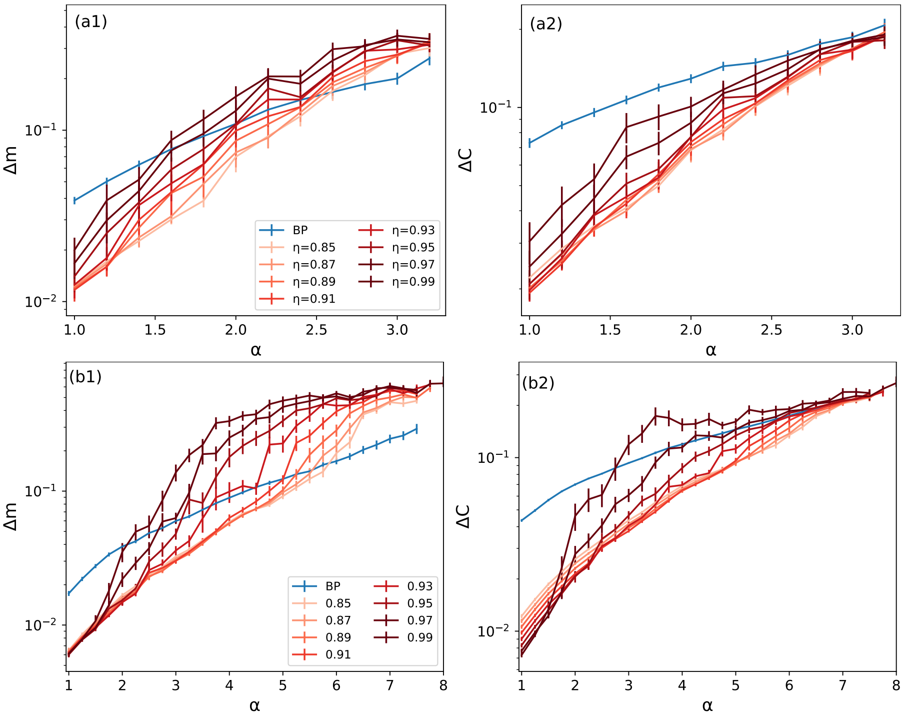

Figure 4.10: Comparison between DC and BP on random $k-$SAT models with $N = 20$. Top panel (a) refers to $k = 3$, bottom panel (b) to $k = 4$. Each panel shows the normalized over magnetizations (left) and over correlations (right) w.r.t. the exact trace as a function of $\alpha = M/N$. DC runs are shown for different values of η, averaged over 40 instances.

4.3 Thermodynamic limit for the Ising ferromagnet on hypercubic lattices

In this section, we exploit Density Consistency to derive a semi-analytic solution for the homogeneous, ferromagnetic Ising model, defined on a hypercubic lattice in d dimensions with periodic (toroidal) boundary conditions (PBC), in the thermodynamic limit. The key feature of such a model, i.e. its traslational invariance, will be exploited by DC in order to provide a finite set of fixed point equations to be solved, independently on the system size. We denote with L the number of spins on each side of the lattice, so that the total number of variables is $N = L^d$. The Hamiltonian is parametrized by a constant coupling J between nearest neighbours spins and a constant field h acting on each variable. For convenience, we rewrite the Boltzmann law (4.1)of this model in a matrix-vector notation:

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left[\frac{\beta J}{2} \boldsymbol{\sigma}^t \mathcal{A}^{(d)} \boldsymbol{\sigma} + \beta h \boldsymbol{\sigma}^t \mathbf{1}\right] \tag{4.32}$$

where $\mathcal{A}^{(d)}$ denotes the adjacency matrix of the $d-$dimensional lattice, and $\mathbf{1}$is a $N-$dimensional vector where each component is equal to 1. The thermodynamic limit is computed by taking $L \to \infty$ ad fixed d. We rewrite again 4.32 as a factorized distribution of continuous variables over n.n. pairs, in the same way as in 4.2:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{\langle i,j \rangle} \psi_{ij} \prod_i \Delta_i, \qquad \psi_{ij}(x_i, x_j) = \exp\left[\beta J x_i x_j + \frac{\beta h}{2d}(x_i + x_j)\right] \tag{4.33}$$

where on each factor ψ_{ij} the field term acting on each spin is divided by the number of its neighbours (equal to $2d$ for a $d-$dimensional lattice with PBC). Since the model defined by 4.33 is translational invariant, the expectation values of a subset of variables will depend only on their relative positions in the lattice. As a consequence, magnetizations will be equal for all spins, and nearest neighbours correlations will be equal among all spins that are connected by a direct link in the lattice, namely:

$$\langle x_i \rangle_p \hat{=} m \quad \forall i \tag{4.34}$$

$$\langle x_i x_j \rangle_p \hat{=} \chi \quad \forall (i,j) \in E \tag{4.35}$$

We now apply the machinery of Density Consistency: in this case, the translational invariance of the model is exploited to construct a family of equivalent approximating Gaussian distribution ϕ_{ij}, all of them being parametrized by the same set of quantities as follows:

$$\phi_{ij}(x_i, x_j) = \exp\left[-\frac{1}{2}\Gamma_0\left(x_i^2 + x_j^2\right) - \Gamma_1 x_i x_j + \gamma(x_i + x_j)\right] \quad \forall (i,j) \in E \tag{4.36}$$

The above expression follows directly from (4.5): here Γ_0 is the diagonal self-coupling , $\Gamma_1$is the approximate pairwise interaction and γ is the Gaussian field. Notice that Eq. (4.36) is invariant under the exchange $x_i \leftrightarrow x_j$. In this way, there will be only 3 parameters to be determined to estimate the equilbrium behaviour of the model, independently on the system size. Taking the product of all the approximating factors we construct the full Gaussian measure as in (3.7), shown below in a matrix-vector notation:

$$q(\boldsymbol{x}) \propto \prod_{\langle i,j \rangle} \phi_{ij}(x_i, x_j) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^t \mathcal{K}^{(d)} \boldsymbol{x} + 2d\gamma \boldsymbol{\sigma}^t \mathbf{1}\right] \tag{4.37}$$

where

$$\mathcal{K}^{(d)} = 2d\Gamma_0 \mathbb{I}_{L^d} + \Gamma_1 \mathcal{A}^{(d)} \tag{4.38}$$

and $\mathbb{I}_N$ denotes the identity matrix of size N. In this case, the distribution (4.37) is expressed by a quadratic form in its exponent, rather than w.r.t. first and second moments. The matrix $\mathcal{K}^{(d)}$ is a combination of a diagonal term (coming from the self-coupling contributions for each neighbour) and the adjacenty matrix of the lattice $\mathcal{A}^{(d)}$, multiplied by the approximate coupling Γ_1. To compute the moments of (4.37), one needs to invert the matrix $\mathcal{K}^{(d)}$, whose size will eventually become infinity as we take the thermodynamic limit. However, the homogeneous structure of the graph allows for an analytic evaluation of the adjacency matrix's eigenspectrum, so that the inverse matrix elements of $\mathcal{K}^{(d)}$ can be easily computed even at the thermodynamic limit. Notice that, in order to apply DC scheme, we are only interested in two types of elements of the inverse of (4.38), namely the diagonal entries and the terms corresponding to nearest neighbours spins, denoted respectively with Σ_0 and Σ_1:

$$\Sigma_0 \hat{=} \left[\mathcal{K}^{(d)}\right]^{-1}_{ii} \quad \forall i \qquad \Sigma_1 \hat{=} \left[\mathcal{K}^{(d)}\right]^{-1}_{ij} \quad \forall\left(i,j\right) \in E \tag{4.39}$$

A detailed computation of (4.39) is discussed in Appendix (B), we report here the final result in the thermodynamic limit $L \to \infty$:

$$\Sigma_0 = \frac{1}{\Gamma_0} R_d\left(r\right) \tag{4.40}$$

$$\Sigma_1 = \frac{1}{r\Gamma_0}\left[\frac{1}{2d} - R_d\left(r\right)\right] \tag{4.41}$$

with

$$r = \frac{\Gamma_1}{\Gamma_0} \qquad R_d\left(r\right) = \frac{1}{2}\int_0^\infty dt\left[e^{-t}\mathcal{I}_0\left(rt\right)\right]^d \tag{4.42}$$

The parameter r plays the role of a normalized effective coupling: in particular, if $|r| < 1$ the matrix $\mathcal{K}^{(d)}$ is diagonally dominant and it can be inverted. The function $R_d\left(r\right)$ in (4.42) has a physical interpretation in terms of the return probability of a random walk in the infinite lattice: more specifically, there is a direct connection between $R_d\left(r\right)$ to the Lattice Green Functions (LGFs) [61], i.e. the probability generating function of a random walk on the lattice. By a simple inspection of $R_d\left(r\right)$, we find that the quantity $1 - \left[2dR_d\left(-1\right)\right]^{-1}$ is nothing but the return probability to the origin of a (symmetric) a random walk on the infinite d-dimensional lattice. Figure 4.11 displays the behaviour of the ratio Σ_1/Σ_0 as a function of the normalized coupling r. Σ_1 is and odd function of r (on the contrary, Σ_0 is even and therefore their ratio is still an odd function): since in the ferromagnetic model correlations are positive, i.e. two nearest neighoburs spin prefer to have the same sign, we are interested in the regime $r < 0$. The regime of positive values of r corresponds instead to the AntiFerromagnetic Ising model, where $J_{ij} = J < 0$ for all n.n. spins.

Duality Ferro-AntiFerro

It is important to remark that in the case of hypercubic lattices there exist an equivalence between the Ferromagnetic and the Antiferromagnetic model: indeed, the hypercubic lattice is a bipartite graph and it is characterized by the absence of odd-length loops[2] [15]; as a consequence, the antiferromagnetic model shows the same critical behavior than its ferromagnetic counterpart (in the absence of external fields) and it is always possible to map one model onto the other by a simple transformation, i.e by mapping half of the spins (in one of the two disjoint sets of the bipartite graph). The latter reasoning explains why Σ_1 is simply an odd function of r. There are other regular structures where such equivalence does not hold: for instance, on a triangular lattice the shortest loops corresponding to triangular plaquettes have an odd length, so that a geometric frustration arises from the fact that the product of nearest neighbours couplings over the (triangular) plaquettes is negative [82]. In this case, the behaviour of the anti-ferromagnetic

[2] The term “bipartite” denotes a generic graph whose vertex set can be divided in two disjoints set A and B, in a such way that all the edges (i,j) always connect two spins from the two sets A, B. Specifically, the hypercubic lattice in d dimension with PBC is bi-partite only if L is even (on the other hand, it is always bi-partite if the boundary conditions are open). Even for odd values of L, the contribution of odd-length cycles becomes negligible in the thermodynamic limit and therefore the spectrum tends to be symmetric for $L \to \infty$. Without loss of generality, one could define $L = 2L'$ so to have always a lattice with an even number of spins on each side.

model is notably different from the ferromagnetic counterpart [135]. The analytic DC scheme proposed in this Section can in principle be applied to any lattice with a known eigenspectrum in the thermodynamic limit, like the already cited triangular lattice or other regular structures (fcc,bcc [69, 128]): we leave this investigation to future works.

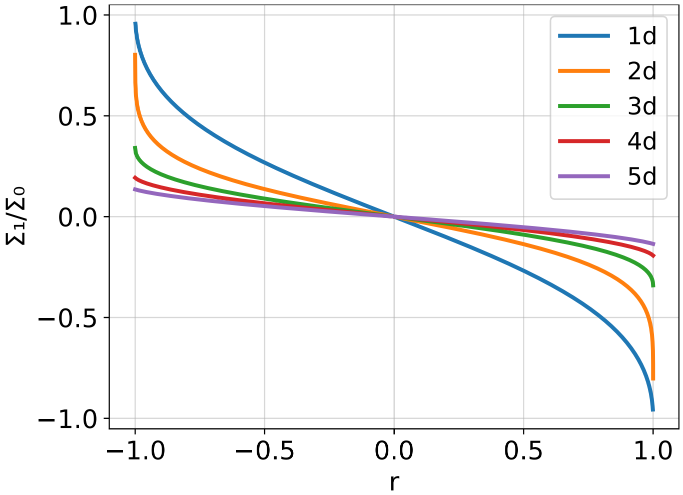

Figure 4.11: Plot of Σ_1/Σ_0 vs r for different values of d.

The Gaussian mean vector $\boldsymbol{\mu}$ of 4.37 can be easily computed by solving the linear system $\mathcal{K}^{(d)}\boldsymbol{\mu} = 2d\gamma\mathbf{1}$, that has the following trivial solution:

$$\mu = \frac{\gamma}{\Gamma_0 + \Gamma_1} = \frac{\gamma}{\Gamma_0\,(1+r)} \tag{4.43}$$

Following the derivation of Density Consistency, we define the set of tilted distributions and Gaussian cavities, one for each link (ij) in the lattice. Their expression follow directly from the ones described in Sec 4.1.1. In particular, we are interested in the marginal tilted distribution over (ij), that can be expressed as:

$$q^{(ij)}\,(x_i, x_j) \propto g^{\setminus(ij)}\,(x_i, x_j)\,\psi_{ij}\,(x_i, x_j)\,\Delta_i\Delta_j. \tag{4.44}$$

The marginal cavity density $g^{\setminus(ij)}$ is parametrized as follows:

$$g^{\setminus(ij)}\,(x_i, x_j) \propto \exp\left[-\frac{1}{2}S_0\left(x_i^2 + x_j^2\right) - S_1 x_i x_j + w\,(x_i + x_j)\right] \tag{4.45}$$

and the parameters w and S_1 denote respectively the cavity field and coupling, shown below[3]:

$$w = \mu\left(\frac{1}{\Sigma_0 + \Sigma_1} - (\Gamma_0 + \Gamma_1)\right) \tag{4.46}$$

$$S_1 = -\frac{\Sigma_1}{\Sigma_0^2 - \Sigma_1^2} - \Gamma_1 \tag{4.47}$$

[3] the diagonal term S_0 does not enter in the computation of tilted moments and it is not shown for simplicity.

The marginal tilted moments can be analytically computed and their expression has the same structure of 4.14:

$$m = \langle x_i \rangle_{q^{(ij)}} = \tanh\left[a + \text{atanh}\left(\tanh b \tanh a\right)\right] \tag{4.48}$$

$$\chi = \langle x_i x_j \rangle_{q^{(ij)}} = \tanh\left[b + \text{atanh}\left(\tanh^2 a\right)\right] \tag{4.49}$$

where

$$a = \frac{\beta h}{2d} + w, \qquad b = \beta J - S_1 \tag{4.50}$$

and w, S are respectively the field and the coupling coming from the marginal cavity distribution, given by (4.46)-(4.47) respectively. We recall now the expression of the DC closure equations 3.34 in this simplified setup (the interpolation parameter appearing in Eq. (3.34c) is set to 1 from now on):

$$\begin{aligned} m &= \frac{\gamma}{\Gamma_0 + \Gamma_1} \\ \Sigma_0 &= \frac{m}{\text{atanh} m} \\ \Sigma_1 &= \frac{\chi - m^2}{1 - m^2}\Sigma_0 \end{aligned}$$

4.3.1 Simplified DC equations

At this point, the system of 3 equations 4.51 can be iteratively solved w.r.t. the 3 unknowns $\gamma, \Gamma_0, \Gamma_1$ at a fixed inverse temperature β. In this way, the fixed point equations share the same structure of the ones discussed in Chapter (3). However, for reasons that will be clear in the following, we rewrite the fixed points equations and simplify the original system (4.51) in order to get a self-consistent equation for the magnetization $m = M\left(m\left(r\right), r\right)$ at a certain inverse temperature $\beta = B\left(m\left(r\right), r\right)$. By eliminating the variable γ and using the above definitions of $\Sigma_0, \Sigma_1, w, S, m, \chi$ together with DC closure equations, after some algebra we get:

$$\begin{aligned} \beta = B\left(m(r), r\right) = \text{atanh}\left[k_d\left(r\right)\left(1 - m^2\right) + m^2\right] + \\ - g_d\left(r\right)\frac{\text{atanh} m}{m} - \text{atanh}\left[\tanh^2\left(f_d\left(r\right)\text{atanh} m + \frac{\beta h}{2d}\right)\right] \end{aligned} \tag{4.52}$$

$$\begin{aligned} m = M\left(m\left(r\right), r\right) = \tanh\Bigg[f_d\left(r\right)\text{atanh} m + \frac{\beta h}{2d} + \\ + \text{atanh}\left(\tanh\left(\beta J + g_d\left(r\right)\frac{\text{atanh} m}{m}\right)\tanh\left(f_d\left(r\right)\text{atanh} m + \frac{\beta h}{2d}\right)\right)\Bigg] \end{aligned} \tag{4.53}$$

where we defined for simplicity the following functions:

$$k_d\left(r\right) = \frac{1 - 2dR_d\left(r\right)}{2drR_d\left(r\right)} \tag{4.54}$$

$$g_d\left(r\right) = \frac{k_d\left(r\right)}{1 - k_d^2\left(r\right)} + rR_d\left(r\right) \tag{4.55}$$

$$f_d\left(r\right) = \frac{1}{1 + k_d\left(r\right)} - \left(r + 1\right)R_d\left(r\right) \tag{4.56}$$

Such equations can be solved at fixed r in the variables β, m, as opposed to the previous fixed point system (4.51) where $r = \Gamma_1/\Gamma_0$ is found at fixed β. In particular, for $h = 0$ the system reduces to a single fixed point equation for $m = M\left(m\left(r\right), r\right)$ while β is fixed by (4.52).

4.3.2 Paramagnetic phase and critical temperature β_p

The paramagnetic solution can be found by eliminating the external field h and setting $m = 0$ in (4.52)(4.53). As a consequence, Eq. (4.53) becomes an identity and we get the following expression for $\beta(r)$ (from now on, we set $J = 1$ without loss of generality):

$$\beta_d(r) = \text{atanh}\left(\frac{1}{r}\left[\frac{1}{2dR_d(r)} - 1\right]\right) - g_d(r) \tag{4.57}$$

The function $\beta_d(r)$ is plotted in Figure 4.12 for different values of d. The right plot corresponds to value of $d \geq 3$: in this regime, the function $\beta_d(r)$ is limited and monotonically in $r \in [-1,1]$. Therefore, under the DC approximation there exists a maximum temperature at which the paramagnetic phase exists, and for $d \geq 3$ it corresponds to the point $r = -1$. On the contrary, in the Bethe Approximation the paramagnetic solution $m = 0$ always exists, although it becomes unstable for $\beta > \beta_{BP}$. We can now define the DC critical temperature β_p as the limit for $r \to -1$ of Eq.(4.52):

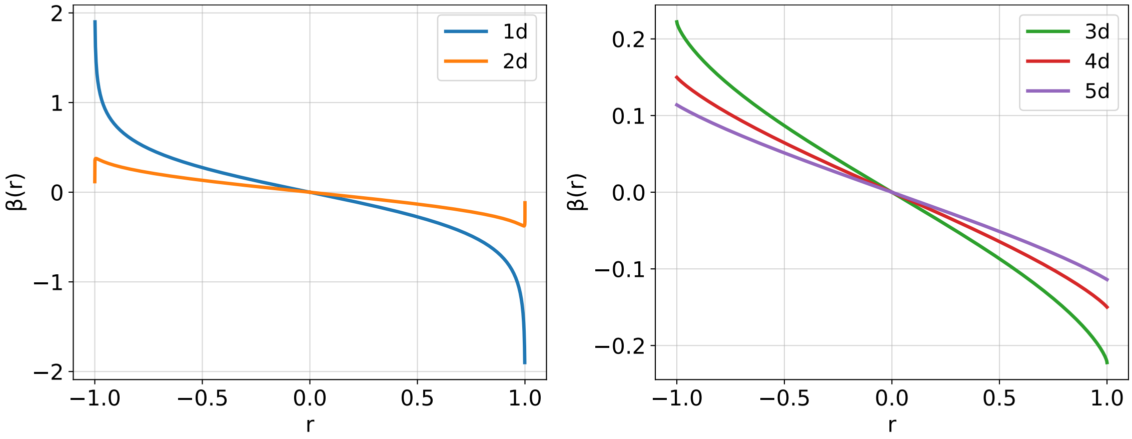

Figure 4.12: Plot of $\beta(r)$ for different values of the lattice dimension d. In particular, we separate the behaviour for $d \leq 2$ (left plot) and for $d > 2$ (right plot)

$$\beta_p = \text{atanh}\left(1 - \frac{1}{z}\right) - z\left(\frac{z-1}{2z-1}\right) + \frac{z}{2d} \tag{4.58}$$

where $z = 2dR_d(-1)$, and we dropped the d-dependency for simplicity. The values of β_p are shown in Table 4.1, where we compared the DC estimate to the best known values in the literature (denoted with β_c) and to other approximation methods in statistical physics: the Mean-Field (MF) solution, the Bethe-Peierls Approximation (BP), plaquette Cluster Variational method (PCVM, [43]), Loop Corrected Bethe (LCB, [91]). For all these methods there exist closed-form expressions for the critical temperature, depending on the lattice dimension d: in particular, the MF and Bethe critical temperature have alredy been discussed in Chapter 2. In particular, we found that for $d \geq 3$ the DC result obtained by using (4.58) gives the closest estimate to the best known value β_c. In addition, the paramagnetic fixed point turns out to be stable in the whole interval $\beta \in (0, \beta_p)$ for $d \geq 3$.

4.3.3 Ferromagnetic phase

The ferromagnetic phase can be investigated by solving the system of Eqs. (4.52)-(4.53) at fixed r. The resulting behaviour of the order parameter $m(\beta)$ is shown in Figure 4.13 for $d = 3$.

d	β_{MF}	β_{BP}	β_{PCVM}	β_{LCB}	β_m	β_p	β_c
2	0.25	0.34657	**0.412258**	-	0.388448	0.37693	0.440687[102]
3	0.16667	0.20273	0.216932	0.238520	0.218908	**0.222223**	0.221654(6)[9]
4	0.125	0.14384	0.148033	0.151650	0.149835	**0.149862**	0.14966(3)[54]
5	0.1	0.11157	0.113362	0.114356	**0.113946**	**0.113946**	0.11388(3)[107]
6	0.083333	0.09116	0.092088	0.092446	**0.092304**	**0.092304**	0.0922530[51]

Table 4.1: Values of the Ferromagnetic Ising Model's critical temperature obtained with different approximation schemes, for different lattice dimension d (shown in the first column). The values of β_{MF}, β_{BP}, β_{PCVM} and β_{LCB} respectively refer to the Mean-Field, Bethe-Peierls, Plaquette Cluster Variational Method (PCVM, [43],) and Loop Corrected Bethe (LCB,[91]) approximations. The values of β_p are computed by using Eq. (4.58). β_m is the minimum value at which a magnetized DC solution exists and it is stable. β_c indicates the currently best known approximation up to numerical accuracy (for $d = 6$ we used the series expansion provided in [51]). Results in bold indicate the closest value to β_c.

Surprisingly, there exist a temperature interval $\beta_m < \beta < \beta_p$ where two magnetized solution exist, together with the paramagnetic solution $m = 0$. However, the lower part of the magnetized phase turns out to be unstable, and this is the reason why we implemented the simplified DC equations discussed before: indeed, the unstable ferromagnetic branch cannot be found by iteratively solving the original system of DC closure equations 4.51. The quantity β_m can be considered as another critical point of DC approximation, and its value can be estimated by computing the minimum of the function $B\left(m\left(r\right),r\right)$ with respect to r; alternatively, one can compute the stability of the ferromagnetic fixed point $m = M\left(m\left(r\right),r\right)$. Both strategies are discussed in details in Appendix C. We report the values of β_m Table 4.1: notice that the difference between the two critical points decreases for large values of d (in particular, for $d \geq 5$ the values coincide up to the numerical precision shown in Table 4.1). The presence of an unstable ferromagnetic phase makes impossible to compute critical exponent within DC approximation; moreover, in the interval $\beta_m < \beta < \beta_p$ DC approximation has both magnetized and a paramagnetic stable solutions, suggesting a phase coexistence that should be absent in the real model [79].

In the right plot of Figure 4.13 we compared the DC estimate with other approximations (again, for $d = 3$, even if the same qualitative behaviour holds for larger dimensions): the Bethe Approximation, univariate Gaussian EP (discussed in Chapter 2) and Density Consistency obtained by using the moment matching closure discussed in 3.2.2. A ferromagnetic unstable branch is found also within univariate EP approximation: this is a signal that the instability might be due to the Gaussian ansatz for the cavity distribution.

4.3.4 $d < 3$

For $d < 3$ the DC solution is qualitatively different w.r.t. the case $d \geq 3$ discussed so far, because the function $R_d\left(r\right)$ is not bounded. The reason can be understood by re-calling the connection with Lattice Green Functions: indeed, for $d < 3$ the random walk defined on the lattice is *recurrent*, i.e the probability of return to the origin is 1, which in turn implies that the quantity $R_d\left(-1\right)$ diverges. In particular, for $d = 1$ the paramagnetic phase described by the function $\beta_1\left(r\right)$ in Eq. (4.57) goes to $+\infty$ at $r \to -1$ (as shown in Figure 4.12. This means that the paramagnetic phase always exist in $1d$ and no phase transition occurs at finite temperature, as it happens on the exact solution [66]. In the two-dimensional case the function $\beta_2\left(r\right)$ defined by Eq. 4.57 has a maximum at a certain $r_p > -1$ and then it diverges to $-\infty$, which has no clear physical meaning; however, we can still define the critical temperature β_p as the maximum value at which the paramagnetic solution exists (and it is stable), which corresponds to the point

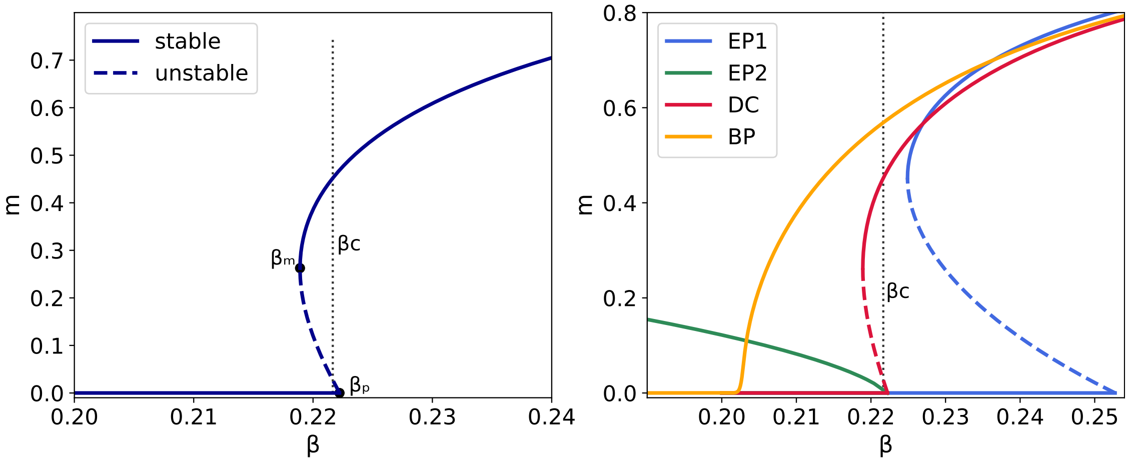

Figure 4.13: Plot of the order parameter m as a function of the inverse temperature β in the 3-dimensional case (in zero field $h = 0$). The left plot shows DC solution: the dashed line represents the unstable ferromagnetic phase. The right plot shows the DC phase diagram compared to the Bethe Approximation, univariate EP (EP1, [90, 103]) discussed in Chapter (2), and DC solution obtained by using the EP closure (EP2) discussed in Sec. 3.2.2

$(r_p, \beta_p) = (-0.994843,0.37693)$. On the other hand, the ferromagnetic solution turns out to be stable for $r_m < r < 0$ with $r_m = -0.99405$, corresponding to $\beta_m = 0.388448$. Since $\beta_p < \beta_m$, there exists a temperature interval $\beta_p < \beta < \beta_m$ in which no stable DC solution can be found. We can conclude that DC is not suited to analyze this kind of low-dimensional models, and one has to rely to other methods like the already cited Cluster Variational method [43].

4.3.5 Finite size corrections

We conclude the discussion on finite dimension d by showing some results about finite size corrections w.r.t. the thermodynamic limit. Indeed, as described in Appendix B, the adjacency matrix of the lattice can be diagonalized exactly even for a finite size L, so that the system of equations (4.51) can be iteratively solved as previously discussed. Figure 4.14 shows the behaviour of n.n. correlations as a function of the lattice size L for the two and three dimensional Ising model, at certain inverse temperatures β close to the transition point. DC solution turns out to be in good agreement with MC results and it rapidly converges to the infinite dimensional solution. The same estimation is carried out using BP and Susceptibility Propagation (SP), as discussed at the end of Section 4.1.2. BP does not take into account at all finite size corrections because of the local character of the approximation, so that its estimation is independent on the lattice size L. On the other hand, SP takes implicitly into account the structure of the graph, but it seems to overestimates the n.n. correlations at all sizes. A possible way to improve it is to exploit the normalization trick discussed in [116]: once the full correlation matrix $\boldsymbol{C}$ is estimated, all its elements are rescaled by $\hat{C}_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$; such a rescaling is introduced to heal the wrong estimation of self-correlations $C_{ii} \neq 1$ occurring in graphs with short loops, and it typically improves also the estimates of off-diagonal correlations. This is confirmed by the two plots in Figure (4.14), where the normalized version of SP (labelled as `SPnorm`) significantly improves the estimate w.r.t. its un-normalized version, still overestimating correlations if compared to DC/MC.

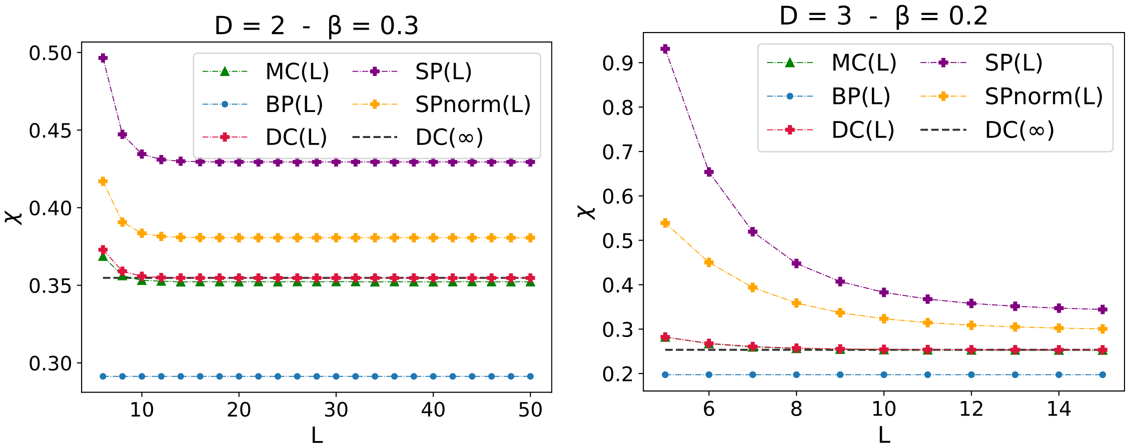

Figure 4.14: Plot of n.n. correlations with respect to the lattice size L. Comparison between DC, BP and SP (both the un-normalized and the normalized version of it) w.r.t. Monte Carlo simulations ($M = 10^6$). Left: 2−d lattice at $\beta = 0.3$, with size L^2. Right: 3−d lattice at $\beta = 0.2$, with size L^3. The black dashed line denotes the DC solution at thermodynamic limit.

4.3.6 Critical temperature scaling in the high dimensional limit

As a final comparison, we compute the series expansion of the critical temperature in the limit where the number of lattice dimensions d goes to infinity. In this limit, the Mean Field solution for the ferromagnetic Ising model is exact, and one can compute perturbative corrections in powers of $1/d$. Notice that in this case one has to rescale the couplings by their connectivity, namely $J \to J/2d$, in order to have a correctly normalized free energy density in the thermodynamic limit. For what concerns the true critical temperature β_c, there is an exact result by Fisher and Gaunt [51] where the authors perfomed a series expansion by exploting the properties of self-avoiding walks in the hypercubic lattice. Their expression is exact up to the fifth order in $1/d$, and we report it here for convenience:

$$\frac{1}{2d\beta_c} = 1 - \frac{1}{2}d^{-1} - \frac{1}{3}d^{-2} - \frac{13}{24}d^{-3} - \frac{979}{720}d^{-4} - \frac{2009}{480}d^{-5} + O\left(d^{-6}\right) \tag{4.59}$$

Starting by the analytic expression of the critical temperature of DC given by (4.58), we perform the same computation by setting $x = 1/d$ and expanding around $x = 0$. The result is shown below up to the fifth order:

$$\frac{1}{2d\beta_p} = 1 - \frac{1}{2}d^{-1} - \frac{1}{3}d^{-2} - \frac{13}{24}d^{-3} - \frac{979}{720}d^{-4} - \frac{2039}{480}d^{-5} + O\left(d^{-6}\right). \tag{4.60}$$

Comparing (4.59) and (4.60) we conclude that DC expansion is exact up to the d^{-4} order. For comparison, the Mean Field is exact up to the d^0 order; the Bethe Approximation is exact up to the d^{-1} order; Loop-Corrected Bethe [91] and Plaquette-CVM [43] are exact up to the d^{-2} order. Therefore, DC is able to correctly estimate two additional orders of magnitude of the critical temperature scaling, with respect to other state-of-the-art methods: qualitatively, this is a signal that, at least in the limit of high d, the loop corrections included by the Gaussian cavity distribution can better describe the behaviour of long-range correlations arising at the critical point.

Chapter 5

The Inverse Ising Problem

In this chapter we discuss the Inverse Ising Problem in statistical physics, providing an approximate closed-form solution for the maximum likelihood parameters through Density Consistency. After some introductory remarks and background motivations in 5.1, we formulate the Inverse Ising Problem (IIP) in a standard Bayesian setting in Section 5.2. Section 5.3 summarizes several state-of-the-art approaches to solve the IIP that will be used to compare our results. Section 5.4 presents the Density Consistency solution and its connection to other approximations. Finally, results of numerical simulation on syntethic data are shown and discussed in Section 5.5.

5.1 Motivations and applications

Inverse statistical problems attempt to reconstruct microscopic parameters describing effective interactions among the degrees of freedom of a certain system, starting from a set of measurements. Inverse problems are gaining more and more interest in recent years, thanks to improved experimental capabilities in several reserach domains and the consequent availability of large-scale datasets. This class of problems is highly interdisciplinary and applications can be found in many fields of applied science, from computational biology, neuroscience, finance, sociology, finance, non-linear optics.

In most cases, the "true" model describing the physical system is not known a priori: as a consequence, it is necessary to define an effective graphical model that tries to capture the relevant interactions between the degrees of freedom, while being compatible in some way with the observed data. In this context, the most common approach is based on maximum entropy modelling [68], that allows to construct the "least-biased" probability distribution while constraining it to reproduce some low-order statistics, observed experimentally.

The Inverse Ising Problem (IIP) represents the simplest - and yet non-trivial - scenario where we want to reproduce the first and second order statistics of the data (namely, magnetization and correlations): as shown in Section 5.2, the resulting maximum-entropy distribution takes the form of the Boltzmann measure for the Ising Hamiltonian, and the parameters to be inferred are the set of fields $\{h_i\}$ and pairwise couplings $\{J_{ij}\}$ of the Hamiltonian. This scenario applies whenever we deal with binary observations: a typical example occurs in neuroscience, where measurements of neural activity provide - in the simplest setting - a binary information where each unit (neuron) is firing some signal $(+1)$ or not (measured as 0 or -1 in terms of spin variables), and we want to recover the underlying structure of the interactions between the set of neurons under study. More in general, when measurements take values in a finite alphabet with more than 2 states, the corresponding maximum entropy distribution corresponds to the Potts model [142]. The Inverse Potts problem has several applications in computational biology, in particular in the inference

of tri-dimensional protein structures from co-evolutionary sequences: see for instance [36] and references therein for a review.
From a pure statistical physics standpoint, the IIP can be considered as the "dual" version of the forward (direct) problem: in the latter, at given Hamiltonian one wants to compute accurate estimates for the marginal probabilities - as extensively discussed in the previous chapters -, while on the former the observations are given experimentally and we need to infer the parameters encoded in the Hamiltonian.
The IIP arises as a general method to infer effective interaction networks between binary units, and applications can be found in several applied fields of pure and applied science: in neuroscience, as already cited, to reconstruct neural connections from times series of neuro spikes [35, 133]; in molecular biology, to reconstruct gene regulatory networks [81, 84]; in econophysics, to analyze stock market data and predict the behaviour of financial markets [17, 24].
In the rest of the chapter, we will address the IIP as a standard Bayesian inference problem by defining the posterior distribution over the model parameters (couplings and fields) given the observed data. However, an exact computation of such quantities scales exponentially with the system size and it can be carried out only when the number of variables is small. For this reason, in the statistical physics community several approximation schemes have been developed to estimate model parameters: we will brefly review some of them in section 5.3.
In particular, we will distinguish between a class of mean-field like approximations and other iterative methods: the formers allow to compute closed-form expressions for the model parameters depending only on the first two empirical moments. Among the iterative methods, we will mainly focus on Pseudolikelihood Maximization [6] that is widely considered as the best algorithmic approach to solve the Inverse Ising Problem [101].
The purpose of this chapter is to show that Density Consistency allows to approximate the maximum likelihood equations for the model parameters, providing a closed-form expression that resembles a known result found by Sessak and Monasson in [124]: in particular, depending on the closure condition chosen for the DC update rules, different expressions can be obtained, that allow to improve the reconstruction quality with respect to other methods, especially in a regime with a small number of observations.
In this chapter we consider only equilibrium reconstruction, i.e. assuming that observation come from an equilibrium model where couplings are symmetric and detailed balance holds. In principle, the Inverse Ising Problem can be exploited to infer a non-equilibrium models from time-series data: in this case, couplings are expected to be asymmetric and the corresponding dynamics leads to a non-equilibrium steady state, different from the Boltzmann distribution. See for instance [101] for an exhaustive review.

5.2 Problem setup

5.2.1 Maximum-entropy modeling

Suppose we are provided with a set of M measurements $\{\boldsymbol{\sigma}^\mu\}^{\mu=1,\dots,M}$, each sample $\boldsymbol{\sigma}^\mu$ being a $N-$dimensional vector of binary variables, $\sigma_i^\mu \in \{-1,1\}$. Data might be provided from experiments or artificially generated from a known model by using Monte Carlo sampling techniques. We are interested in deriving the maximum-entropy distribution with specified first and second order statistics. In the following, we will denote with $\mathcal{D} = \{\boldsymbol{\sigma}^\mu\}^{\mu=1,\dots,M}$ the set of configurations, whose first two moments are given by:

$$m_i \hat{=} \langle \sigma_i \rangle_{\mathcal{D}} = \frac{1}{M} \sum_{\mu=1}^{M} \sigma_i^{\mu}, \tag{5.1}$$

$$\chi_{ij} \hat{=} \langle \sigma_i \sigma_j \rangle_{\mathcal{D}} = \frac{1}{M} \sum_{\mu=1}^{M} \sigma_i^{\mu} \sigma_j^{\mu} \tag{5.2}$$

where the symbol $\langle \cdot \rangle_{\mathcal{D}}$ is understood as the average w.r.t. the empirical distribution $p_{\mathcal{D}}$, formally expressed as

$$p_{\mathcal{D}}(\boldsymbol{\sigma}) = \frac{1}{M} \sum_{\mu=1}^{M} \delta(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\mu}). \tag{5.3}$$

In the rest of the chapter, we will denote with $\boldsymbol{m} \in \mathbb{R}^N$ the vector of empirical magnetizations and with $\boldsymbol{C} = \boldsymbol{\chi} - \boldsymbol{m}\boldsymbol{m}^t \in \mathbb{R}^{N \times N}$ the (symmetric) matrix of connected correlations. Following the approach developed by Jaynes [68], we define the following constrained variational entropy:

$$\begin{aligned} \mathcal{S}[p] = & - \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \log p(\boldsymbol{\sigma}) + \gamma \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) - 1 \right) \\ & + \sum_i h_i \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sigma_i - m_i \right) + \sum_{i<j} J_{ij} \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sigma_i \sigma_j - \chi_{ij} \right) \end{aligned} \tag{5.4}$$

which is a functional of the distribution p. The parameters $\gamma, \{h_i\}, \{J_{ij}\}$ included as Lagrange multipliers constraint the probability p to be normalized (γ) and to have prescribed first and second order statistics. By setting to 0 the functional derivative of (5.4), namely $\frac{\delta \mathcal{S}}{\delta p} = 0$, it is straightforward to recover the Boltzmann measure for the Ising model, the Lagrange parameters playing the role of local fields $\{h_i\}$ and pairwise couplings $\{J_{ij}\}$:

$$p(\boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{J}) = \frac{1}{Z} \exp\left[\sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \right] = \frac{1}{Z} \exp\left[-H(\boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{J}) \right] \tag{5.5}$$

where $H(\boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{J}) = -\sum_{i<j} J_{ij} \sigma_i \sigma_j - \sum_i h_i \sigma_i$ is the Ising Hamiltonian and

$$Z(\boldsymbol{h}, \boldsymbol{J}) = \sum_{\boldsymbol{\sigma}} e^{-H(\boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{J})} \tag{5.6}$$

is the partition function. There are two important differences to notice w.r.t. Eq. (1.14): first, the summation over couplings runs over all distinct pairs $i < j$ (again, assuming that the graph is undirected), because we do not know a priori the interaction network and all couplings need to be estimated. Secondly, in inverse problems the temperature cannot be determined in general as a free parameter, since the statistics (5.5) depends on β only through the products βh_i and βJ_{ij}: for this reason, we will assume β to be implicitly absorbed into the model parameters, namely $\beta h_i \to h_i$ and $\beta J_{ij} \to J_{ij}$[1].

This is the simplest setup where topological connection between pairs of degrees of freedom can be inferred: for instance, constraining only the first moments would result into a factorized

[1] Neverthless, in Section 5.5 we will still use the parameter β to vary the strenght of couplings/fields in order to quantify the reconstruction quality with respect to the temperature.

model over single spins (*independent spin model*), which is poorly relevant for applications. On the other hand, one could reasonably think about constraining also high-order statistics. For instance, matching also 3rd-order moments in the maximum-entropy distribution would result into an Ising model with 3-body interactions $\sum_{i\neq j\neq k} K_{ijk}\sigma_i\sigma_j\sigma_k$, and the corresponding number of parameters to be inferred would scale as $O\left(N^3\right)$: as a consequence, a much larger amount of configurations would be required in order to accurately infer all the parameters, which is not possible in many experimental setups. On the other hand, in many applications there is no need of such overparametrization, since third-order experimental moments can be well explained by a model with pairwise interactions [127]. In addition, a model with third order interactions can be approximated by retaining only pairwise couplings, especially when the third order couplings are dense [88]. In this perspective, constraining only 2-body correlations results into the simplest and yet non-trivial effective model where only pairwise couplings have to be determined.

5.2.2 Bayesian approach

The maximum-entropy approach is used to derive the simplest (in terms of entropy) model able to explain a set of empirical statistics. In practice, by using the above principle the problem of finding the best model to describe the dataset turns into the problem of finding the optimal set of parameters.

From now on, we can equivalently assume that the dataset $\mathcal{D}$ has been generated from the distribution (5.5), thus we are left with the pratical problem of estimating the couplings and fields, given the observed configurations. Following a standard Bayesian approach, the *posterior* probability of the model parameters $\boldsymbol{\theta} = (\boldsymbol{h}, \boldsymbol{J})$ given the data is given by the Bayes Theorem:

$$p\left(\boldsymbol{\theta}|\mathcal{D}\right) = \frac{p\left(\mathcal{D}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)}{p\left(\mathcal{D}\right)} \tag{5.7}$$

where $p\left(\boldsymbol{\theta}\right)$ is the prior distribution and $p\left(\mathcal{D}|\boldsymbol{\theta}\right) = p\left(\left\{\boldsymbol{s}^\mu\right\}_{\mu=1}^M|\boldsymbol{\theta}\right)$ is the likelihood function: the latter represents the probability that the set of configurations $\left\{\boldsymbol{\sigma}^\mu\right\}_{\mu=1}^M$ has been drawn from the starting distribution (5.5) with parameters $\boldsymbol{\theta}$, to be interpreted as a function of $\boldsymbol{\theta}$. The prior distribution can take into account additional information about the model parameters: for instance, if we know that the underlying graph is sparse, it is possible to enforce this condition by means of ℓ^p-norm prior (also known as regularization) [63], so to penalize large values for the inferred couplings. In the present work we will discard any prior information (i.e. choosing $p\left(\boldsymbol{\theta}\right)$ to be uniformly distributed), so that the MAP estimator coincides with the maximum likelihood (ML) point. In the IIP, under the assumption that samples are independent and identically distributed (i.i.d), the log-likelihood function[2] can be expressed as:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}\left(\boldsymbol{h},\boldsymbol{J}\right) \hat{=} & \frac{1}{M}\sum_{\mu=1}^{M}\log p\left(\boldsymbol{\sigma}^{\mu}\mid\boldsymbol{h},\boldsymbol{J}\right) \\ & =\sum_i h_i\frac{1}{M}\sum_i s_i^{\mu}+\sum_{i<j}J_{ij}\frac{1}{M}\sum_{\mu=1}^{M}s_i^{\mu}s_j^{\mu}-\log Z\left(\boldsymbol{h},\boldsymbol{J}\right) \\ & =\sum_i h_i m_i+\sum_{i<j}J_{ij}\chi_{ij}-\log Z\left(\boldsymbol{h},\boldsymbol{J}\right) \end{aligned} \tag{5.8}$$

where Z is the partition function defined in Eq. (5.6). The ML estimator can be iteratively found by means of gradient-ascent algorithm, a procedure known as Boltzmann Learning [1], whose update rules are shown below:

$$h_i^{t+1}=h_i^t+\eta\frac{\partial\mathcal{L}_{\mathcal{D}}}{\partial h_i}=h_i^t+\eta\left(\langle\sigma_i\rangle_{\mathcal{D}}-\langle\sigma_i\rangle_p\right)\ \forall i \tag{5.9}$$

$$J_{ij}^{t+1}=J_{ij}^t+\eta\frac{\partial\mathcal{L}_{\mathcal{D}}}{\partial J_{ij}}=J_{ij}^t+\eta\left(\langle\sigma_i\sigma_j\rangle_{\mathcal{D}}-\langle\sigma_i\sigma_j\rangle_p\right)\ \forall i\neq j \tag{5.10}$$

where η is the learning rate. The maximum likelihood point is found by setting to zero the derivatives of (5.8) w.r.t. $\boldsymbol{\theta}$, and it satisfies the following set of moment matching conditions:

$$m_i=\langle\sigma_i\rangle_p\ \forall i, \tag{5.11}$$

$$\chi_{ij}=\langle\sigma_i\sigma_j\rangle_p\ \forall i\neq j \tag{5.12}$$

The above formulas imply that the maximum likelihood point (or estimator) $\boldsymbol{\theta}^{ML}$ is found when the first and second empirical moments coincide with the expectation values over the Boltzmann measure (5.5), provinding a way to fix the parameters derived through the maximum-entropy approach. Furthermore, it can be proven that the maximum likelihood estimator converges in probability to the true value when the number of samples goes to infinity: this property is called *consistency*. Computing expectation values over the equilibrium distribution (5.5) scales exponentially with the system size ($O\left(2^N\right)$) and it can be done explicitly only for very small systems. In general, one must rely on some suitable approximations to estimate the model parameters, that will be discussed in the next section.

5.3 Review of methods

In this section, we are going to review some state-of-the-art techniques commonly employed for the IIP. In particular, we will distinguish between a class of *mean-field*-like approximations and other iterative methods; among the latters we will focus on Pseudolikelihood, that is widely considered as the outperformer for the IIP.

Mean-field like methods can be obtained from suitable approximations of the Ising Free energy.

[2] (5.8) refers to the logarithm of the likelihood normalized over the number of samples: this choice is typically more convenient for numerical reasons. Indeed, the likelihood scales exponentially with the number of samples, while its normalized logarithm is intensive w.r.t. M. Moreover, replacing products with summations thanks to the logarithm is more convenient to avoid dealing with small numbers. Maximizing the likelihood is equivalent to maximize its logarithm since the latter is a strictly monotone function.

In inverse problems, the correct thermodynamic potential to be employed is obtained by performing a Legendre transform of the Helmoltz free energy, both with respect to magnetizations and correlations:

$$S\left(\boldsymbol{m},\boldsymbol{\chi}\right)=\min_{\boldsymbol{h},\boldsymbol{J}}\left[-\sum_i h_i m_i-\sum_{i<j}J_{ij}\chi_{ij}-F\left(\boldsymbol{h},\boldsymbol{J}\right)\right] \tag{5.13}$$

which is nothing but the entropy at fixed $\boldsymbol{m}$,$\boldsymbol{\chi}$. A simple inspection of (5.13) shows that, apart from a sign, it gives the maximum likelihood estimator. In this perspective, the potential (5.13) provides a link between the statistical physics and the Bayesian approach for the Inverse Ising Problem. Couplings and fields can estimated from (5.13) through the following relations:

$$h_i=-\frac{\partial S\left(\boldsymbol{m},\boldsymbol{\chi}\right)}{\partial m_i};\qquad J_{ij}=-\frac{\partial S\left(\boldsymbol{m},\boldsymbol{\chi}\right)}{\partial \chi_{ij}} \tag{5.14}$$

where $\boldsymbol{m}$ and $\boldsymbol{\chi}$ are computed from the empirical dataset. Another useful thermodynamic potential can be obtained by from F by transforming only w.r.t. the fields, leading to the so-called *Gibbs* free energy:

$$G\left(\boldsymbol{m},\boldsymbol{J}\right)=\min_{\boldsymbol{h}}\left[-\sum_i h_i m_i-F\left(\boldsymbol{h},\boldsymbol{J}\right)\right] \tag{5.15}$$

The above expression turns out to be useful in many cases where correlations cannot be easily derived. In this scenario, while fields can be directly estimated by using $h_i=-\partial G\left(\boldsymbol{m},\boldsymbol{J}\right)/\partial m_i$, couplings can instead be reconstructed by using Linear response theory [71]:

$$\left(\boldsymbol{C}^{-1}\right)_{ij}=\frac{\partial h_i}{\partial m_j}=-\frac{\partial G\left(\boldsymbol{m},\boldsymbol{J}\right)}{\partial m_i\partial m_j} \tag{5.16}$$

to be solved w.r.t. J_{ij} at fixed $\boldsymbol{m}$,$\boldsymbol{C}$ (again, computed from the empirical measurements). All the mean-field like approximations discussed in the following allow to compute closed-form expressions for the model parameters in terms of the empirical magnetizations and correlations, so that no iterative procedure is required.

5.3.1 Naive Mean Field

The simplest approximation in statistical physics is the Naive Mean field (MF) approach, already introduced in Section 2.1.1. In the context of the Inverse Ising Problem, the first attempt in using a MF approximation was developed in [71, 111]. Since correlations are not taken into account by MF, one needs to rely to linear response to estimate the couplings, so that the suitable thermodynamic potential to employ is the Gibbs free energy (5.15). In particular, the corresponding MF approximation to (5.15) can be obtained using the ansatz (2.12) and performing a Legendre transform as in (5.15), leading to:

$$G_{MF}\left(\boldsymbol{m},\boldsymbol{J}\right)=\sum_i\left[\mathcal{H}\left(\frac{1+m_i}{2}\right)+\mathcal{H}\left(\frac{1-m_i}{2}\right)\right]+\sum_{i<j}J_{ij}m_im_j \tag{5.17}$$

where $\mathcal{H}\left(x\right)=-x\log x$. Using $h_i=-\partial G\left(\boldsymbol{m},\boldsymbol{J}\right)/\partial m_i$, the local fields are computed as:

$$h_i=\text{atanh}m_i-\sum_{j\neq i}J_{ij}m_j \tag{5.18}$$

Notice that (5.18) coincides with the MF fixed point equation (2.19), this time solved w.r.t. the local fields h_i. By exploiting the linear response relations (5.16) on (5.18), one gets a direct expression for the inferred couplings under the MF approximation:

$$J_{ij}^{MF}=-\left(\boldsymbol{C}^{-1}\right)_{ij} \tag{5.19}$$

Finally, fields can be reconstructed by inserting (5.19) into (5.18) and solving for h_i, leading to:

$$h_i^{MF} = \text{atanh} m_i - \sum_j J_{ij}^{MF} m_j \tag{5.20}$$

5.3.2 TAP

TAP equations are named after Thouless, Anderson and Palmer, who derived a Mean-Field theory for the fully-connected spin glass (SK) model [126, 130]. The TAP formalism can be also employed for the Inverse Ising Problem, as first carried out in [129]. The TAP free energy can be obtained starting from the MF expression (5.17) by adding the so-called Onsager's correction term [130]:

$$\begin{aligned} G^{TAP}(\boldsymbol{m}, \boldsymbol{J}) = & \sum_i \left[\mathcal{H}\left(\frac{1+m_i}{2}\right) + \mathcal{H}\left(\frac{1-m_i}{2}\right) \right] \\ & + \sum_{i<j} \left[J_{ij} m_i m_j + \frac{1}{2} J_{ij}^2 \left(1 - m_i^2\right)\left(1 - m_j^2\right) \right] \end{aligned} \tag{5.21}$$

Its stationary points satisfy the following set of self-consistent equations:

$$m_i = \tanh \left[h_i + \sum_{j \neq i} J_{ij} m_j - m_i \sum_{j \neq i} J_{ij}^2 \left(1 - m_j^2\right) \right] \qquad \forall i \tag{5.22}$$

Eqs. (5.22) are the TAP equations: in the forward problem, they can be iteratively solved at fixed couplings and fields, providing an exact solution in the thermodynamic limit for the fully connected spin glass. Indeed, for the SK model with random (e.g. Gaussian) couplings, in order to have an intensive free energy density one needs to rescale the couplings as $J_{ij}/\sqrt{N}$; in turn, this implies that the second order terms in J_{ij}^2 cannot be neglected in the thermodynamic limit. On the other hand, in the ferromagnetic case the Onsager's reaction term goes to 0 (since $J_{ij} \sim 1/N$ and the second order correction is negligible), so that one recovers the MF theory. The TAP solution for the inverse Ising problem follows the same reasoning discussed in the previous section for the MF inference: starting from (5.22), we rewrite it to isolate the local field h_i, so that couplings can be estimated by solving the Linear response relation (5.16) with respect to J_{ij}. Their expression is finally put back into (5.22) to reconstruct the fields. The final result is shown below:

$$J_{ij}^{TAP} = \frac{\sqrt{1 - 8 m_i m_j \left(\boldsymbol{C}^{-1}\right)_{ij}} - 1}{4 m_i m_j} \tag{5.23}$$

$$h_i^{TAP} = \text{atanh} m_i - \sum_j J_{ij}^{TAP} m_j + m_i \sum_{j \neq i} \left(J_{ij}^{TAP}\right)^2 \left(1 - m_j^2\right) \tag{5.24}$$

As a final remark, notice that in the limit of zero magnetizations the TAP inference for the couplings tends to the MF expression (5.19).

A note on the Plefka expansion

A seminal work by Plefka [112] showed how to derive the MF and TAP free energy approximations (and high-order corrections) starting from the thermodynamic potential $G(\boldsymbol{m}, \boldsymbol{J})$ and performing a perturbative expansion in small couplings. With no aim of giving any computational details and referring to [56, 101, 112] for additional details, such expansion can be performed by setting $J_{ij} \to \lambda J_{ij}$ and expanding in powers of λ:

$$G^{(\lambda)}(\boldsymbol{m}, \boldsymbol{J}) = G^{(0)}(\boldsymbol{m}, \boldsymbol{J}) + \lambda \left.\frac{\partial G}{\partial \lambda}\right|_{\lambda=0} + \frac{\lambda^2}{2} \left.\frac{\partial^2 G}{\partial \lambda^2}\right|_{\lambda=0} + \dots \tag{5.25}$$

In particular, the 0-th order term of (5.25) corresponds to a non-interacting model where each magnetization m_i is fixed by the conjugate local field h_i (this is the independent spin model). The series expansion can be performed since high-order contributions are computed w.r.t. to the non-interacting case $\lambda = 0$, that can be easily handled. By truncating the expansion to a certain k-th order and setting $\lambda = 1$, one can derive closed expressions for the local fields h_i in terms of the magnetizations and couplings; the latter can instead be reconstructed by using linear response theory, as previously discussed. In this notation, truncating to the 1-st order gives the MF inference, while adding the 2-nd order term leads to the TAP approximation. A systematic way to compute high-order terms of the series was carried out by Georges and Yedidia in [56]: however, it has been shown that adding high-order contributions does not necessarily improve the inference w.r.t. to MF/TAP [116].

5.3.3 Independent Pair Approximation

The independent pair approximation (IIP) is one of the easiest approaches to the Inverse Ising problem, where - as the name suggests - the inference is performed for each pair of spin separately, as if they were independent on the others. First developed by Roudi et al. [118, 119], it turns out to be exact if the topology is known and the interaction graph is acyclic (in the limit of infinite M).

Consider the pair (i,j) where $i,j \in \{1,\dots,N\}$, their joint probability measure $p(\sigma_i, \sigma_j)$ defined over $\{-1,1\}^2$ can be written in two equivalent ways, either in terms of its indepedent moments (m_i, m_j, χ_{ij}), or in the Boltzmann form (5.5):

$$p(\sigma_i, \sigma_j \mid m_i, m_j, \chi_{ij}) = \frac{1 + m_i\sigma_i + m_j\sigma_j + \chi_{ij}\sigma_i\sigma_j}{4} \tag{5.26}$$

$$p(\sigma_i, \sigma_j \mid h_i, h_j, J_{ij}) = \frac{e^{J_{ij}\sigma_i\sigma_j + h_i^{(ij)}\sigma_i + h_j^{(ij)}\sigma_j}}{4\left(\cosh h_i \cosh h_j \cosh J_{ij} + \sinh h_i \sinh h_j \sinh J_{ij}\right)} \tag{5.27}$$

In latter formula, $h_i^{(ij)}$ (resp. $h_j^{(ij)}$) represents the local field associated to i (resp. j) when considered only in pair with node j (resp. i). The two sets of parameters are in one-to-one correspondence: in particular, the expression for the moments in terms of the couplings have already been used in the previous Chapter (Sec. 4.1.1), to estimate the marginal tilted moments through Density Consistency. Conversely, the inverse relations are shown below:

$$J_{ij}^{IP} = \sum_{\sigma_i,\sigma_j} \frac{\sigma_i\sigma_j}{4} \log p(\sigma_i,\sigma_j) = \frac{1}{4}\log\frac{[1+m_i+m_j+\chi_{ij}]\,[1-m_i-m_j+\chi_{ij}]}{[1+m_i-m_j-\chi_{ij}]\,[1-m_i+m_j-\chi_{ij}]} \tag{5.28}$$

$$h_i^{(ij)IP} = \sum_{\sigma_i,\sigma_j} \frac{\sigma_i}{4} \log p(\sigma_i,\sigma_j) = \frac{1}{4}\log\frac{[1+m_i+m_j+\chi_{ij}]\,[1+m_i-m_j-\chi_{ij}]}{[1-m_i+m_j-\chi_{ij}]\,[1-m_i-m_j+\chi_{ij}]} \tag{5.29}$$

$$h_j^{(ij)IP} = \sum_{\sigma_i,\sigma_j} \frac{\sigma_j}{4} \log p(\sigma_i,\sigma_j) = \frac{1}{4}\log\frac{[1+m_i+m_j+\chi_{ij}]\,[1-m_i+m_j-\chi_{ij}]}{[1+m_i-m_j-\chi_{ij}]\,[1-m_i-m_j+\chi_{ij}]} \tag{5.30}$$

where at the right-most hand sides the functional form (5.26) is used. (5.28) can be used as it is to infer the coupling between the two nodes i and j, just in terms of their moments. On the other hand, to reconstruct the overall local field h_i one needs to sum all the terms (5.29) for $j \neq i$. As noted in [119], it is necessary to correct the resulting expression by removing the single-site's contribution to (5.29) for each pair of spin, that would be overcounted otherwise. This contribution is equal to $\text{atanh} m_i$ for each distinct pair, so that the final expression for the inferred fields under the Independent Pair approximation can be written as:

$$h_i^{IP} = \sum_{j \neq i} h_i^{(ij)IP} - (N-2)\,\text{atanh} m_i \tag{5.31}$$

From a statistical physics standpoint, the Independent Pair approximation can be derived by plugging in the Bethe ansatz on the thermodynamic potential $S(\boldsymbol{m}, \boldsymbol{\chi})$. Indeed, the Bethe approximation considers exactly the contribution of spin pairs to the entropy, as discussed in Chapter (2). By plugging in the Bethe ansatz introduced in Eq. (2.24) and using (5.13), we get:

$$\mathcal{S}^{\text{Bethe}}(\boldsymbol{m}, \boldsymbol{\chi}) = \sum_{i<j} \sum_{\sigma_i, \sigma_j} \mathcal{H}\left[\frac{1 + m_i\sigma_i + m_j\sigma_j + \chi_{ij}\sigma_i\sigma_j}{4}\right] + \sum_i \sum_{\sigma_i} (2-N)\,\mathcal{H}\left[\frac{1+m_i\sigma_i}{2}\right] \tag{5.32}$$

Using Eqs. (5.14) on the above formula leads exactly to the same expression for the inferred couplings (5.28) and fields (5.31). In this case, the second term in (5.31) naturally arises from the single-site entropy contribution in (5.32).
This method can be also used if the topology is known: in particular, Eq. (5.31) still holds for those coupling we know to be edges in the graph, and the local field on each spin is computed by summing only on its neighbours, i.e. by replacing $\sum_{j \neq i}$ with $\sum_{j \in \partial i}$ and $(N-2)$ with $(d_i - 1)$, d_i being node i's degree. The Independent Pair approximation can in principle be improved by including entropic contributions of clusters with increasing size into (5.32): an iterative approach to compute these corrections was developed by Cocco and Monasson in [33, 34] and it is known as the Adaptative Cluster Expansion (ACE).

5.3.4 Susceptibility Propagation

Another way of exploiting the Bethe approximation is to estimate couplings through linear response. This approach was first developed by Mezard and Mora in [100], specifically for the Inverse Ising Problem: the authors designed an iterative message-passing scheme defined over cavity *susceptibilities* rather than cavity messages, computed by using linear response theory on the BP messages discussed in Section 2.2. This approach resembles a previous work by Welling and Teh in [138, 139] for the forward problem and it is known as Susceptibility Propagation algorithm (SP). The authors of [100] presented an iterative scheme to update also the couplings at fixed empirical magnetizations, so to provide an approximate solution for the inverse Ising Problem. It was further noticed in [116] that the Bethe approximation allows to *analytically* compute couplings through linear response, with no need of running an iterative algorithm. We now follow the latter approach. The key point of [116] is that the BP equations on Ising-like models can be rewritten only in terms of single-node beliefs, without relying on cavity messages. In this way, we get a set of self-consistent equations for the local magnetizations, similarly to what derived for the MF and TAP approximations. We just report the final result, and we refer to [116] for additional details:

$$m_i = \tanh\left[h_i + \sum_{j \neq i} t_{ij} f(m_j, m_i, t_{ij})\right] \tag{5.33}$$

where

$$f\left(m_1, m_2, t\right) = \frac{\left(1-t^2\right) - \sqrt{\left(1-t^2\right)^2 - 4t\left(m_1 - tm_2\right)\left(m_2 - tm_1\right)}}{2\left(m_2 - tm_1\right)} \tag{5.34}$$

and $t_{ij} = \tanh J_{ij}$. Solving (5.33) gives the same fixed points of Belief Propagation on any graph topology, involving only single-node magnetizations with no need of defining cavity messages. From the point of view of the Plefka expansion, (5.33) can be derived from by summing all the 2-spin contributions in the Gibbs free energy (5.15), i.e. all the terms like $\sum_{i<j} J_{ij}^k$ appearing in (5.25). Indeed, if the graph has no loops, additional terms like $J_{ij}J_{jk}J_{ki}$ (and higher-order contributions) do not appear in the Plefka expansion, a further confirmation that the Bethe approximation is exact on acyclic graphs.

In principle, the fixed point equations (5.33) can be also derived by using the Bethe ansatz for the free energy and rewriting the nearest-neighbours correlations in terms of the single-site beliefs. With regard to the Inverse Ising Problem, the advantage of (5.33) is that couplings can be straightforwardly computed by using linear response theory, analogously to the MF and TAP inference. First notice that, by applying the LR relations (5.16) to (5.33), any element of the inverse covariance matrix $\boldsymbol{C}^{-1}$ can be easily computed:

$$\left(\boldsymbol{C}^{-1}\right)_{ij} = \left[\frac{1}{1-m_i^2} - \sum_k \frac{t_{ik} f_2\left(m_k, m_i, t_{ik}\right)}{1 - t_{ik}^2 f^2\left(m_k, m_i, t_{ik}\right)}\right]\delta_{ij} - \frac{t_{ij} f_1\left(m_j, m_i, t_{ij}\right)}{1 - t_{ij}^2 f^2\left(m_j, m_i, t_{ij}\right)} \tag{5.35}$$

where

$$f_1\left(m_1, m_2, t\right) \hat{=} \frac{\partial f\left(m_1, m_2, t\right)}{\partial m_1}; \qquad f_2\left(m_1, m_2, t\right) \hat{=} \frac{\partial f\left(m_1, m_2, t\right)}{\partial m_2} \tag{5.36}$$

and m_i are the fixed points of (5.33). Eq. (4.4) is the same used to compute LR correlations for the forward problem in the previous Chapter (in particular, a preliminar set of results is shown in Figures 4.4 and 4.14). Here instead we are interested in retrieving couplings and fields when both the magnetizations and the full covariance matrix are known. In order to do that, it is sufficient to invert Eq. 5.35 and solving it for $J_{ij}, i \neq j$ to infer the couplings. Notice that each equations (4.4) for $i \neq j$ can be solved independently on the others. We now report the final expression for the couplings and fields inferred under the Bethe approximation in Linear Response (which we will refer to in the following as Susceptibility Propagation, SP):

$$\begin{split} J_{ij}^{SP} = -\text{atanh}\Bigg[& \frac{1}{2\left(\boldsymbol{C}^{-1}\right)_{ij}}\sqrt{1 + 4\left(1-m_i^2\right)\left(1-m_j^2\right)\left(\boldsymbol{C}^{-1}\right)_{ij}^2} - m_i m_j - \\ & \frac{1}{2\left(\boldsymbol{C}^{-1}\right)_{ij}}\sqrt{\left(\sqrt{1 + 4\left(1-m_i^2\right)\left(1-m_j^2\right)\left(\boldsymbol{C}^{-1}\right)_{ij}^2} - 2m_i m_j\left(\boldsymbol{C}^{-1}\right)_{ij}\right)^2 - 4\left(\boldsymbol{C}^{-1}\right)_{ij}^2}\,\Bigg] \end{split} \tag{5.37}$$

$$h_i^{SP} = \text{atanh} m_i - \sum_{j \neq i} \text{atanh}\left[\tanh J_{ij}^{SP} f\left(m_j, m_i, \tanh J_{ij}^{SP}\right)\right] \tag{5.38}$$

where, again, the fields are recovered by inserting (5.37) into (5.33). The above formulas provide the exact expression of the inferred couplings and fields (in the limit of infinite samples) on a tree, even if the topology is not known a priori, at difference with the Independent Pair approximation. As noted in [116], both SP and TAP reconstruction suffer of a domain issue: it might happen that, depending on the particular instance considered, the argument of the square roots of (5.37)-(5.23) become negative, thus resulting into a non-physical solution. As a final remark, notice that Eq. (5.37) can be used to estimate long-range correlations for the forward Ising Problem, by inverting it and solving w.r.t. $\boldsymbol{C}$ at fixed couplings (the same reasoning actually holds for the TAP and MF approximations as well).

5.3.5 Sessak-Monasson approximation

Another statistical physic approach for the IIP exploits a small-correlation expansion (SCE) for the Ising Free energy at fixed magnetizations and correlations, performed by Sessak and Monasson in [124]. The idea is similar to the Plefka expansion [112], with the difference that the series expansion is performed perturbatively with respect to the connected correlation $\chi_{ij} - m_i m_j$, on the thermodynamic potential obtained by Legendre-transforming the Ising Free energy both w.r.t. magnetizations and correlations (5.13): the advantage of such approach is that couplings can be directly inferred, without relying on Linear Response theory. The authors of [124] presented a systematic way to compute arbitrary high-order perturbative contributions to the inferred couplings and fields, and showed how to represent each contribution in terms of loop diagrams. We report below for convenience the expressions for the inferred couplings and fields up to the fourth order (further details can be also found in [123]):

$$\begin{aligned} J_{ij}^{SCE} &= \beta K_{ij} - 2\beta^2 m_i m_j K_{ij}^2 - \beta \sum_k K_{jk} K_{ki} L_k + \\ &+ \frac{1}{3}\beta^3 K_{ij}^2 \left[1 + 3m_i^2 + 3m_j^2 + 9m_i^2 m_j^2\right] + \beta^3 \sum_{k \neq i, j} K_{ij} \left(K_{jk}^2 L_j + K_{ki}^2 L_i\right) L_k \\ &+ \beta^3 \sum_{k \neq i, l \neq j} K_{jk} K_{kl} K_{li} L_k L_l + O\left(\beta^4\right) \end{aligned} \tag{5.39}$$

$$\begin{aligned} h_i^{SCE} &= \text{atanh} m_i - \sum_j J_{ij}^{SCE} m_j + \beta^2 \sum_{j \neq i} K_{ij}^2 m_i L_j + \\ &- \frac{2}{3}\beta^3 \left(1 + 3m_i^2\right) \sum_{j \neq i} K_{ij}^3 m_j L_j - 2\beta^3 m_i \sum_{j<k} K_{ij} K_{jk} K_{ki} L_j L_k + O\left(\beta^4\right) \end{aligned} \tag{5.40}$$

where $K_{ij} = C_{ij} / \left(C_{ii} C_{jj}\right)$, $\boldsymbol{C} = \boldsymbol{\chi} - \boldsymbol{m}\boldsymbol{m}^t$ and β is a perturbative parameter, playing the role of a fictious temperature: in practice, the small correlation expansion can be easily performed by setting $C_{ij} \to \beta C_{ij}$, and then expanding over β. In this perspective, the 0-th order term corresponds to the non-interacting model, that can be easily handled. At any order β^k, the couplings and fields can be estimated by putting $\beta = 1$, in the same way as for the Plefka expansion. Moreover, the authors of [124] were able to sum all the contributions of loop and 2-spin diagrams for the couplings at any order, whose expression was further simplified in [118] and it is shown below:

$$J_{ij}^{SM} = J_{ij}^{IP} - \left(\boldsymbol{C}^{-1}\right)_{ij} - \frac{C_{ij}}{C_{ii} C_{jj} - C_{ij}^2} \qquad \forall i \neq j \tag{5.41}$$

The above formula is a combination of the Independent-Pair approximation J_{ij}^{IP} (5.28), that accounts for all the 2-spin diagrams (i.e. all the terms like K_{ik}^p in (5.39)), and of the Mean-Field inference $-\boldsymbol{C}^{-1}$ (5.19); the last term avoids the overcounting of 2−spin diagrams appearing in the MF term and it corresponds to the coupling inferred through MF on a system with 2 spins only. (5.41) is a non-perturbative expression that accounts for certain types of loop contributions *at any order*. Clearly, an infinite number of diagrams appearing at higher orders w.r.t. $O\left(\beta^4\right)$ (for instance, all the high-order $k-$spin diagrams with $k \geq 3$) are neglected.

In the literature, with the name "Sessak-Monasson" approximation one refers to the closed-form result (5.41) for the couplings, rather than the expansion (5.39). Conversely, summing the same types of loop contribution is not possible for the local fields h_i, whose expression it is typically left in the series form (5.40). However, in Section 5.4.1 we will show that the analogy with the

Density Consistency inference provides a way to write a closed-form expression for the local fields obtained by summing up all the loop and 2−spin contributions.

5.3.6 Pseudo-Likelihood

All the methods described so far give analytic expressions for the inferred parameters as functions of the empirical magnetizations and 2-point correlations and, with the exception of the Independent Pair Approximation (Setion 5.3.3) all of them require a single inversion of the covariance matrix $\boldsymbol{C}$, whose computational cost scales as $O\left(N^3\right)$. Notice also that the Boltzmann learning procedure to maximize the likelihood only involves the first two moments to update fields and couplings by means of gradient-ascent, as clear from (5.9)-(5.10). An alternative and powerful approach to the exact likelihood maximization approach, known as Pseudo-likelihood (PL), allows to exploit all the information encoded in the dataset - i.e. also high-order correlations - to iteratively compute the model parameters with a polynomial running time, both w.r.t. the system size N and the number of samples M. It was first introduced in [13] and then re-discovered in the statistical physics community more recently [6]. The key approximation behind PL is that the likelihood function (5.8) can be simplified by taking into account the effect of one spin explicitly and approximate the rest of the degrees of freedom with the empirical distribution. Let us start from the Boltzmann measure (5.5): for a certain spin i, we can exploit the chain rule to write the conditional probability of node i, given all the others:

$$p\left(\boldsymbol{\sigma}\right)=p\left(\sigma_i,\boldsymbol{\sigma}_{\setminus i}\right)=p\left(\sigma_i\mid\boldsymbol{\sigma}_{\setminus i}\right)p\left(\boldsymbol{\sigma}_{\setminus i}\right) \tag{5.42}$$

where $\setminus i=\{j=1,\ldots,N\mid j\neq i\}$ denotes the set of all degrees of freedom but i, and the dependency on $(\boldsymbol{h},\boldsymbol{J})$ is dropped for simplicity. The quantity $p\left(\sigma_i\mid\boldsymbol{\sigma}_{\setminus i}\right)$ is the probability distribution of spin i conditioned on all the others, explicitly given by:

$$p\left(\sigma_i\mid\boldsymbol{\sigma}_{\setminus i}\right)=\frac{e^{\sigma_i\tilde{h}_i\left(\boldsymbol{\sigma}^{\mu}_{\setminus i}\right)}}{2\cosh\tilde{h}_i\left(\boldsymbol{\sigma}^{\mu}_{\setminus i}\right)};\qquad \tilde{h}_i\left(\boldsymbol{\sigma}_{\setminus i}\right)=h_i+\sum_{j\neq i}J_{ij}\sigma_j \tag{5.43}$$

where $\tilde{h}_i$ is a local effective field depending on all the $\boldsymbol{\sigma}_{\setminus i}$'s states. Exploiting (5.42) and (5.43) it is possible to express moments of spin i, shown below:

$$\langle\sigma_i\rangle_{p\left(\sigma_i\mid\boldsymbol{\sigma}_{\setminus i}\right)p\left(\boldsymbol{\sigma}_{\setminus i}\right)}=\left\langle\tanh\tilde{h}_i\left(\boldsymbol{\sigma}_{\setminus i}\right)\right\rangle_{p\left(\boldsymbol{\sigma}_{\setminus i}\right)} \tag{5.44}$$

$$\langle\sigma_i\sigma_j\rangle_{p\left(\sigma_i\mid\boldsymbol{\sigma}_{\setminus i}\right)p\left(\boldsymbol{\sigma}_{\setminus i}\right)}=\left\langle\sigma_j\tanh\tilde{h}_i\left(\boldsymbol{\sigma}_{\setminus i}\right)\right\rangle_{p\left(\boldsymbol{\sigma}_{\setminus i}\right)} \tag{5.45}$$

The above expressions are derived by explicitly performing the summation over σ_i, so that the right hand sides still have to be averaged with respect to the distribution of $\boldsymbol{\sigma}_{\setminus i}$, requiring 2^{N-1} operations. So far, (5.44)-(5.45) are exact relations, also known as Callen's identities [25]. The key approximation behind PL is to approximate the probability $p\left(\boldsymbol{\sigma}_{\setminus i}\right)$ in (5.42)-(5.44)-(5.45) with its empirical counterpart, denoted with $p_{\mathcal{D}}\left(\boldsymbol{\sigma}_{\setminus i}\right)$. In this way, for each spin we consider only the contribution of its conditional probability to the likelihood, and the rest is approximated by using the empirical dataset. This allows to define an approximate log-likelihood for spin i as:

$$\tilde{\mathcal{L}}_i\left(h_i,\boldsymbol{J}_{i*}\right)=\frac{1}{M}\log\prod_{\mu=1}^{M}p\left(\sigma_i^{\mu}\mid\boldsymbol{\sigma}^{\mu}_{\setminus i}\right)=\frac{1}{M}\sum_{\mu=1}^{M}\log p\left(\sigma_i^{\mu}\mid\boldsymbol{\sigma}^{\mu}_{\setminus i}\right) \tag{5.46}$$

Note that $\tilde{\mathcal{L}}_i$ is a function of N parameters, namely the external field h_i and the i-th row of the coupling matrix, $\boldsymbol{J}_{i*}=\{J_{ij}\}_{j\neq i}$. Performing derivatives of (5.46) w.r.t. these parameters one finds that the maximum of (5.46) satisfies the following conditions:

$$\langle \sigma_i \rangle_{\mathcal{D}} = \left\langle \tanh \tilde{h}_i \right\rangle_{\mathcal{D}} \tag{5.47}$$

$$\langle \sigma_i \sigma_j \rangle_{\mathcal{D}} = \left\langle \sigma_j \tanh \tilde{h}_i \right\rangle_{\mathcal{D}} \quad \forall j \neq i \tag{5.48}$$

that correspond to the Callen identities (5.44)-(5.45), where the average on both sides is here performed with respect to the data. Since the right hand sides of (5.47)-(5.48) both contain a non linear-function of all the degrees of freedom, high-order correlations are implicitly taken into account into the PL maximization, at a difference with the exact maximum likelihood procedure which only involves the first two moments. The (log)-*pseudolikelihood* is then defined as the sum of (5.46) over all spins:

$$\mathcal{L}^{\mathrm{PL}}(\boldsymbol{h}, \boldsymbol{J}) = \sum_{i=1}^{N} \tilde{\mathcal{L}}_i (h_i, \boldsymbol{J}_{i*}) \tag{5.49}$$

The overall computational complexity to compute $\left(\boldsymbol{h}^{PL}, \boldsymbol{J}^{PL}\right) = \arg\max_{\boldsymbol{J},\boldsymbol{h}} \mathcal{L}^{\mathrm{PL}}(\boldsymbol{h}, \boldsymbol{J})$ scales as $O\left(N^2 M\right)$ per iteration, and the maximization can be carried out using standard gradient-ascent algorithms, equivalently to (5.9)-(5.10). However, by construction the optimization algorithm can be split into N separate procedures, i.e. one for each spin i in which h_i and all its couplings J_{i*} are inferred: as a consequence, the optimization can be easily parallelized by maximizing each single-site pseudolikelihood $\mathcal{L}_i$ independently, thus reducing the computational cost by a factor N. The pseudolikelihood has the same maximum as the exact likelihood in the limit $M \to \infty$ [101]. As a final remark, note that the inferred coupling matrix is not symmetric by construction at finite M: the simplest solution, used also in the present work, is to symmetrize it to get a symmetric matrix, i.e. $J_{ij}^{PL} \leftarrow \frac{1}{2}\left(J_{ij}^{PL} + J_{ji}^{PL}\right)$. Alternatively, one might constraint the optimization to the subspace of symmetric matrices, but in this case a parallel computation would no longer be possible. Pseudolikelihood can also be implemented by using regularization terms on the model parameters, in order to penalize large values. However, in the present work we choose not to do so, in order to have a more fair comparison w.r.t. the other techniques. Further developments to Pseudolikelihood have been developed in order to design efficient decimation procedures to iteratively select only the most significant couplings [40]. It should be stressed that PL exploits in principle all the information contained in the dataset, including all the high-order correlations, whereas all the previous techniques employ only the first and second empirical moments.

5.4 Density Consistency solution

This section shows how to compute an approximate solution to the Inverse Ising Problem by using Density Consistency. As discussed in Section 5.2.2, the maximum likelihood point is found by matching the first two moments of the empirical statistics and the equilibrium expectation values computed at $\left(\boldsymbol{h}^{ML}, \boldsymbol{J}^{ML}\right)$. The simplest way to exploit DC is to replace the equilibrium expectation values at the right-hand sides of (5.11)-(5.12) with the DC estimates:

$$\langle \sigma_i \rangle_{DC} = \langle \sigma_i \rangle_{\mathcal{D}} \quad \forall i \tag{5.50}$$

$$\langle \sigma_i \sigma_j \rangle_{DC} = \langle \sigma_i \sigma_j \rangle_{\mathcal{D}} \quad \forall i \neq j \tag{5.51}$$

where $\langle \cdot \rangle_{DC}$ is computed by averaging over the tilted distributions introduced in Section 3.1.1. We now follow the same approach discussed in Section 4.1.1 for the Ising Model. For convenience, let us start by rewriting the expressions for the moments computed under the tilted distributions (already shown in Eqs. (4.14)):

$$\langle x_i \rangle_{q^{(ij)}} = \tanh\left[a_i^{(ij)} + \operatorname{atanh}\left(\tanh b^{(ij)} \tanh a_j^{(ij)}\right)\right] \tag{5.52a}$$

$$\langle x_j \rangle_{q^{(ij)}} = \tanh\left[a_j^{(ij)} + \operatorname{atanh}\left(\tanh b^{(ij)} \tanh a_i^{(ij)}\right)\right] \tag{5.52b}$$

$$\langle x_i x_j \rangle_{q^{(ij)}} = \tanh\left[b^{(ij)} + \operatorname{atanh}\left(\tanh a_i^{(ij)} \tanh a_j^{(ij)}\right)\right] \tag{5.52c}$$

where

$$a_i^{(ij)} = h_i^{(ij)} + w_i^{(ij)}; \qquad a_j^{(ij)} = h_j^{(ij)} + w_j^{(ij)}; \qquad b^{(ij)} = J_{ij} - S_{ij}^{(ij)} \tag{5.53}$$

In the above formulas, $h_i^{(ij)}$(resp. $h_j^{(ij)}$) represents a portion of the local field h_i (resp. h_j) associated to the pair (ij); conversely, $w_{i(j)}^{(ij)}$, $S_{ij}^{(ij)}$ are the fields and couplings coming from the Gaussian cavity distribution, given by:

$$w_i^{(ij)} = \frac{\Sigma_{jj}\mu_i - \Sigma_{ij}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \gamma_i^{(ij)} \tag{5.54}$$

$$w_j^{(ij)} = \frac{-\Sigma_{ij}\mu_i + \Sigma_{ii}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \gamma_j^{(ij)} \tag{5.55}$$

$$S_{ij}^{(ij)} = \frac{-\Sigma_{ij}}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \Gamma_{ij}^{(ij)} \tag{5.56}$$

where $\gamma_i^{(ij)}, \gamma_j^{(ij)}$, $\Gamma_{ij}^{(ij)}$are the parameters of the approximating Gaussian factor ϕ_{ij} in Eq. (4.5). Eqs. (5.52) are essentially equal to (4.14), with the only difference that now tilted distributions are defined for each pair of spins $i < j$ since the topology is unknown, and β has been adsorbed into the model parameters. At this stage, the Gaussian moments $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are not yet specified, and they will be fixed by choosing a suitable set of closure conditions. A first consequence of using Density Consistency on all the spin pairs is that all the entries of $\boldsymbol{\Sigma}$ can be determined; in this way, by using the definition (5.45) we can replace $\Gamma_{ij}^{(ij)}$ with $\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}$. Using Eqs. (5.52), the approximate maximum likelihood point obtained by using DC will satisfy the following matching conditions for any pair of spins $i < j$:

$$\langle \sigma_i \rangle_{q^{(ij)}} = m_i \tag{5.57a}$$

$$\langle \sigma_j \rangle_{q^{(ij)}} = m_j \tag{5.57b}$$

$$\langle \sigma_i \sigma_j \rangle_{q^{(ij)}} = C_{ij} + m_i m_j \tag{5.57c}$$

to be solved w.r.t. $\left\{\left(h_i^{(ij)}, h_j^{(ij)}, J_{ij}\right)\right\}_{i<j}$. In principle, one could use a gradient-ascent algorithm similar to the Boltzmann learning procedure to iteratively update the fields and couplings; however, in this case Density Consistency needs to be run at each iteration in order to estimate the left-hand sides of (5.57). Such iterative scheme turns out to be unnecessary, since it is possible to invert (5.57) analytically. First notice that Eqs. (5.57) can be solved using (5.52) w.r.t. $\left\{\left(a_i^{(ij)}, a_j^{(ij)}, b_{ij}\right)\right\}_{i<j}$; by construction, their expression coincide to the Independent Pair estimates discussed in Section 5.3.3:

$$a_i^{(ij)} = h_i^{(ij)IP}; \qquad a_j^{(ij)} = h_j^{(ij)IP}; \qquad b_{ij} = J_{ij}^{IP} \tag{5.58}$$

Then, by exploiting the definitions of $a_i^{(ij)}, a_j^{(ij)}, b_{ij}$ we can write a closed-form solution for $\left(h_i^{(ij)}, h_j^{(ij)}, J_{ij}\right)$:

$$h_i^{(ij)*} = h_i^{(ij)IP} - w_i^{(ij)} = h_i^{(ij)IP} - \frac{\Sigma_{jj}\mu_i - \Sigma_{ij}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} + \gamma_i^{(ij)} \tag{5.59}$$

$$h_j^{(ij)*} = h_j^{(ij)IP} - w_j^{(ij)} = h_j^{(ij)IP} - \frac{-\Sigma_{ij}\mu_i + \Sigma_{ii}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} + \gamma_j^{(ij)} \tag{5.60}$$

$$J_{ij}^{*} = J_{ij}^{IP} + S_{ij}^{(ij)} = J_{ij}^{IP} - \frac{\Sigma_{ij}}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} - \left(\mathbf{\Sigma}^{-1}\right)_{ij} \tag{5.61}$$

At this stage, all the couplings can be inferred using (5.61), after fixing a suitable set of closure conditions. Conversely, the local external field on spin i will be computed by summing all the contributions $h_i^{(ij)*}$ for $j \neq i$:

$$\begin{aligned} h_i^{*} &= \sum_{j\neq i} h_i^{(ij)} = \sum_{j\neq i}\left(h_i^{(ij)IP} - \frac{\Sigma_{jj}\mu_i - \Sigma_{ij}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2}\right) + \sum_{j\neq i}\gamma_i^{(ij)} && (5.62)\\ &= \sum_{j\neq i} h_i^{(ij)IP} - \sum_{j\neq i}\frac{\Sigma_{jj}\mu_i - \Sigma_{ij}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} + \left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right)_i && (5.63)\\ &= h_i^{IP} + (N-2)\,\mathrm{atanh} m_i - \sum_{j\neq i}\frac{\Sigma_{jj}\mu_i - \Sigma_{ij}\mu_j}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} + \left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right)_i && (5.64) \end{aligned}$$

where the Indepedent Pair expression (5.31) is used in the last line; in the second line, we used $\left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right)_i = \sum_{j\neq i}\gamma_i^{(ij)}$ as in Eq. (3.11).

In order to get an explicit expression in terms of first and second empirical moments $(\boldsymbol{m}, \boldsymbol{C})$, it is necessary to fix a set of closure equations. In the following, we exploit both the DC closure (3.34) and the moment matching (or EP) closure (3.55), for reasons that will be clear in the next section. Combining the closure equations and the maximum likelihood matching conditions (5.57) we get the following expressions for the Gaussian moments $(\boldsymbol{\mu}, \mathbf{\Sigma})$ w.r.t. the empirical ones:

$$\begin{aligned} \mu_i^{EP} &= m_i \quad \forall i, && (5.65a)\\ \Sigma_{ii}^{EP} &= 1 - m_i^2 \quad \forall i && (5.65b)\\ \Sigma_{ij}^{EP} &= C_{ij} \quad \forall i \neq j. && (5.65c) \end{aligned}$$

$$\begin{aligned} \mu_i^{DC} &= m_i \quad \forall i, && (5.66a)\\ \Sigma_{ii}^{DC} &= \frac{m_i}{\mathrm{atanh} m_i} \quad \forall i && (5.66b)\\ \Sigma_{ij}^{DC} &= C_{ij}\sqrt{\frac{\Sigma_{ii}^{DC}}{C_{ii}}\frac{\Sigma_{jj}^{DC}}{C_{jj}}} \quad \forall i \neq j && (5.66c) \end{aligned}$$

respectively for the EP and DC closures. Note that, despite Σ_{ii}^{DC} is not well defined for $m_i = 0$, it has a finite limit $\lim_{m_i \to 0}\Sigma_{ii}^{DC} = 1$. Finally, by inserting the above formulas into (5.61)-(5.64) we get the final expression for the inferred parameters under DC or EP approximations, shown in the next paragraphs.

EP (moment matching) closure

$$J_{ij}^{EP} = J_{ij}^{IP} - \left(\boldsymbol{C}^{-1}\right)_{ij} - \frac{C_{ij}}{C_{ii}C_{jj} - C_{ij}^2} \qquad \forall i \neq j \tag{5.67}$$

$$h_i^{EP} = h_i^{IP} + (N-2)\,\text{atanh} m_i - \sum_{j\neq i} \frac{C_{jj}m_i - C_{ij}m_j}{C_{ii}C_{jj} - C_{ij}^2} + \left(\boldsymbol{C}^{-1}\boldsymbol{m}\right)_i \qquad \forall i \tag{5.68}$$

where $C_{ii} = 1 - m_i^2$ is the variance of spin i.

DC closure

$$J_{ij}^{DC} = J_{ij}^{IP} - \left[\left(\boldsymbol{\Sigma}^{DC}\right)^{-1}\right]_{ij} - \frac{C_{ij}}{C_{ii}C_{jj} - C_{ij}^2}\sqrt{\frac{C_{ii}}{\Sigma_{ii}^{DC}}\frac{C_{jj}}{\Sigma_{jj}^{DC}}} \qquad \forall i \neq j \tag{5.69}$$

$$h_i^{DC} = h_i^{IP} + (N-2)\,\text{atanh} m_i - \sum_{j\neq i} \frac{\Sigma_{jj}^{DC}m_i - \Sigma_{ij}^{DC}m_j}{\Sigma_{ii}^{DC}\Sigma_{jj}^{DC} - \left(\Sigma_{ij}^{DC}\right)^2} + \left[\left(\boldsymbol{\Sigma}^{DC}\right)^{-1}\boldsymbol{m}\right]_i \qquad \forall i \tag{5.70}$$

where Σ_{ii}^{DC},Σ_{ij}^{DC} are given by Eqs. 5.66. The above formulas allow to reconstruct the model parameters in terms of the empirical magnetizations and correlations, by requiring a single matrix inversion ($\boldsymbol{C}$ or $\boldsymbol{\Sigma}^{DC}$, depending on the closure used), in the same way as for all the other mean-field like methods previously discussed.
It is important to notice that, independently on the closure, the first term equal to the Independent Pair estimate does not depend on the closure used. This is a contribution coming from the direct link (ij) and the (approximate) moment matching condition to the empirical moments (5.57), while all the others depend on the cavity distribution and explicitly depend on the chosen closure. Another consequence is that, neglecting cavity correlations in the tilted distributions (i.e. asssuming that the Gaussian covariance matrix $\boldsymbol{\Sigma}$ is diagonal), the inferred couplings J_{ij}^* coincide to the Independent Pair estimates (5.28), independently on the closures. For what concerns the fields, this is true only by selecting a set of closures satisfiying DC condition (3.28). Indeed, starting from (5.64) and setting $\Sigma_{ij} = \delta_{ij}\Sigma_{ii}$, we get:

$$h_i^* = h_i^{IP} + (N-2)\,\text{atanh} m_i - (N-2)\frac{m_i}{\Sigma_{ii}} \tag{5.71}$$

The above expression is equal to h_i^{IP} only if $m_i/\Sigma_{ii} = \text{atanh} m_i$, which is precisely the DC condition (3.28). Remembering that Density Consistency coincides with the Bethe Approximation when neglecting cavity correlations, we conclude that the above reasoning is actually equivalent to solve the maximum likelihood equations by using the Bethe Approximation, so that the corresponding inferred parameters coincide with the Independent Pair approximation, as we could expect.

5.4.1 Relation to Sessak-Monasson approximation

Surprisingly, the expression of the couplings obtained by using EP closure (5.67) is equal to the Sessak-Monasson expression (5.41). This analogy suggests that also the expression h_i^{EP} should be similar to what we would get by summing the same loop contributions as in (5.41): indeed, we found that h_i^{EP} (5.68) gives the exact series expansion presented in [124] at least up to order β^3, apart from a 0-th order term in the magnetization. Therefore, we can state that the expression of inferred fields under the Sessak-Monasson (SM) approximation, obtained by summing all the loop and 2-spin diagrams, can be re-phrased as:

$$h_i^{SM} = h_i^{EP} + (N-2)\left[\frac{m_i}{1-m_i^2} - \text{atanh} m_i\right] + O\left(\beta^4\right), \tag{5.72}$$

where h_i^{EP} is defined in (5.68). The second term in (5.72) is present simply because DC with plain moment matching (EP closure) is not exact even for a non-interacting model. Since additional loop contributions not summed up in (5.41) appear at higher orders in β, further investigation would be needed to verify the correctness of (5.72) at higher orders in the series expansion, and we leave this point for future works.
The DC closure provides an expression similar in structure to the Sessak-Monassson approximation: expanding (5.69) for small correlations leads to the same loop expansion presented in the previous section, but the coefficients multiplying each term (that depend only on the magnetizations m_i) differ from (5.39) because of the different closure used. This analogy suggests that Density Consistency might be closely related to the Sessak-Monasson approximation, to be investigated in future works. Finally, notice that in the limit of zero magnetizations, $\mathbf{\Sigma}^{DC} \to \boldsymbol{C}$ and the expression of the inferred couplings tends to the Sessak-Monasson approximation, namely $J_{ij}^{DC} \to J_{ij}^{EP} = J_{ij}^{SM}$. Therefore, in order to highlight different reconstruction performances all the simulations presented in the next section will be performed in regimes where the empirical magnetizations differ from 0.

5.5 Results

In this final section we provide an extensive comparison between all the inference methods discussed so far on synthetic generated data. All the simulations are performed by generating configurations from a known model, defined on a certain graph topology with arbitrary distributions of fields and couplings; in addition, we will modify the inverse temperature β to tune their strength.
In section 5.5.1 we first compare the DC performances with the EP closure and the Sessak-Monasson approximations, in order to establish under which conditions DC improves the other two (related) approaches. This comparison is carried out on a fully connected spin glass (Sherrington-Kirkpatrick model [126]) with $N = 20$ nodes, so that an exact computation of the equilibrium observables can be performed in a reasonable time.
A second comparison among all the mean-field like methods is carried out in Section 5.5.2 on sparse topologies and small system sizes, where again the exact trace over all the configurations is feasible. Finally, in Section 5.5.3 we evaluate the performances of all the methods on larger systems, where samples are collected by using Monte Carlo Gibbs sampling [55]; this time, the comparison includes also Pseudo-Likelihood, which is widely considered as the out-performer for the IIP [101]. In all the cases illustrated hereafter, the inference quality will be measured in terms of the reconstruction errors of the inferred couplings and fields:

$$\Delta_J = \sqrt{\frac{\sum_{i<j}\left(J_{ij}^t - J_{ij}\right)^2}{\sum_{i<j}\left(\beta J_{ij}^t\right)^2}}, \qquad \Delta_h = \sqrt{\frac{\sum_i \left(h_i^t - h_i\right)^2}{\sum_i \left(\beta h_i^t\right)^2}}, \tag{5.73}$$

where $\left(h_i^t, J_{ij}^t\right)$ are the true model parameters. For each scenario, we run n different instances by varying the seed used to generate the topology and/or the model parameters. As a consequence, all the following plots show the average and standard error for both measures (5.73):

$$\bar{f} = \frac{1}{n}\sum_{\alpha=1}^{n} f^\alpha \qquad \delta f = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{\alpha=1}^{n}\left(f^\alpha - \bar{f}\right)^2} \tag{5.74}$$

where $f \in \{\Delta_J, \Delta_h\}$ and the summation runs over the different instances. From a technical point of view, a small diagonal regularization $\varepsilon = 10^{-10}$ will be added to the empirical/exact covariance matrix $\boldsymbol{C}$ to prevent numerical issues arising when inverting it, mainly relevant at low temperatures.

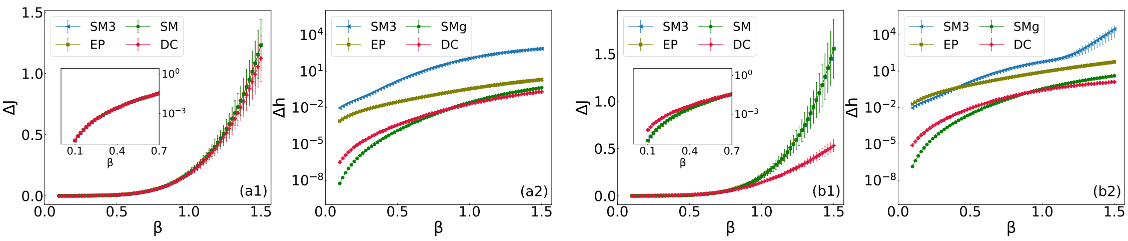

Figure 5.1: Reconstruction quality on a fully connected graph of $N = 20$ nodes with binary couplings $J_{ij} \sim \pm\beta/\sqrt{N}$ and uniform fields $h_i \sim h_0 U(0, \beta)$. Panel a, left: $h_0 = 0.1$; (b): $h_0 = 0.5$. For each scenario, we show the error w.r.t the couplings Δ_J and the fields Δ_h, averaged on 30 instances. The insets show the log-scale behaviour at high temperatures (low β).

5.5.1 Comparison to Sessak-Monasson approximation

We start from a comparison between SM, EP and DC on a fully connected frustrated model in the presence of random external fields. As discussed in the previous section, the EP and SM inference of couplings coincides and it is given by Eq. (5.67); for what concerns the fields' reconstruction, we compare the EP and DC expressions (resp. Eq. (5.68) and Eq. (5.70)) to the small correlation expansion up to 3−rd order given by Eq. (5.40) (labelled as SM3) and the closed-form guess given by Eq. (5.72) (labelled as SMg). We show in Figure 5.1 the reconstruction performances for a fully connected graph of $N = 20$ spins, with a binary distribution of couplings, namely $J_{ij} \sim \pm\beta/\sqrt{N}$; conversely, fields are uniformly distributed, i.e. $h_i \sim \beta h_0 U(0,1)$ with a certain scale h_0. Figure 5.1 shows the reconstruction error on couplings and fields for two values of the scale, respectively $h_0 = 0.1$ (left) and $h_0 = 0.5$ (right), averaged over $n = 30$ instances. At the smaller value of $h_0 = 0.1$, DC and SM have a similar behaviour on Δ_J, since the magnetizations are relatively small and the DC estimate tends to SM, as previously discussed. At larger values of h_0 (right panel), DC reconstruction improves the estimate of the couplings, if compared to SM, and the gap between the two increases as the temperature is lowered. Looking at the fields' reconstruction, it is evident that the small correlation expansion truncated to the 3rd order and the EP estimate have poor performances. On the other hand, the SM reconstruction given by our guess (5.72) gives a good estimate at high temperatures, while being out-performed by DC at lower temperatures, as it happens for the couplings. The same qualitative behaviour can be observed by choosing other combination for the fields/couplings' distributions (e.g. Gaussian/binary, Gaussian/Gaussian, etc.).

5.5.2 Reconstruction using exact statistics

In this section we show results on different sparse topologies, i.e. graphs with low average connectivity, by comparing DC reconstruction to the other mean-field like methods for the IIP: model parameters are reconstructed through the Independent Pair (IP) approximation (5.28)-(5.31), Mean-Field (MF) (5.19)-(5.20), TAP equations (5.23)-(5.24), Susceptibility Propagation (SP) (5.37)-(5.38), in addition to DC and SM. From now on, Eq. (5.72) is used to reconstruct the external fields under the SM approximation, since the 3−rd order truncated series typically gives too large errors.
We performed simulations for several combinations of the couplings and fields distributions (Gaussian $\mathcal{N}(0,1)$, uniform positive $U(0,1)$, binary ±1 and constant) and graph topology: we used

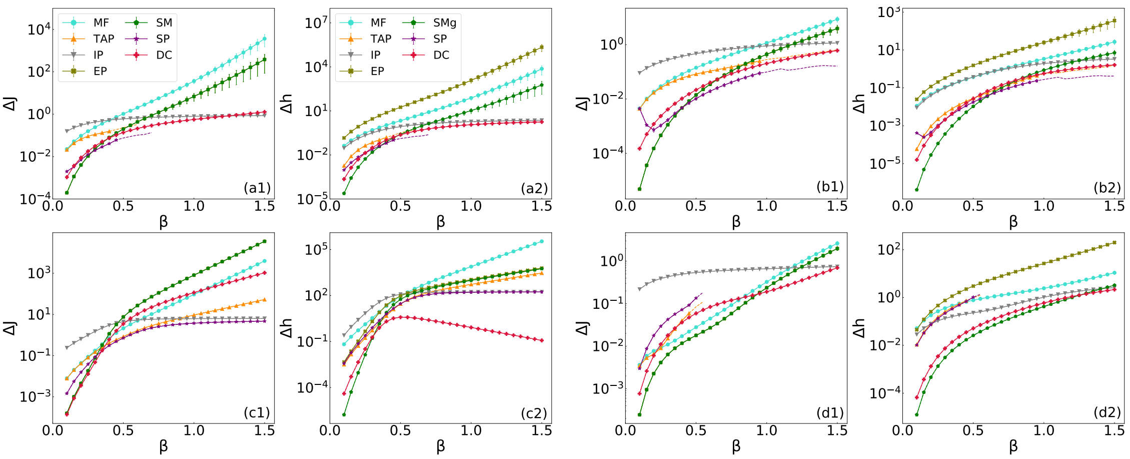

Figure 5.2: Reconstruction in sparse topologies from exact observables. (a) Erdős-Reny graph with $N = 20$ and mean connectivity $k = 3$, Gaussian couplings $J_{ij} \sim N\left(0, \beta^2\right)$, constant fields $h_i = \beta$. (b) $2d$ square lattice with periodic boundary conditions of $N = 4^2$ nodes and dilution coefficient $p = 0.7$, uniform positive couplings $J_{ij} \sim U\left(0, \beta\right)$, binary fields $h_i \sim \pm 0.5\beta$. (c) Random regular graph with $N = 20$ and fixed connectivity $k = 4$, constant couplings $J_{ij} = \beta$, constant fields $h_i = 0.1\beta$. (d) Triangular lattice with $N = 4^2$ anti-ferromagnetic couplings $J_{ij} = -\beta$, binary fields $h_i \sim \pm 0.6\beta$. All the plots show respectively the error over couplings and over fields, averaged on $n = 50$ instances. SP and TAP reconstructions are shown as dashed lines when the number of instance giving physical solution n^* is $n/2 \leq n^* < n$, and it is not shown for $n^* < n/2$.

Erdős-ény and random regular graphs with different mean (resp. fixed) connectivity, regular lattices with a diluition coefficient $p \in (0{,}1]$ [3].

A selected subset of results is shown in Figure 5.2. Each scenario shows the reconstruction errors averaged over $n = 50$ instances. We can identify a general behaviour of the different methods: SM is always better than MF but both of them give large errors at low temperatures. TAP and SP outperform MF but suffer from numerical problems when the fields are large and the temperature is low (see for instance the results for the Erdos-Rényi and graph the triangular lattice in Figure 5.2 (a) and (d) respectively). In thes regimes, TAP and SP equations have no fixed points as the arguments of the square roots appearing in the expressions of the couplings become negative. In particular, in Figure 5.2 we separate the regimes in which TAP and SP provide physical solutions on all the n instances (by using both dots and lines), from those in which at least one time they do not find a solution (here we plot lines, without dots). No result is shown when these methods fail to provide a physical solution on more than half of the n instances. DC turns out to significantly outperform SM in almost all cases at low temperatures and it provides comparable estimates to SP at small β. A different behaviour can be noted in Figure 5.2 (c), where DC performs worse than other methods for the couplings reconstruction, but it gives a very good estimation for the fields.

[3] At $p \in (0{,}1)$, only a random fraction p of the links of the full structure is considered

5.5.3 Inference using sampled configurations

Finally, we performed another set of simulations on large system sizes where a set of M equilibrium observables are generated through Monte Carlo Gibbs sampling [55]. In this case, we compare all the previous methods also against Pseudolikelihood. The MCMC dynamics starts from a random configuration and it is let to equilibrate for an initial M steps. Then, the same dynamics is run for other Md steps, and a total M of samples is collected (1 every d). In this way, in principle we slightly reduce the effect of the autocorrelation time in the Monte Carlo dynamics, that is relevant especially at low temperatures. We remark that each MC "step" here corresponds to N sequential Gibbs-sampling sweeps, one for each spin performed on a random permutation of the indices. In all the simulations, $M = 10^5$, $d = 10^2$. Pseudo-likelihood maximization is performed up to a numerical precision of 10^{-4} with no regularization term, using the implementation available at [105].

Pseudo-count The estimation of the first and the second empirical moments, used by all the mean-field like methods, is slightly modified by the addition of a small pseudo-count $\lambda \in (0{,}1)$ [45]. In general, the effect of the pseudo-count is to modify the distribution of a set of N binary variables (in this case, the empirical distribution) as a mixture between the starting one and a uniform density in the range $\{-1{,}1\}^N$. The new empirical density, denoted with $p_{\mathcal{D}}^{(\lambda)}$ is given by:

$$p_{\mathcal{D}}^{(\lambda)}(\boldsymbol{\sigma}) = (1-\lambda)\, p_{\mathcal{D}}(\boldsymbol{\sigma}) + \frac{\lambda}{2^N} \tag{5.75}$$

As a consequence, adding a pseudo-count $\lambda \in (0{,}1)$ modifies the expectation values as:

$$\langle \sigma_i \rangle_{\mathcal{D}}^{(\lambda)} = (1-\lambda)\, \langle \sigma_i \rangle_{\mathcal{D}} \quad \forall i \tag{5.76}$$

$$\langle \sigma_i \sigma_j \rangle_{\mathcal{D}}^{(\lambda)} = (1-\lambda)\, \langle \sigma_i \sigma_j \rangle_{\mathcal{D}} \quad \forall i \neq j \tag{5.77}$$

and the same holds for higher-order (non-connected) moments. The strength of the pseudocount is chosen in such a way that its effect becomes negligible as when the number of samples goes to infinity. We chose $\lambda = 1/M$ as a reasonably good value. It is important to remark that most of the methods are slightly affected by the addition of pseudocount: in particular, only MF gives significantly worse performances when turning off λ, as noted in [10]. We show in Figure 5.3 the reconstruction error of the couplings and the fields for all the methods on different graph topologies and different model parameters (see the details in the caption of Figure 5.3). The behaviour is qualitatively similar to the results shown in Figure 5.2. In particular, SM/EP and MF typically have a very large reconstruction error at low temperatures; on the other hand, SP and TAP fail to find physical solutions for the couplings and fields at low temperatures, as previously discussed. Although PL seems to outperform the other methods in all the regimes considered here, DC gives comparable performances to PL for small β and it often provides the best estimates among all the methods that, similarly to DC, use only the information about first and second moments.

Effect of the number of samples on the inference quality

Finally, we test the robustness of the inference methods when the number of samples is lowered: in the previous section we used $M = 10^5$ for all the simulations, which in many applications is a too optimistic value for the number of available data. The sampling is carried out with the same setting as in the previous section; then, we select only the first subset of these samples for different values of $\mathcal{M} < M$; in this way, the final dataset will be characterized by the same equilibration and de-correlation time as before.

In this setting, another measure used to compare the reconstruction performances is the area under the Receiver Operating Characteristic (ROC) curve, denoted with AUC. The latter quantifies how

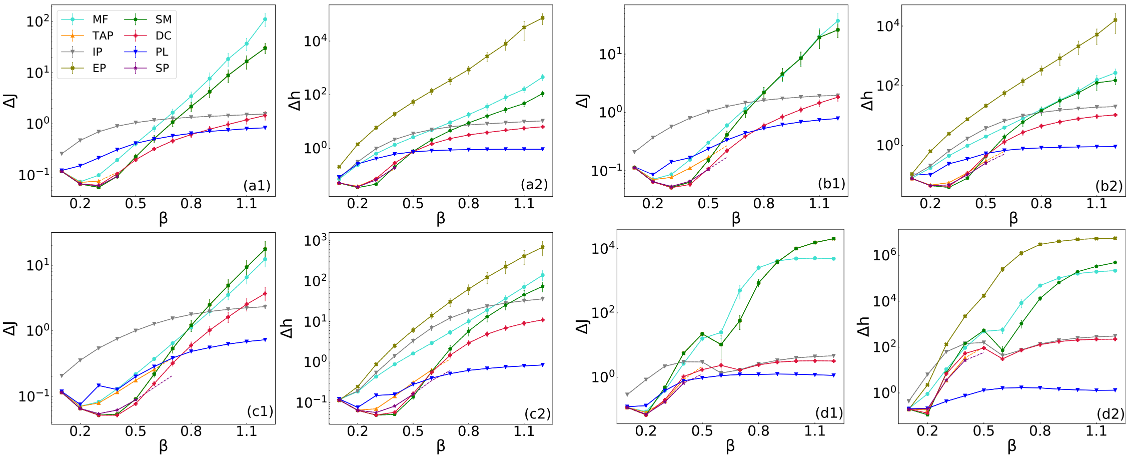

Figure 5.3: Reconstruction in sparse topologies using sample averages. (a) Barabasi-Albert [8] graph with $N = 50$, $n_0 = k = 2$, binary couplings $J_{ij} \sim \pm\beta$ and constant fields $h_i = 0.7\beta$. (b) $2d$ square lattice with $N = 7^2$ and PBC, binary couplings $J_{ij} \sim \pm\beta$ and constant fields $h_i = 0.5\beta$. (c) Random regular graph with $N = 50$ and fixed connectivity $k = 4$, binary couplings $J_{ij} \sim \pm\beta$, binary fields $h_i \sim \pm 0.3\beta$. (d) Erdos-Reny graph with $N = 50$ and mean connectivity $k = 4$, constant couplings $J_{ij} = \beta$, uniform fields $h_i \sim 0.3U(0,\beta)$. All the plots show respectively the (log-scale) error over couplings and over fields, averaged on 20 instances.

good the detection of present/absent couplings is on a sparse topology, independently on their strength [48]. Figures 5.4-5.5 show the results on the same instance of Figure 5.3, panels (a) and (b) respectively, for $\mathcal{M} \in \left[10^2, 10^5\right]$. We also report as a vertical and dashed line at $M^* = \binom{N}{2}$, that corresponds to the number of couplings to be inferred. This value can be consider as a natural threshold which separates a harder (resp. easier) regime in which the number of configuration used to compute the data statistics is smaller (resp. higher) than the number of unknowns. In the large temperature regime, all methods show comparable performances with the exception of IP which provides the best estimates, in terms of both Δ_J and AUC when the statistics is extremely poor, that is for $\mathcal{M} = 10^2$. Then the accuracy of the predictions deteriorates for large $\mathcal{M}$ values. Here, PL looses its predictive power only for $\beta = 0.4$ and $\mathcal{M} < M^*$. In the large β regime, the fixed-statistics methods outperform PL in a wide range of $\mathcal{M}$ values and, among them, DC seems to be preferred as suggested by both metrics, Δ_J and AUC. Conversely, when the amount of data is large, PL performances are the most accurate and the error decreases with increasing $\mathcal{M}$: this is reasonable since PL is the only consistent approach, being exact in the limit $\mathcal{M} \to \infty$. At extreme small values of $\mathcal{M} < M^*$ the Independent Pair approximation has the best performance if compared to the other fixed-statistics inference methods; this is reasonable since, unlike the other methods, it requires no matrix inversion. In a large and intermediate range of $\mathcal{M}$, even for $\mathcal{M} < M^*$, and, especially for large β, DC seems to be preferred among all the methods.

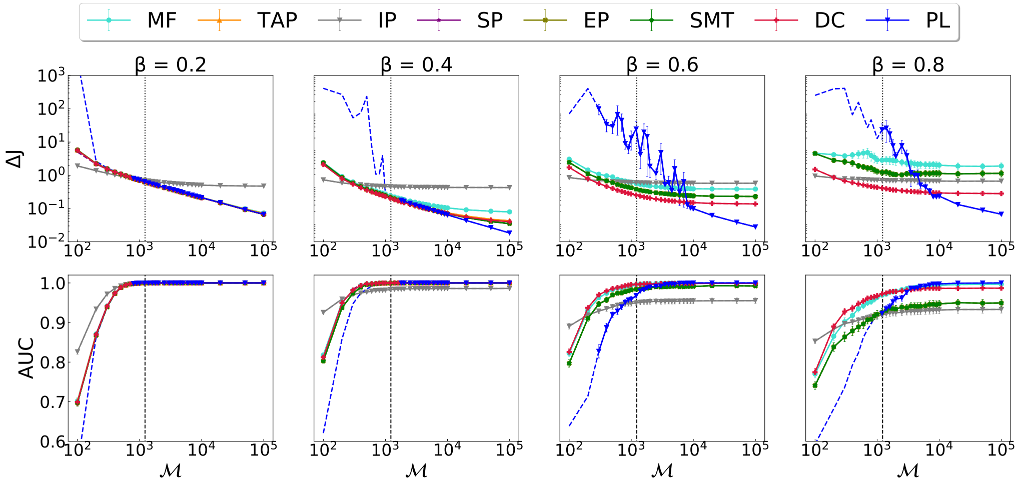

Figure 5.4: Effect of the number of samples on the reconstruction quality. The model is a Barabasi-Albert [8] graph with $N = 50$, $n_0 = k = 2$, binary couplings $J_{ij} \sim \pm\beta$ and constant fields $h_i = 0.7\beta$, i.e the same regime of Figure 5.3 (a). Each plot shows the average error over the couplings Δ_J (top panel) and the AUC (bottom panel) w.r.t. the number of samples $\mathcal{M}$ for different temperatures (values are shown at the top of each subplot). The black dotted line corresponds to $M^* = \binom{N}{2}$, i.e. the number of couplings to be inferred

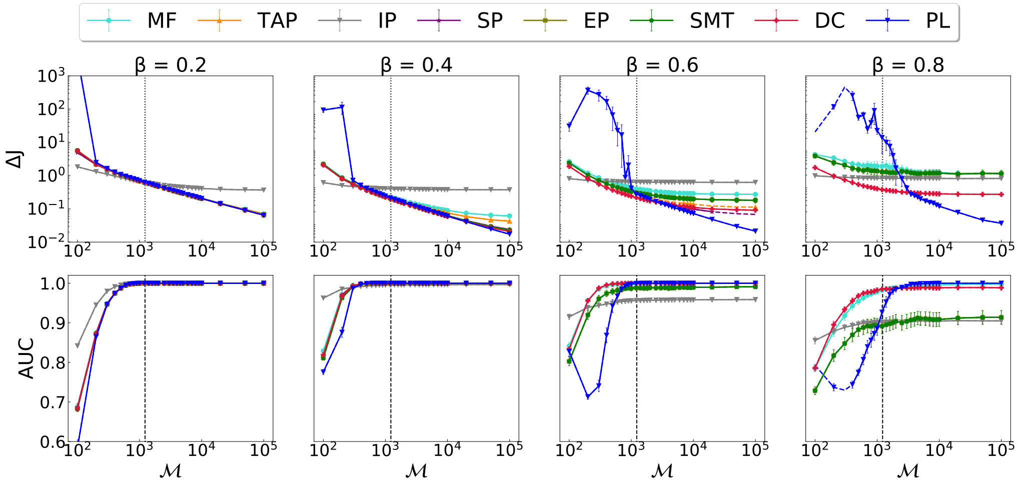

Figure 5.5: Effect of the number of samples on the reconstruction quality. The model is a $2d$ square lattice with $N = 7^2$ and PBC, binary couplings $J_{ij} \sim \pm\beta$ and constant fields $h_i = 0.5\beta$., i.e the same regime of Figure 5.3 (b). Each plot shows the average error over the couplings Δ_J (top panel) and the AUC (bottom panel) w.r.t. the number of samples $\mathcal{M}$ for different temperatures (values are shown at the top of each subplot). The black dotted line corresponds to $M^* = \binom{N}{2}$, i.e. the number of couplings to be inferred

Chapter 6

Conclusions and future perspectives

The first part of this manuscript presents a new class of iterative schemes to compute marginal distributions on discrete graphical models, called Density Consistency (DC). A general derivation, yet restricted to binary degrees of freedom, is carried out in chapter 3. The derivation follows the same reasoing used for Expectation Propagation (EP) in section 2.4: in this perspective, DC can be considered as a generalization of the EP scheme where the Gaussian approximation is used onto each factor node of the original model. This parametrization allows to include approximate loop corrections into the marginal probabilities, coming from all the cycles present in the graph. DC is constructed in such a way to give the exact marginals on trees, so to be an extension of the Bethe approximation: in particular, this is realized by imposing a *consistency* between the *density* values of single-node marginal distributions, that gives the name to the method. The connection with the Bethe approximation is proven rigorously in Section 3.2.1 on acyclic graphs, thanks to the DC condition (3.28).
As opposite to other approaches (e.g. the Cluster Variational method) that include larger regions of the graph while still assuming a factorized probability on these, DC relaxes the factorization assumption of the cavity distribution in the Bethe ansatz by including pairwise effective interactions, that are computed from a unique multivariate Gaussian distribution: by construction, this assumption approximately takes into account the effect of loops of any length, whose contribution is easily included into the marginal tilted distributions. In particular, the Gaussian ansatz for the cavity distribution takes into account effective pairwise interactions for each pair of neighbours spin, that corrects the "direct" link's contribution in the original factor graph. The correlations induced by each Gaussian cavity are easily incorporated in a self-consistent way, in an iterative scheme that scales polynomially (in particular, as the 3rd power) with respect to the number of variables.
Thanks to this correction, DC turns out to give significant improvements to the BP marginal estimates, as discussed in Chapter 4.1 for the forward problem, and it gives comparable performances with other loop corrected schemes. In particular, DC provides very good estimates of the equilibrium observables especially on models with long-range correlations (e.g. with ferromagnetic couplings) on several architectures.
The strength of the loop contributions can be tuned by decreasing the value of the interpolation parameter introduced in Section 3.2.2, that can improve convergence in all those regimes where a "full" DC solution, obtained by matching the Pearson correlation coefficient, cannot be found. In this perspective, neglecting at all cavity correlations by choosing a null value of the interpolation parameter gives the exact fixed points of Belief Propagation on any graph topology, as proven in

Section 3.2.1.

DC can be also exploited to estimate the equilibrium behaviour of homogenous models in the thermodynamic limit: in this regard, the analytic solution carried out on the ferromagnetic Ising model shows how the critical temperature is estimated significantly better than other mean-field like approximations. Despite the nature of the ferromagnetic phase is not clear yet, due to unstable branches and phase cohexistence observed in 3 dimensions, the critical temperature estimation is a signal that the loop corrections induced by the Gaussian cavities can better describe the long-range correlations arising at the critical point: this feature is further confirmed by the high-dimensional expansion 4.3.6, that predicts exactly two additional perturbative orders with respect to other mean-field like approaches.

In general, we expect DC to perform worse on factor graph models with dominant multi-body interaction terms (e.g. $p-$spin models), especially in presence of many short loops (this happens for instance in k-SAT models with a high density of clauses). A first reason is that the approximating family of distributions is Gaussian: therefore, the cavity distribution will carry explicitly information just about the first two moments, with higher-order correlations being dependent on them in a trivial way. Moreover, on these models, if the graph contains short loops it might happen that two nodes are connected to more than one factor node: if this scenario, we believe that DC condition needs to be generalized in such a way to require consistency between two-nodes marginals. By construction, this issue does not arise in pairwise graphical models, but a similar argument is discussed in 4.1.2, where a first attempt to include larger regions of the graph (plaquettes) is discussed for the square lattice Ising model: in this case, DC currently works only if the regions are chosen in such a way to have only single-nodes intersections.

In this perspective, DC can be in principle extended to include larger regions of the graphs explicitly (along the lines described in in Section 2.3.1 for the CVM): however, the interesection of graph regions on more than 1 variable node suggests that DC condition should be generalized in such a way to guarantee consistency between 2(or more)$-$body marginals.

In pairwise (Ising-like) models in sparse structured graphs with mean connectivity ≤ 4 DC breaks down slightly before the transition temperature: indeed, on such systems (like square lattice Ising model) the Gaussian covariance matrix becomes non-invertible close to the transition point, and at thermodynamic limit the inverse matrix elements diverge: as discussed in Section 4.3.4, this phenomenon can be physically interpreted thanks to the analogy with lattice random walks, where at low connectivity the latter becomes recurrent in the infinite dimensional limit.

In chapter 5 DC is exploited to provide an approximate closed-form solution to the Inverse Ising problem, where the maximum likelihood estimator is found by using the marginal tilted estimations; the resulting expression shares many similarities to the Sessak-Monasson approximation [124], depending on the set of closures used to fix DC's parameters; furthermore, simulations on synthetic data show how DC provides significant improvenements to the Sessak-Monasson approximation in the presence of random external fields on several topologies, and comparable performances with other methods (even Pseudolikelihood), especially in the presence of small amounts of data.

The connection to the Sessak-Monasson approximation suggests a more quantitative understading of the loop contributions encoded by DC which, however, is still to be clarified: future directions might exploit existing literature on the topic, e.g. the work by Georges and Yedidia [56], or the loop calculus by Chertkov [31], as a way to go. The comparison with the Sessak-Monasson expansion suggests that DC could be interpreted as a non-perturbative approximation. This might explain the lack of convergence at low temperatures, mainly due to the covariance matrix becoming singular and thus non-invertible due to high long-range correlations.

The generalization to non-binary degrees of freedom discussed in Section 3.4.1 opens several interesting directions to be investigated: for instance, similar analytic calculations for the ferromagnetic Ising model could be carried out on other discrete models like Blume-Emery-Griffiths

[16]. With regard to inference problem, the extension to Potts-like models can be in principle used to solve the *inverse* problem, by generalizing the procedure used for the Ising variables in Chapter 5; this direction can provide further developments with interesting applications in computational biology, in the context of protein structure inference from co-evolutionary sequences [36].

As an overall final remark, the freedom in selecting the set of closure equations might allow for further developments with better performances and/or convergence properties. In this regard, the preliminar variational formulation presented in Section 3.4.2 might be helpful in understanding which set of closure equations should be used, that up to now is chosen heuristically.

With regard to spin glass systems with quenched disorder, another potential future direction regards the possibility to extend DC at the *ensemble* level, analogously to the cavity method.

Part II

Bayesian inference approaches for epidemic mitigation

Chapter 7

Bayesian inference approaches to epidemic mitigation via digital contact tracing

7.1 Introduction and motivation

The outbreak of SARS-COV-2 at the beginning of 2020 turned into a world-wide pandemic that spread rapidly across all countries and, at the time this manuscript is being written, more than 100 million cases have been confirmed. One of the major issues with this virus is the possibility to be infected without manifesting any symptoms, so that the detection of infected indivuals become harder. One of the many tools national health-care systems and public authorities currently employ to mitigate the epidemic spreading is the trace-test-isolate strategy: given an individual who has just been tested positive, her/his recent contacts can be identified, tested and isolated to prevent further transmission events. However, this strategy turns out to be efficient only at the very early stage of the epidemic, when the number of new infected people is small enough to be managed "manually" by public health infrastructures; on the contrary, it cannot be applied when the epidemic starts to grow, mainly because the average number of contacts of a typical individual has before he/she is tested positive can be large and not all her/his recent contacts might be known to the individual; moreover, public health infrastructures and personnel are tipycally under-dimensioned to perform manual tracing in this regime.

For these reasons, digital contact tracing provides an alternative way to mitigate the epidemic spreading, and it has been deployed in some countries at the early stage of the outbreak [11]. Digital contact tracing exploits mobile-phone technology to automatically register proximity events between individuals, for instance via Bluetooth, so to have access to the time-evolving contact network within a population; moreover, current technologies allow to estimate both the distance and the duration of the contact, that can be used to quantify the infection probability. Digital contact tracing raise several privacy issues, so that a lot of attention has been devoted to design privacy preserving protocols [29, 32, 115, 132]. On the other hand, less work has been carried out to quantify the efficiency of these protocols in terms of epidemic mitigation: indeed, most considered systems use the tracing data simply as a fast and scalable device to identify all recent contacts, in order to notify and eventually isolate all of them.

7.1.1 Bayesian Epidemic Tracing

The scope of the project discussed in this chapter is to show that probabilistic inference techniques allow to exploit the data collected by tracing applications about proximity events (i.e. contacts) to provide accurate estimates for the probability that a certain individual is infected; we will refer to this problem as *risk estimation* in the rest of the chapter. Probabilistic risk estimation provides a criterion that can be used by public health authorities to implement optimized sanitary protocols, by suggesting tests and other interventions on the individuals with the highest risk to be infected: this can be particularly relevant to detect individuals who do not show symptoms, as it happens with SARS-CoV-2. The Bayesian inference approach discussed in the following is based on a prior description of the epidemic process in terms of a probabilistic model of propagation, that is used to define a likelihood function, further conditioned by the additional information provided by observational data, as a result of tests (e.g. PCR, serology) and/or symptoms appearance.
Risk estimation is carried out approximately by using two different message-passing algorithms: the most accurate approximation scheme we use is based on Belief Propagation (BP), already introduced in Sec. 2.2 and presented in Section 7.3 for this application; a simpler approach that relies on a Mean-Field (MF) heuristics is discussed in Appendix D.
Both of them are based on systems of equations that need to be solved iteratively to provide a direct estimation of the individual's risk in terms of marginal probabilities, and require individuals that have been in contact in the recent past to be able to exchange messages about their risk level. When two individuals meet, they exchange a small amount of information (typically through Bluetooth); later on, they exchange messages carrying information about their current status, e.g. an increased risk due to presence of syndromes or due to their history of past contacts. Notice that this information is highly more accurate than standard contact tracing implementations, where the exchanged information is binary (i.e. two individuals have been in contact or not). Probabilistic inference then concatenates this information from all past contacts locally on the individuals phones and sends updates of the status to their contacts.
Both approaches have been already developed in the last years: in the present work, we adapt them to the current intervention scenario. Both methods have and have pros and cons in terms of efficiency, amount of data exchanged and privacy compatibility, and they differ by the accuracy of the approximation and on the way observations are encoded into the probabilistic model.

7.1.2 Propagation and epidemic spreading models

In order to apply the probabilistic approaches just introduced, it is necessary to use some mathematical description of the epidemic dynamics. Epidemic spreading models provide a simple characterization of the mechanisms behind disease transmission [7], and can be employed to investigate the sources of large epidemic outbreaks, as well as to analyze the dynamical properties of disease spreading over networks. Epidemic spreading models are tipycally defined in terms of a finite set of compartmental states, in such a way that each individual belongs to one of them at each time-step of the dynamics. One of the simplest and widest used is the so-called Susceptible-Infected-Recovered (SIR) model [74], that provides an accurate description of all those diseases where some immunity level to future infections is acquired after recovery (e.g. measles, chicken pox, influenza) but also for lethal diseases (HIV, Ebola), where the Recovered state is understood as "Removed". Apart from epidemiology, the SIR model has been applied to analyze spreading of viruses over networks of computers, as well as rumors over social networks [117].
The probabilistic approach based on BP/MF is carried out within the SIR description of the dynamics. In principle, the BP and/or MF equations can be generalized more complex compartmental models, provided that the dynamics is irreversible (recovered individuals acquire a long-lasting immunity and cannot become susceptible again). The latter assumption can be considered fairly realistic in the case of SARS-CoV-2: despite the presence of a long-term immunization on

recovered individuals is still under debate, we can assume that the time scale at which an eventual re-infection occurs ($\sim$ months) is much longer than the time-scale associated with the infection dynamics in a population ($\sim$ weeks).
On the other hand, the SIR model is unfeasible to capture some of the features observed in SARS-CoV-2 spreading, e.g. the presence of asymptomatic/symptomatic individuals. The most accurate mathematical descriptions of SARS-CoV-2 propagation are based on much richer compartment models, in which infected individuals are not immediately contagious upon infection, may be asymptomatic or develop mild/severe symptoms with some delay, and the ages, households and workplaces are also taken into account. Moreover, additional details about non-trivial distributions of incubation and recovery times, as well as of time-dependent viral transmission capacity are included in these models, as observed in SARS-CoV-2 [12, 50, 52, 73].
For these reasons, the quality of the inference protocols will be validated by using a more complex epidemic spreading model [114], specifically designed to reflect the main features of SARS-CoV-2 dynamics. It is important to remark that the inferential model used to develop risk assestment is much simpler than the one used to simulate the dynamics, at difference with other recently proposed approaches [58, 64]; neverthless, some of the ingredients of the realistic epidemic propagation can be included in the BP scheme, such as non-Markovian evolution between states and time-dependent infectiousness.

7.2 Bayesian inference in the SIR model

This section provides a general description of the Bayesian approach used to perform inference within the SIR model. Let us consider a graph $G = (V, E)$ representing the time-evolving contact network for a set of V individuals. Each node is associated to a time-dependent variable representing its state at each time t and denoted with x_i^t, taking values on a finite set of epidemic states: in the SIR model, $x_i^t \in \mathcal{X} = \{S, I, R\}$, where the three states refer respectively to Susceptible (S), Infected (I) and Recovered/Removed (R). In general, if the dynamics is non-Markovian, the quantity x_i^t will depend on individual i's state at all previous times, as well as on the states of all the individuals j that have been in contact with node i in the interval $[0, t]$. For simplicity, we here suppose that the dynamics is discretized on a sequence of finite time-steps $t \in \{0, \dots T\}$ (it is convenient to think to t as a number of days).
The SIR dynamics is fully specified by two sets of parameters: the individuals' recovery rate μ_i, defining the daily probability that the (infected) node i will recover, and the transmission rates $\{\lambda_{k\to i}\}$, that represent the probability that i will be infected by another node k at time t. In general, μ_i and $\lambda_{i\to j}$ might depend on the absolute time t of the dynamics and on the time elapsed since node i's infection: the latter dependency can be used to describe a time-dependent infectiousness (for instance, the initial incubation period of the virus in the organism), as well as clinical interventions (recovery, treatments, appearance of symptoms and so on, that influence the time-dependency of the recovery rate). We denote with $t_i = \min\{t : x_i^t = I\}$ the infection time of node i. Depending on the individual's current state x_i^t, between t and $t+1$ the following events can take place:

- if node i is susceptible ($x_i^t = S$), it can be infected by another individual j it has been in contact with with probability $\lambda_{j\to i}(t_j)$; if this happens, then $x_i^{t+1} = I$;
- if node i is infected, it can recover with a probability $\mu_i(t_i)$; if this happens, then $x_i^{t+1} = R$;

If none of these events happen then node i remains in its state, namely $x_i^{t+1} = x_i$: in particular, recovered individuals will always remain in this state. The above expressions define the transition rules for the SIR model, whose dynamics is schematized in Figure 7.1. In the following, we denote with $\boldsymbol{x}_i = \left(x_i^0, \dots, x_i^T\right)$ (resp. $\boldsymbol{x}^t = (x_1^t, \dots, x_N^t)$) the overall trajectory of node i (resp

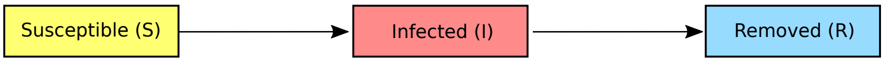

Figure 7.1: Dynamics of SIR model

the state of all nodes at time t). The above rules allow to compute the transition probability $p\left(x_i^{t+1} \mid \boldsymbol{x}^0, \dots, \boldsymbol{x}^t\right)$ for node i occurring between time t and time $t+1$:

$$p\left(x_i^{t+1} = S \mid \boldsymbol{x}^0, \dots, \boldsymbol{x}^t\right) = \mathbb{I}\left[x_i^t = S\right] \prod_{k \neq i} \left(1 - \lambda_{k \to i}\left(t_k\right) \mathbb{I}\left[x_k^t = I\right]\right) \tag{7.1a}$$

$$p\left(x_i^{t+1} = I \mid \boldsymbol{x}^0, \dots, \boldsymbol{x}^t\right) = \left[1 - \mu_i\left(t_i\right)\right] \mathbb{I}\left[x_i^t = I\right] + \mathbb{I}\left[x_i^t = S\right] \left\{1 - \prod_{k \neq i} \left(1 - \lambda_{k \to i}\left(t_k\right) \mathbb{I}\left[x_k^t = I\right]\right)\right\} \tag{7.1b}$$

$$p\left(x_i^{t+1} = R \mid \boldsymbol{x}^0, \dots, \boldsymbol{x}^t\right) = \mu_i\left(t_i\right) \mathbb{I}\left[x_i^t = I\right] + \mathbb{I}\left[x_i^t = R\right] \tag{7.1c}$$

where for simplicity we neglected the dependency on the absolute time t. In the above expressions, $\lambda_{k \to i}(t) = 0$ if k and i are not in contact at time t, and $\mathbb{I}[\cdot]$ denotes the indicator function of the condition given by its argument. In the simpler setup where μ_i and $\lambda_{i \to j}$ do not depend on the time since infection, the above dynamics becomes Markovian, so that $p\left(x_i^{t+1} \mid \boldsymbol{x}^0, \dots, \boldsymbol{x}^t\right) = p\left(x_i^{t+1} \mid \boldsymbol{x}^t\right)$.

Let us denote with $\mathbb{X} = \{x_i^t\}_{i=1,\dots,N}^{t=0,\dots,T}$ the collective time-trajectory of the epidemy, describing the status of each individual i at each time $t \in [0, T]$. For simplicity, we assume that the initial condition probability $p\left(\boldsymbol{x}^0\right)$ is factorized over single nodes, namely $p\left(\boldsymbol{x}^0\right) = \prod_i p\left(x_i^0\right)$; the prior probability associated with a trajectory $\mathbb{X}$ is given by:

$$p\left(\mathbb{X}\right) = \prod_i p\left(x_i^0\right) \prod_{t=0}^{T-1} p\left(\boldsymbol{x}^{t+1} \mid \boldsymbol{x}^0, \dots, \boldsymbol{x}^t\right) , \tag{7.2}$$

The Bayesian approach allows to easily include the effect of observations, providing some information about the an individual's state at a given time (for instance, given the results of test or the appearence of symptoms). We denote with $\boldsymbol{\mathcal{O}} = \{\mathcal{O}_r\}$ the set of observations; assuming that these are statistically independent, the posterior probability of the trajectory $\mathbb{X}$ can be expressed using Bayes theorem as:

$$\begin{aligned} p\left(\mathbb{X} \mid \boldsymbol{\mathcal{O}}\right) &= \frac{1}{p\left(\boldsymbol{\mathcal{O}}\right)} p\left(\mathbb{X}\right) p\left(\boldsymbol{\mathcal{O}} \mid \mathbb{X}\right) \\ &= \frac{1}{p\left(\boldsymbol{\mathcal{O}}\right)} p\left(\mathbb{X}\right) \prod_r p\left(\mathcal{O}_r \mid \mathbb{X}\right) \end{aligned} \tag{7.3}$$

where one identifies the $p\left(\boldsymbol{\mathcal{O}} \mid \mathbb{X}\right) = \prod_r p\left(O_r \mid \mathbb{X}\right)$ as the likelihood function. In order to estimate each individual's risk to be infected, we need to compute marginal distributions of the posterior:

$$p\left(x_i^t \mid \boldsymbol{\mathcal{O}}\right) = \sum_{\substack{t',j \\ (j,t') \neq (i,t)}} \sum_{x_j^{t'}} p\left(\mathbb{X} \mid \boldsymbol{\mathcal{O}}\right) \tag{7.4}$$

and the risk estimate is simply $p\left(x_i^t = I \mid \boldsymbol{\mathcal{O}}\right)$. For a comparison, the patient zero problem, i.e. the detection of the individuals who were infected at time 0 and generated the epidemic cascade, is mathematically translated into the computation of $p\left(x_i^0 = I \mid \boldsymbol{O}\right)$.

As extensively discussed in the whole thesis, the evaluation of (7.4) is computationally unfeasible when the number of variables is large, and one has to rely on suitable approximations. In the next section, the Belief Propagation approach to inference processes within the SIR model is presented.

7.3 Belief Propagation Approach to SIR

The Belief Propagation (BP) approach to epidemic spreading models has been developed in our group and extensively applied to several inference problems [2–5], e.g. the patient zero problem with noisy/partial observations and the reconstruction of causality chains of disease transmission. In this work, we adopted the more general continous-time approach developed in [19]. One of the advantage of the BP approach is that it can deal with non-Markovian processes, which is fundamental to capture some of the features of the SARS-CoV-2 spreading (as discussed in the next section). Moreover, the continous-time version can be used whenever interactions are registered without a prescribed time-frame, as it happens on realistic scenarios. The BP approach is based on the assumption that an irreversibile epidemic dynamics can be represented in terms of a static graphical model, defined by a finite number of transition times. In the SIR model, each individual's trajectory is uniquely determined by two quantities: the infection time $t_i = \min\{t : x_i^t = I\}$ and the recover time $r_i = \min\{t : x_i^t = R\}$. Indeed, given the full set of infection and recovery times, respectively denoted with $\boldsymbol{t} = \{t_i\}_{i\in V}$ and $\boldsymbol{r} = \{r_i\}_{i\in V}$, one can reconstruct the state of all the nodes at any time t:

$$p\left(\boldsymbol{x}^t \mid \boldsymbol{t}, \boldsymbol{r}\right) = \prod_i \xi_i\left(x_i^t, t_i, r_i\right) \tag{7.5}$$

where

$$\xi_i\left(x_i^t, t_i, r_i\right) = \mathbb{I}\left[x_i^t = S, t < t_i\right] + \mathbb{I}\left[x_i^t = I, t_i \leq t < r_i\right] + \mathbb{I}\left[x_i^t = R, t \geq r_i\right] \tag{7.6}$$

The BP approach can be defined on a time-evolving contact network without a prescribed time window. Each pair of individuals i, j is assumed to interact in a finite set of real times, denoted with $\mathcal{T}_{ij} \subset \mathbb{R}_\infty = \mathbb{R} \cup \{+\infty\}$; we assume that $\infty \in \mathcal{T}_{ij}$ for reasons that will become clear in the following. Suppose now that, among these contacts, one of these will give rise to an instantaneous contagion, at a time $s_{ij} \in \mathcal{T}_{ij}$, that occurs if i is infected and j is susceptible (on the other hand, j might infect i at a time $s_{ji} \in \mathcal{T}_{ji} \equiv \mathcal{T}_{ij}$). The infection probability is denoted $\lambda_{i\to j}^{s_{ij}}(t_i)$: in general it can depend on the absolute time s_{ij} on the event, and on the infection time t_i of the infector individual i (from now on, we use the short notation $\lambda_{i\to j} = \lambda_{ij}$, the directionality of the contact being understood). Individual i can thus become infected in one instant t_i in the set $t_i \in \mathcal{T}_i = \cup_{j\in\partial i}\mathcal{T}_{ji}$. In particular, when $t_i = \infty$ means that the individual never gets infected within the dynamics' time-frame. The recovery time $r_i \in \mathbb{R}_\infty$ is assumed to be drawn from a continuous distribution: since recovery can happen only if node i gets infected, i.e. for $t > t_i$, it is useful to define the recovery delay $r_i - t_i$ as the time interval occurring between infection and recovery. We denote with $R_i(r_i - t_i)$ the recovery delay's distribution. At given infection and recovery times (resp. t_i and r_i), the (conditional) probability distribution of the transmission times s_{ij}, denoted with $S_{ij}(s_{ij} \mid t_i, r_i)$ is given by:

$$S_{ij}\left(s_{ij} \mid t_i, r_i\right) = \mathbb{I}\left[t_i < s_{ij} < r_i\right] \lambda_{ij}^{s_{ij}}(t_i) \prod_{t_i < s < s_{ij}} \left[1 - \lambda_{ij}^{s}(t_i)\right] + \mathbb{I}\left[s_{ij} = \infty\right] \prod_{s \geq r_i} \left[1 - \lambda_{ij}^{s}(t_i)\right] \tag{7.7}$$

Indeed, i will be infectious in the open time interval (t_i, r_i) and can transmit the disease to j only for $s_{ij} \in [t_i, r_i]$, if the transmission has never occured before. In this representation, transmission occur independently on node j's state, while infenfection takes place only if j susceptible at time s_{ij}. The second term in (7.7) is justified by noticing that, if node i recovers at time r_i before a transmission occurs, it will never transmit the disease through that link; as a consequence, the transmission delay on that link will be nominally $s_{ij} = \infty$. The standard markovian (i.e memory-less) dynamics of the SIR model is recovered by setting a time-independent infection probability, i.e. $\lambda_{ij}^{s_{ij}}(t_i) \equiv \lambda_{ij}$ and an exponential distribution for the recovery delay, namely $R_i(r_i - t_i) = \mu_i e^{-\mu_i(r_i - t_i)}$, where μ_i is the (constant) recovery rate of node i.

7.3.1 Modelling auto-infections

The approach presented so far represents a closed systems in which infections can occur only through an existing contact between two nodes, one of them being infected. In order to take into account the presence of an initial number of infected seeds, it is necessary to introduce a probability that individuals can infect spontaneously, i.e. without having a contact with other infected people. Moreover, it will be useful to model the scenario in which a full knowledge of the contact network cannot be possible, as it happens in realistic situations. A simple solution is to add a series of contacts with a virtual node who is always infected. In particular, for each node i and for each time $t \in \mathcal{T}_i \setminus \{\infty\}$, we will add a series of contacts with this always-infected neighbour $\tilde{i}^t$ at all times $t \in \mathcal{T}_i \setminus \{\infty\}$, so that with probability $p_t\left(s_{\tilde{i}^t,i} = t\right) = \gamma_i^t$ (resp. $p_t\left(s_{\tilde{i}^t,i} = \infty\right) = 1 - \gamma_i^t$), node i will spontaneously self-infect at time t with probability γ_i^t (resp. it never gets self-infected at t), provided that it is susceptible at that time. Let us define the quantity $A_i\left(\boldsymbol{s}_{i^*}\right) = A_i\left(\left\{s_{\tilde{i}^t i}\right\}_{t\in\mathcal{T}_i}\right) = \prod_{t\in\mathcal{T}_i\setminus\{\infty\}} p_t\left(s_{\tilde{i}^t i}\right)$, that describes the overall self-infection probability for node i. For convenience, we also define with $\partial^* i$ the enlarged neighborhood of i including all virtual nodes $\left\{\tilde{i}^t\right\}_{t\in\mathcal{T}_i}$. The infection time t_i can thus be expressed deterministically through the following condition:

$$t_i \quad = \quad \min_{k\in\partial^* i} s_{ki}, \tag{7.8}$$

The joint probability distribution of the SIR dynamics can now be rewritten in terms of the infection times $\boldsymbol{t} = \{t_i\}$, the recovery times $\boldsymbol{r} = \{r_i\}$ for all nodes, and the transmission times $\boldsymbol{s} = \{s_{ij}\}_{i,j=1\dots N}$:

$$p\left(\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{s}\right) \quad \propto \quad \prod_i \delta\big(t_i, \min_{k\in\partial^* i} s_{ki}\big) A_i\left(\boldsymbol{s}_{i^*}\right) R_i\left(r_i - t_i\right) \prod_{(ij)} S_{ij}\left(s_{ij} | t_i, r_i\right) \tag{7.9}$$

In the above expression, the set of delta functions enforce the constraints (7.8) and the product over (ij) runs over all the contacts for each pair of nodes (including also the extra-neighbour interactions).

7.3.2 Factor graph representation

Combining all the previous terms together and adding the observations, the posterior probability distribution of an epidemic trajectory given the set of observations can be written as:

$$p\left(\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{s} \mid \boldsymbol{\mathcal{O}}\right) \quad \propto \quad \prod_i \delta\big(t_i, \min_{k\in\partial^* i} s_{ki}\big) A_i\left(\boldsymbol{s}_{i^*}\right) R_i\left(r_i - t_i\right) p_{O,i}\left(\mathcal{O}_i \mid t_i, r_i\right) \prod_{(ij)} S_{ij}\left(s_{ij} \mid t_i, r_i\right) \tag{7.10}$$

where $\prod_i p_{O,i}\left(O_i \mid t_i, r_i\right)$ is the likelihood function. Observations can carry out structured information on individual i's state, for instance about its current infectivity and viral charge, as well as its symptoms. However, we will restrict to the simplest case where only the "SIR" state of node i can be detected: therefore, an individual can be observed in one of the three states S, I, R at any time. In the absence of noise, observations simply put bounds on the infection and recovery times (for instance, if node i is observed S at time t, its infection time t_i must be greater than t). On the other hand, in the case of noisy observations these constraints are relaxed, depending on the false/positive negative rates associated to test accuracies.
The factor graph of (7.10) is composed by a set of factors ψ_i, each one depending on variables $t_i, r_i, \{s_{ij}\}_{j\in\partial i}, \{s_{ji}\}_{j\in\partial^* i}$. However, such representation contains many short loops, as shown by the central panel of Fig 7.2: for instance, pairs $\left(t_i, s_{ji}\right), \left(t_i, s_{ij}\right), \left(t_j, s_{ij}\right), \left(t_j, s_{ji}\right)$ share respectively factors with indices $i, (ij), j, (ji)$, effectively forming a small cycle. The presence of short

loops would make BP approach not exact even if the contact network is a tree. However, a simple solution can be constructed by grouping together variables s_{ij} and s_{ji}, and consider them as a single variable (s_{ij}, s_{ji}): in the resulting factor graph, each variable (s_{ij}, s_{ji}) has degree two and lives in the middle of the original edges, while t_i, r_i have degree 1. In this way, the resulting topology (shown in Figure 7.2, right panel) closely reflects the one of the original contact network. With this parametrization, the posterior probability can be written as:

$$p\left(\boldsymbol{t},\boldsymbol{r},\boldsymbol{s}\mid\boldsymbol{\mathcal{O}}\right)=\frac{1}{Z}\prod_{i}\psi_i\left(t_i,r_i,\{s_{ki},s_{ik}\}_{k\in\partial^*i},O_i\right) \tag{7.11}$$

where

$$\psi_i\left(t_i,r_i,\{s_{ki},s_{ik}\}_{k\in\partial^*i},O_i\right)=\delta\big(t_i,\min_{k\in\partial^*i}s_{ki}\big)A_i\left(\boldsymbol{s}_{i^*}\right)R_i\left(r_i-t_i\right)p_{O,i}\left(\mathcal{O}_i|t_i,r_i\right)\prod_{k\in\partial i}S_{ik}\left(s_{ik}|t_i,r_i\right) \tag{7.12}$$

Each factor ψ_i depends on the infection and recovery times of node i, on the set of observations on it, and on all the transmission times (both inwards and $\{s_{ki}\}_{k\in\partial^*i}$ outwards $\{s_{ik}\}_{k\in\partial i}$. The message-passing equations are then defined over variables (s_{ij}, s_{ji}), while the equations for t_i, r_i can be computed straighforwardly by using (7.8) and the distribution of recovery delays $R_i\left(r_i - t_i\right)$. The derivation of the message-passing equations follows directly as described in Section 2.2, and their explicit form (with further details about the computational complexity) is discussed in Appendix E.

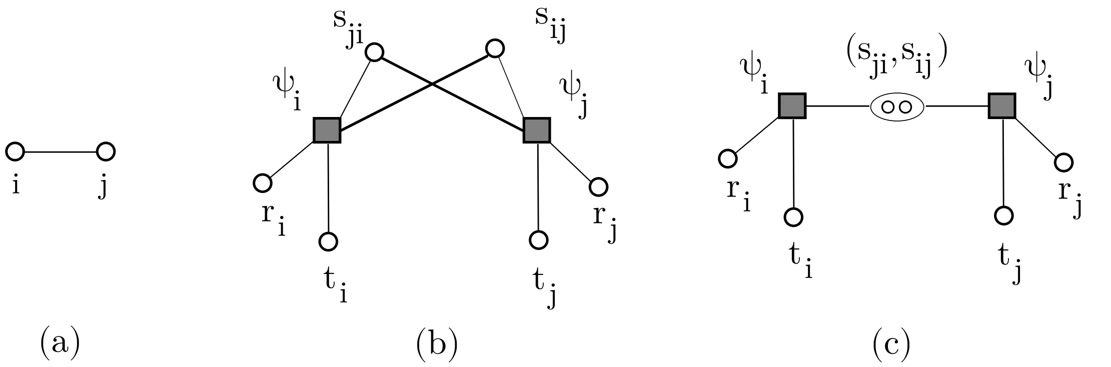

Figure 7.2: Factor graph representation of the BP posterior. (a) Simple graph with two nodes: the edge includes all the contacts between i and j at all times. (b) Naive interpretation of the posterior with short loops. (c) Adjusted factor graph with variables (s_{ij}, s_{ji}) paired together, so that the topology reflects the one of the starting graph.

7.4 Agent Based Model for SARS-COV-2 dynamics

This section provides a detailed description of the propagation model we used to simulate a realistic spreading of SARS-COV-2 [49, 114]. The Agent Based Model (referred in the following as ABM) was among the first ones published at the beginning of the european outbreak and it was designed to quantify and predict the epidemic evolution on an age-stratified population; moreover, it allows to evaluate the effect of different containment measures, from global lockdowns to local isolations for symptomatic individuals, as well as standard contact tracing procedures. The ABM is defined by a more complex infection dynamics w.r.t. the standard SIR model just discussed,

in order to better take into account for the presence of different infective individuals, namely the presence of asymptomatic and symptomatic individuals with symptoms of different strenghts, and it also includes non-trivial distribution of infectiousness and transmission times. The ABM will be used to simulate a more realistic epidemic spreading onto which the Bayesian inference prodecure discussed before will be exploited to estimate the individual's risk to be infected.
The ABM is defined on a fixed population of N individuals, categorized into 9 age-groups by decade, from $(0-9)$ to $(80+\text{years})$; in the following, we will denote with $a_i \in \{1,\ldots,9\}$ the age group of node i. Individual are divided into houses of different sizes, which is necessary to model one of the three types of interaction domains considered within the model. The distribution of individuals in the different age group is sampled in such a way to match UK demographics data (additional information can be found in Table 1 of [114]). Age stratification influences the households' composition (for instance, elderly people are more likely to live with other elderly, while children will preferably live with young adults), the social activity level of individuals, as well as the viral transmission capacity.

7.4.1 Contact Network

The contact network is constructed as the superposition of three graphs, each one describing a different interaction domain:

1. **Household network**: each individual interacts with all the others in the same household, so that the contact network in each household is fully connected (for each day). The size of each household runs from 1 (individuals living on their own) to 6. Once again, the distribution of individuals in houses of different sizes is sampled so to match UK demographics data.

2. **Workplace network**: each individual is a member to one workplace, that models the interaction inside schools (for children), offices and similar workplaces for adults and recurrent social activities for elderly people. The interactions within each workplace are generated by using the Watts-Strogatz small-world network model [137]. Each workplace network is static, and individuals are assigned to a specific workplace at the beginning: however, daily interactions are down-sampled, so that an individual interacts with a random fraction (typically half) of its workplace connections on each day.

3. **Random network**: this network models spurious interactions that occur once in a while independently for each day. The number of random contacts for each individual is the same each day and it is drawn from a negative-binomial distribution. In this way, since this distribution is over-disperse, the ABM takes into account the presence of super-spreaders in the network.

Additional details about the parameters used to sample the workplace and random network can be found in Table 3-4 of [114]. Mean values of the number of daily interactions are fitted according to recent studies of social interactions [121]. To give an order of magnitude, the mean value of the daily interactions for an individual with age a_i - obtained by summing the workplace and random contacts - runs from 6 for elderly individuals to 12 young and adult people (without counting household contacts).

7.4.2 Dynamics

The ABM dynamics is modelled as a discrete-time stochastic process with a temporal resolution of 1 day and it is schematized in Fig 7.3. At each day, individuals belong to one of the 11 comparmental states. Individuals typically are initialized as suscetible (S) at time (day) 0 (except

a small number of infective seeds, otherwise there would be no epidemic spreading). Upon infection, an individual i enters into an irreversible cascade of different infection routes, depending on its age a_i, ending in one of two absorbing states: removed (R) or dead (D). In particular, suppose that at a certain time t individual i get infected (details about the infection probability will be discussed in the next paragraph), there are three possible infective pathways: it can become asymptomatic (A), mild-symptomatic (SM) or severe symptomatic (SS), respectively with age-dependent probabilities $\phi_A(a_i)$, $\phi_{SS}(a_i)$, $\phi_{SM}(a_i)$ [1]. These probabilities take into account that the disease is more likely to affect elderly individuals with severe symptoms, with respect to young people. The asymptomatic route is the simplest: once infected, these individuals can eventually recover after a time $\tau_{A \to R}$ drawn from a Gamma distribution with a mean of 15 days. On the other hand, symptomatic people first enter into a pre-symptomatic state, denoted with pSS and pSM for severe (resp. mild) individuals, used to describe the starting incubation period of the virus; neverthless, they are already potentially infective at this stage. In both cases, symptoms appear after a random characteristic time τ_{sym} drawn from a Gamma distribution with a mean of 6 days (equal for mild and severe symptomatics). Mild symptomatic individuals can then recover after $\tau_{SM \to R}$ days. The infection dynamics for severe symptomatics individuals is more complex: a fraction of these might get hospitalized (H) with a (age-dependent) probability $\phi_H(a_i)$ after a random number of days τ_H drawn from a shifted Bernoulli distribution (in particular, $\tau_H \in \{5{,}6\}$ days with equal probability). The remaining fraction $1 - \phi_H(a_i)$ recovers after $\tau_{SS \to R}$ days. Hospitalized individuals can either recover (after $\tau_{H \to R}$ days), die (after τ_{crit} days) or get to a more critical state where intensive care unit (ICU) is needed (again, after τ_{crit} days): these transitions occur respectively with age-dependent probabilities $1 - \phi_{crit}(a_i)$,$\phi_{crit}(a_i)(1 - \phi_{ICU}(a_i))$, $\phi_{crit}(a_i)\,\phi_{ICU}(a_i)$. Hospitalized individuals can eventually die (with probability $\phi_D(a_i)$ after τ_D days) or recover. Recovery from ICU is modelled in two steps, in order to quantify the occupation time of a ICU seat in a realistic scenario. In particular, ICU individuals who recover enter first into a survival state (denoted with HR, i.e hospitalized recovering) with probability $1 - \phi_D(a_i)$ after τ_{surv} days, and then recover definitely after $\tau_{HR \to R}$ days.
The compartmental states describing individuals in hospitals (or ICU) are unrelevant from the point of view of the infection dynamics, and they are introduced to quantify the amount of hospital/ICU seats needed by the National Healthcare System as the epidemy grows. We remark that, with the exception of the waiting time for severe individuals to go to the hospital (τ_H), all the transition times are sampled for each individual from different Gamma distributions with fixed mean and variances. All these parameters (and the age-dependent probabilities $\phi_*(a_i)$) can be found in Table 7 of [114] and they have been fitted from epidemiologic/medical analysis carried out on the first outbreak in the province of Hubei (China).

Infection probability

The infection spreads from infected individuals through their interactions with other susceptible nodes. Contacts are assumed to be instantaneous and the viral transmission capacity of an individual starts at the moment it has been infected, denoted with t_i. The infectivity rate has an explicit time dependency, used to model reflect the time-dependent transmission capacity of SARS-COV-2 and its starting incubation period [77]. In addition, the overall infection probability is modelled by taking into account the status of the infector node i, $s_i \in \{A, SM, SS\}$, the susceptibility of the potential infected node j (that depends in turn on its age a_j) and the contact network in which the contact occurs. The daily probability of infection $\lambda_{i \to j}$ is computed in the

[1] for each age group a these probabilities are normalized, namely $\phi_A(a) + \phi_{SS}(a) + \phi_{SM}(a) = 1$

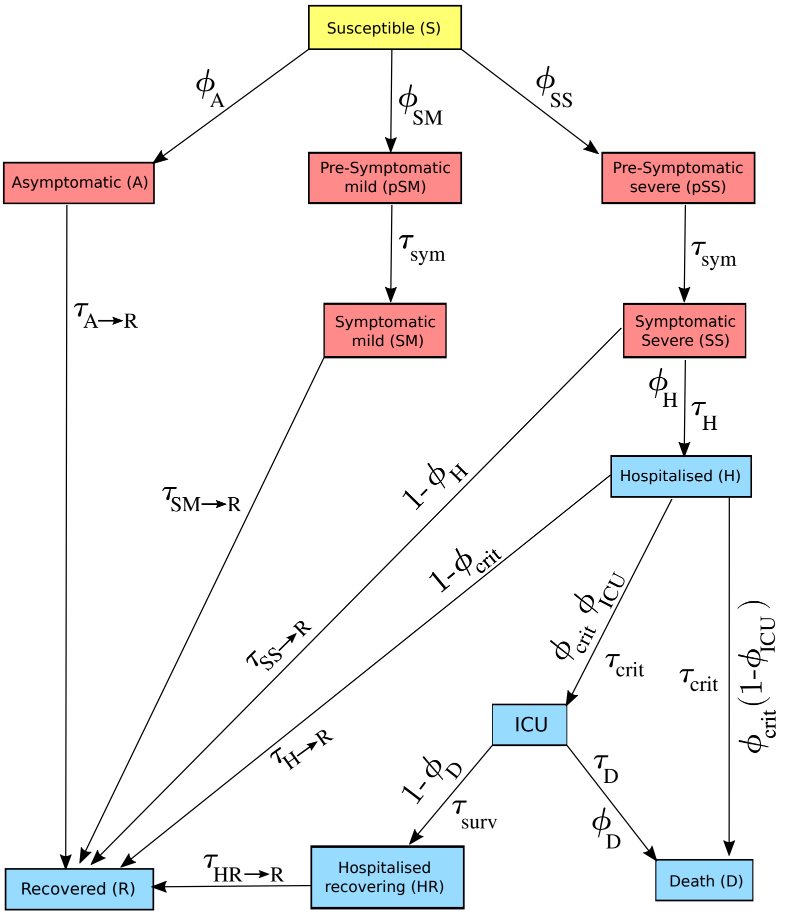

Figure 7.3: Infection dynamics of the ABM model. All the transition probabilities ϕ_* are functions on the individuals's age a_i, but their dependecy has been dropped to avoid cluttering the notation. The color of each compartment describes how it is considered within the SIR parametrization used for the inference, according to Figure 7.1.

ABM as follows:

$$\lambda_{ij}(t-t_i) = \Theta(t-t_i)\left\{1-\exp\left[-\Lambda_0(a_i,a_j,\mathcal{G}_{ij},\mathcal{I}_i)\,\Lambda(t-t_i)\right]\right\} \tag{7.13}$$

where

$$\Lambda(t_i) = \int_{t_i-1}^{t_i} f_\Gamma(z \mid \mu_I,\sigma_I)\,dz \tag{7.14}$$

The time dependent part $\Lambda(t)$ given by (7.14) can be considered as an integrated continuous infection rate over a 1-day window, where f_Γ is a Gamma function with fixed mean and standard deviation. The behaviour over time of $\Lambda(t)$ is shown in the right plot of Figure (7.4): the infectiousness start at 0 at the time some individual is infected, then it reaches a maximum value and it starts decreasing afterwards. Notice also that the maximum value of $\Lambda(t)$ is reached, on average, after the same number of days at which symptomatic individuals start to show symptoms (the random variable is denoted with τ_{sym} in the previous section and it is drawn from a Gamma distribution with the same parameters as f_Γ). On the other hand, the prefactor Λ_0 mainly depends on the age of the two individuals i and j and the status of the infector node (i), and it is used to model the susceptibility of the potential attacked node j as a function of its age, as well as the attack rate of the infector node: for instance, elderly individuals are more susceptible to infection, and symptomatic individuals typically have a larger transmission capacity w.r.t. asymptomatics. The explicit formula is not shown here for simplicity and we refer to [114] for further details. The left plot of Figure (7.4) shows the histogram of a tipycal realization of infective scales Λ_0 on a population of $N = 10^4$ individuals, simulated with the ABM. It is important to remark that such individual's based information cannot be accessed by standard contact tracing implementations, mainly for privacy issues. The only information we keep track of is the prefactor depending on the contact network over which the contact is realized, denoted with $\mathcal{G}_{ij}$: in particular, this multiplicative term it is equal to 1 if the nodes i and j have a contact in the workplace or in the random network, and 2 for contacts in the same house. In this way, even if the contacts are considered istantaneous it is implicitly assumed that household interactions have a longer duration and the corresponding infection probability is larger (doubled).

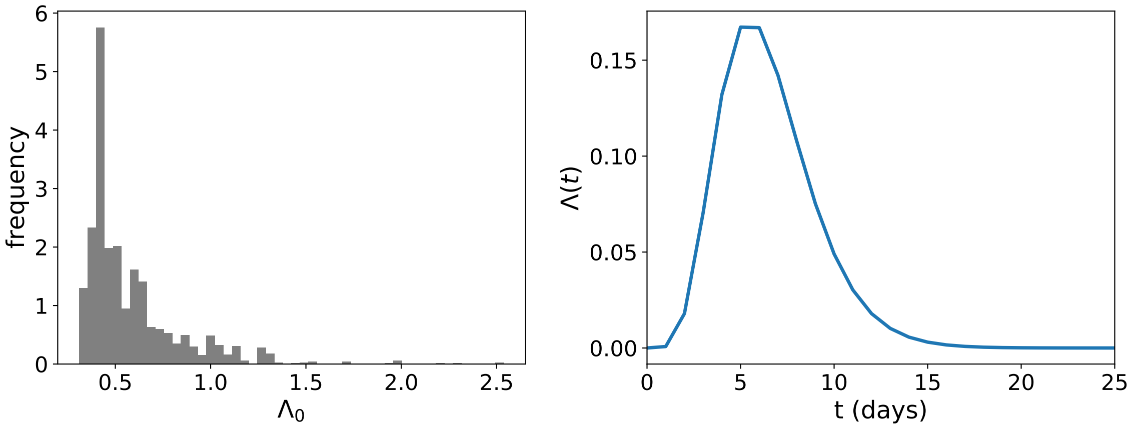

Figure 7.4: Left: normalized histogram for the distribution of $\Lambda_0$values for a tipycal realization of the ABM in a population of $N = 10^4$ individuals. Right: plot of the time-dependent part of the infection rate $\Lambda(t)$, modelled as a (integrate) gamma function with $\mu_I = 6$ days, $\sigma_I = 2.5$ days.

7.5 Wrapping the ABM onto a SIR model

The ABM dynamics discussed in the previous section is remarkably more complex than the usual compartmental models used in statistical physics, with a total of 11 states vs the 3 states of SIR model. However, the purpose of this work is to show that even a simpler description of the dynamics within the inference procedure is still able to detect the individuals with the highest risk to be infected, even on top of a complex dynamics that tries to mimic a realistic epidemic spreading. In this section, we discuss how to map the ABM dynamics onto the SIR procedure used for the inference within BP; the MF case is discussed in Appendix D. The color map in Figure 7.3 describes the mapping: the 5 infective states of the ABM (A,pSM,pSS,SM,SS) are naturally considered as infected (I) in the SIR model. On the other hand, all the states from Hospitalized to the absorbing ones (H,ICU,HR,D,R) are considered as Removed (R): the underlying hypothesis is that individuals do not have contacts from the time they get hospitalized, which is - or at least, it should be - a reasonable assumption in a real-case scenario and thefore, from the point of view of the SIR, they can be considered as removed[2].

Infection probability

As previously discussed, in the ABM the infection probability has a complex structure that depends both on the information about the two individuals in contact, and on the time elapsed since node's i infection. The information about the two nodes cannot be in general accessed by standard contact tracing information because of privacy issues (for instance, we cannot know the age of the two nodes i and j). The only two ingredients we use to define the scale factor of the infection probability is the multiplier $\mathcal{G}_{ij}$ of the contact network: we recall that $\mathcal{G}_{ij}$ is equal to 2 for intra-house contacts and 1 otherwise. Moreover, the BP scheme allows to take into account the time-dependency of the infectivity, which we assume to be known from the literature about SARS-CoV-2. Combining all these effects, the daily infection probability λ_{ij}^{BP} is given by:

$$\lambda_{ij}^{BP}\left(t-t_i\right)=\Lambda_0^{BP}\times f_\Gamma\left(t-t_i \mid \mu_I,\sigma_I\right); \quad \Lambda_0^{BP}=0.25\times\mathcal{G}_{ij} \tag{7.15}$$

where t_i is the infection time of node i, as estimated by BP. The numerical prefactor in Λ_0^{BP} is obtained by averaging the scale factors of a typical realizations of the ABM on a population of $N=10^4$ nodes. The parameters of the Gamma function f_Γ are the same of (7.14).

Recovery time

In the standard SIR model, a constant recovery probability/rate is typically assumed, thus resulting into a memory-less exponential (or geometric on discrete time) distribution for the recovery time. However, some of the information about COVID-19 transmission can be exploited to design a more refined recovery time to be used for the BP inference. Let us denote with τ_r^{BP} the averaged recovery time used in BP. From the ABM we can carry out a coarse-grained estimation for the recovery distribution, by tracing over the different infective routes. In the ABM there are 4 different infection pathways an individual can enter, as shown by Figure 7.3: the asymptomatic and mild-symptomatic paths, and the severe symptomatic path, the latter ending in either a Recovered or Hospitalized state (further transitions are neglected as previously discussed). For each path, we compute the expected recovery time distribution: then, an averaged rate is computed by weighting each path by the probability that an individual with age a_i will undergo through that specific path. The resulting probability distribution for τ_r^{BP} is given by:

$$\begin{aligned} p\left(\tau_r^{BP}\right)=&\left\langle\phi_A p\left(\tau_r^A\right)+\phi_{SM}p\left(\tau_r^{SM}\right)+\right.\\ &\left.+\phi_{SS}\left(1-\phi_H\right)p\left(\tau_r^{SS}\right)+\phi_{SS}\phi_H p\left(\tau_H^{SS}\right)\right\rangle_{p(a)} \end{aligned} \tag{7.16}$$

where

$$\tau_r^A=\tau_{A\to R} \tag{7.17a}$$

$$\tau_r^{SM}=\tau_{sym}+\tau_{SM\to R} \tag{7.17b}$$

$$\tau_r^{SS}=\tau_{sym}+\tau_{SS\to R} \tag{7.17c}$$

$$\tau_h^{SS}=\tau_{sym}+\tau_{SS\to H} \tag{7.17d}$$

and the + symbol is intended as a sum of 2 random variables, whose p.d.f. is given by the convolution of the two addends' densities. Furthermore, since we want to define a single recovery

[2] In principle, the ABM takes into account the possibility to include interactions among hospitalized individuals. In that case, in the H state transmissions can still occur and it has to be treated as infective within the SIR parametrization.

time independently on the individual's age, a further average on the age groups distribution $p\left(a\right)$ is performed. The empirical distribution of τ_r^{BP} (7.16) is then fitted by a Gamma p.d.f.: indeed, since all the transition times are gamma-distributed (except for $\tau_{SS\rightarrow H}$) it is reasonably to assume that the averaged recovery time is well described by a Gamma distribution. Therefore, $\tau_r^{BP} \sim \text{Gamma}\left(\mu_r^{BP}, \sigma_r^{BP}\right)$ with $\mu_r^{BP} \cong 18$ days and $\sigma_r^{BP} \cong 5.6$ days as a result of the fitting.

Finite window approximation

The number of exchanged messages between two individuals grow quadratically with the number of temporal contacts occurred between them. However, it is reasonably to assume that only recent contacts are important to determine marginal probabilities at current time: in this perspective, we choose to discard past information and keep a short time window ($\Delta t \sim$ 2,3 weeks), in order to obtain quasi-optimal results with a fixed computational cost. Therefore, at each time step t we apply the BP inference on the time interval $t' \in [t - \Delta t, t]$. The information about contacts and observations at the previous dropped times is included approximately as simple factorized priors applied at the start of the window $t - \Delta t$: for each node, the corresponding prior contains the posterior probability at the first non-dropped time computed only using contacts and observations at the dropped time (and the prior computed in the previous step). All simulations have been performed using a $\Delta t = 21$ days time window.
The self-infection probability is chosen as $\gamma^0 = 1/N$ at time $t = 0$ (k/N where k is the number of patient zeros would bring slightly better results, but would use inaccessible information) and 0 for $t > 0$, except for the cases with partial adoption (discussed in Sec. 7.7.3).

7.6 Ranking Strategies and intervention protocol

In this section, we describe the implementation used to run the online mitigation strategy. The ABM dynamics is initialized by choosing a small number of infected individuals $n_0 > 1$, also called patient zeros or infective seeds. This setting describes either an early-stage epidemic spreading or a post-lockdown scenario with several small outbreaks. As time advances, we assume that all the individuals that develop severe symptoms will immediately be tested and quarantined at the time symptoms appear. A node i is quarantined at time t_i^q meaning that it will only have contacts inside its households for $t \in [t_i^q, T]$ where T is the simulation window (in practice, since the dynamics is irreversible, it is sufficient to keep a node quarantined until recovery or hospitalization occurs). The same protocol is used for a fraction ρ_{SM} of mild symptomatic individuals, that are assumed to self-report symptoms and get tested. For both mild and severe symptomatics, $t_i^q \equiv t_i^{sym}$ where t_i^{sym} is the time at which node i shows symptoms: however, in the time window $[t_i, t_i^{sym}]$ where t_i is the infection time, they are still are free to move and spread the disease, and their detection can be made only by contact tracing. On the other hand, asymptomatic individuals can be identified only by using the contact tracing based inference. The simulation is run with these settings in the window $[0, t_{start}]$, after which the inference-guided intervention protocol starts. Every day, we run the inference algorithm (either BP or MF) to estimate each individual's risk to be infected: then, we perform a fixed amount n_{obs} of tests to the individuals with the highest risk. In this framework we ideally suppose that the test's result is immediately available. Individuals whose test turns out to be positive are immediately quarantined. In addition to BP and MF, we implement two other strategies, a random procotol and another one based on standard contact tracing implementations. We show below the details of each strategy:

- Random Guessing (RG): for each time step t, we randomly select n_{obs} individuals to be tested, among those who have not been previously tested positive;

- Contact Tracing (CT): for each time step t, individuals who have not been tested positive previously according are ranked according to the number of contacts with confirmed positive individuals during the time interval $[t - \tau_{CT}, t]$; then the n_{obs} individuals with the largest number of contacts are tested. This procedure reflects standard contact tracing implementations available so far [65]. In all the simulations, $\tau_{CT} = 5$ days.
- Mean-Field (MF): we run the MF update equations (discussed in Appendix D) at each day in order to estimate the marginal probability $q_i^t = p\left(x_i^t = I\right)$ for node i to be infected at time t. Individuals who have not previously been tested positive are ranked according to their risk q_i^t, so that the top n_{obs} individuals are tested. The test results are then included in the observations used to adjust the probabilities of risk on the next time step. In all the simulations we used constant values for the two algorithmic parameters: $\delta_{MF} = 5$ days and $\tau_{MF} = 10$ days. Notice that $\delta_{MF} \approx \tau_{sym}$ where the latter is the typical time at which symptoms appear in the ABM.
- Belief Propagation (BP): we run the message passing equations described in appendix E in order to estimate the marginal probabilities $q_i^t = p\left(x_i^t = I\right)$ for node i to be infected at time t. Individuals who have not previously been tested positive are ranked according to their risk q_i^t, so that the top n_{obs} individuals are tested. For BP, the rank is computed by summing the probability of infection in the last δ^{BP} days. In this way, we assume that the most recent infections are more dominant than older ones. The test results are then included in the observations used to adjust the probabilities of risk on the next time step. The equations are run for a fixed number of iteration n_{it} each day. In all the trials, $n_{it} = 40$, $\delta^{BP} = 10$ days.

The expected behaviour is that, the better the ranking strategy used, the more effective the mitigation will be. In this perspective, the random guessing strategy is understood as a worst-case scenario, as the ranking is performed by random shuffling the list of individuals. The intervention quality will be measured by considering the number of infected individuals over time. Algorithm 7.1 describes the pseudocode implementation used to run all the simulations presented in the next section. It is important to remark that all quarantined individuals keep having contacts with their households: as a stronger intervention protocol, we can choose to quarantine the whole household once one of its cohabitants is tested positive (this option is shown in Algorithm 7.1 as "HH", and its effect will be discussed in Sec. 7.7.1). The main drawback of the method presented so far is that the test result is immediately available (or at least, in the same day). This is not a truly realistic option, especially for PCR tests, or when a large number of tests are performed on a daily basis, with a consequent slow-down of the analysis procedure. However, we can reasonably assume that individuals who are tested choose to self-isolate until the outcome is available, so that the dynamics is not severly affected by the waiting time of the test results.

7.7 Results

In this section, we quantify the effect of the different intervention protocols (RG, CT, MF, BP) on the ABM dynamics. All the simulations are made on a population of $N = 500.000$ individuals and a time window of $T = 100$ days. We consider two realistic scenarios in which the dynamics is initialized with a small number of infected patient zeros n_0 and the contact-tracing based interventions starts after some time t_{start} (expressed in days). In the scenario A, $n_0 = 50$, $t_{start} = 10$, while in the scenario B, $n_0 = 20$, $t_{start} = 7$. All the results shown in the following are obtained by implementing the quarantination protocol discussed before: for $t > t_{start}$ one the 4 ranking strategies (RG, CT, MF, BP) are used to identify the individuals with the highest probability to be infected at each time step t, by performing n_{obs} daily tests on the

Algorithm 7.1 Intervention strategy on the ABM dynamics.

```
Input: $n_0$ (number of patient-zero), $T$ (verall simulation window), $n_{obs}$ (number of daily observations), ranking strategy
while $t < T$
    evolve the ABM dynamics for 1 time-step (day)
    select all (new) SS individuals → quarantine (+HH)
    choose a fraction $\rho_{SM}$ of the (new) SM individuals → quarantine (+HH)
  if $t > t_{start}$
      run the ranking procedure to estimate the infection risk of each node
      rank individuals according to their risk, select the top $n_{obs}$ of them
      test: if test positive → quarantine (+HH)
```

highest risk nodes. On top of that, all the severe symptomatic and a fraction $\rho_{SM} = 0.5$ of mild-symptomatics individuals are immediately tested and quarantined at the day symptoms appear, starting from $t = 0$. Notice that, while the number of medical tests performed on individuals detected by the inference is fixed for each day, there is no limitation on those performed to the fraction of symptomatic people. Simulations are run for a number n_i of different instances by varying the random seed used to inizialize the ABM dynamics; in this way, at fixed seed and for $t \leq t_{start}$ the evolution is independent on the ranking strategy applied. For the time being, we assume that the results of the test is noise-less and the full knowledge of the contact network is available: the latter conditions mimic an ideal scenario in which all individuals in a population use the contact-tracing app and the results of the test is 100% reliable. The first results are shown in Figure 7.5-7.6 respectively for the scenario A and B, for different values of the number of daily observations (in particular, $n^A_{obs} \in \{625{,}1250{,}2500{,}5000\}$ and $n^B_{obs} \in \{250{,}500{,}1000\}$.

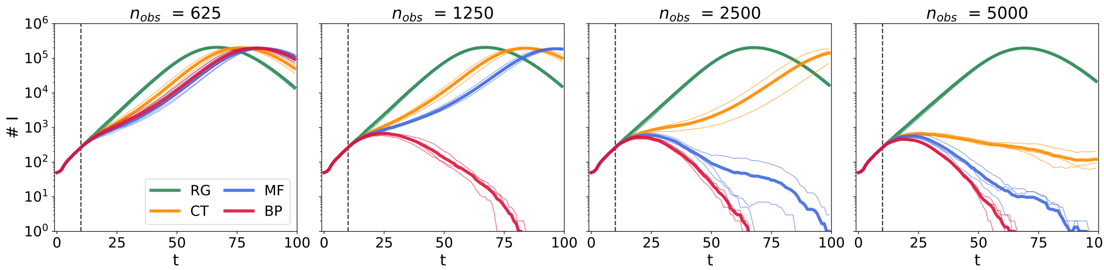

Figure 7.5: Effect of the mitigation strategies on the ABM dynamics on scenario A: $n_0 = 50$ and $t_{start} = 10$ (represented by the vertical dotted line). Each panel shows the number of infected individuals w.r.t. time, for different values (increasing from left to right) of daily obervations $n_{obs} \in \{625{,}1250{,}2500{,}5000\}$. In each plot, the thin lines show 3 different instances obtained by varying the random seed of the ABM dynamics (thin lines), while the thick lines represent their average.

Results show that, in both scenarios, the intervention protocol guided by MF/BP inference is able to contain the epidemic spreading with a significant reduction of the overall number of infected individuals; on the other hand, Contact Tracing's main effect is to delay the infection peak with respect to the random guessing protocol. In particular, BP always outperforms all the other methods, being the only protocol to efficiently mitigate the epidemic spreading in a wide range of parameters: for instance, in the scenario A for $n_{obs} = 1250$ or B for $n_{obs} = 500$. Scenario B is more interesting because it represents a more feasible situation for a early-growing epidemic spreading, in which the BP-based inference can be effective in mitigating the spreading with a reasonably

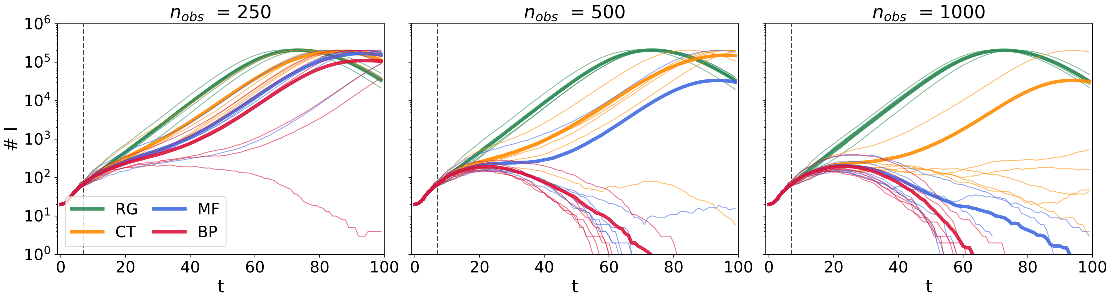

Figure 7.6: Effect of the mitigation strategies on the ABM dynamics on scenario B: $n_0 = 20$ and $t_{start} = 7$ (represented by the vertical dotted line).Each panel shows the number of infected individuals w.r.t. time, for different values (increasing from left to right) of daily obervations $n_{obs} \in \{250{,}500{,}1000\}$. In each plot, the thin lines show 3 different instances obtained by varying the random seed of the ABM dynamics (thin lines), while the thick lines represent their average.

low amount of tests (notice that in the scenario B the number of daily observations is roughly 1 order of magnitude lower than scenario A). For a comparison, we remark that the number of tests daily performed in Italy during the second wave (October-November 2020) is approximately 2×10^5 on a country-wide population of 6×10^7 individuals, that gives a comparable fraction of tests over population size compared to scenario B (i.e. $1/3 \times 10^{-2}$ vs 10^{-3} when $n^B_{obs} = 500$). We can conclude by saying that BP needs the least amount of daily tests to effectively mitigate the epidemic with respect to the other strategies; in this perspective, MF is generally better than CT which in turns is still better than the worst-case scenario given by RG. Despite its simplicity, the MF is able to mitigate the exponential growth when the number of observations is sufficiently large (for instance, $n_{obs} = 2500{,}5000$ in scenario A and $n_{obs} = 1000$ on scenario B); however, even in these regimes, a full suppression occurs about 20 days later than BP strategy.

7.7.1 Quarantining the households

We now discuss the effect of a stronger intervention protocol: as soon as an individual gets tested positive (either if it was detected by the ranking algorithm or if symptoms appear), it is immediately quarantined together with all its cohabitants, the latters without being tested. Figure 7.7 shows the results of such procedure for the scenario A (top panel) and B (bottom panel), for the same number of daily observations used before. In this setting, all the methods perform quantitatively better than in the previous case: for instance, even Contact Tracing (CT) is able to efficiently mitigate the spreading when the number of observation is sufficiently large (as in the top panel of Figure 7.7 for $n^A_{obs} = 2500{,}5000$ or the bottom panel for $n^B_{obs} = 1000$). For what concerns the scenario B, BP seems to be able to suppress the spreading even when the minimum amount of test chosen (the lower-left panel of Figure 7.7 with $n^B_{obs} = 250$) on at least 5 out of 6 instances (notice that, given the orders of magnitude of the number of infected, the average is dominated by the only instance where the epidemy grows exponentially). We can conclude that household contacts are one of the preferred pathways for the disease spreading in a population, so that the intervention protocol in which all the cohabitants are quarantined at the same time can be more efficient in a realistic-scenario.

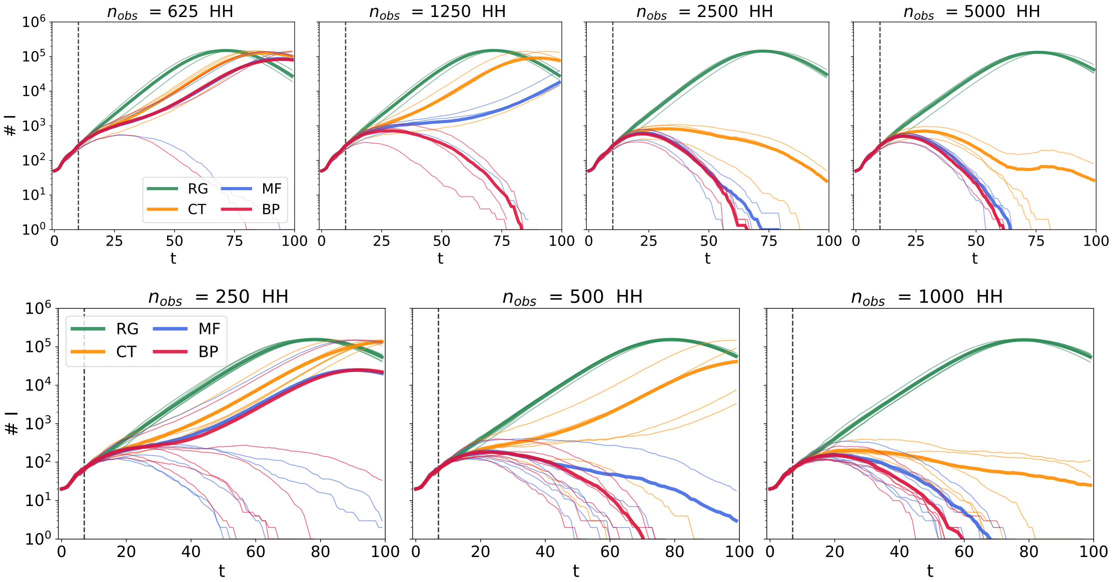

Figure 7.7: Effect of the mitigation strategies on the ABM dynamics when the whole household is quarantined once one of its members is tested positive. The top row refers to the scenario A of Figure 7.5 ($n_0 = 50$, $t_{start} = 10$, $n_{obs} \in \{625{,}1250{,}2500{,}5000\}$). The bottom row refers to the scenario B of Figure 7.6 ($n_0 = 20$, $t_{start} = 7$, $n_{obs} \in \{250{,}500{,}1000\}$). Each plot shows the number of infected individuals w.r.t. t. The thin lines represent a single instance, thick lines represent their average.

7.7.2 Effect of noisy observations

In this section, we present a set of results in a more realistic scenario, in which the observations are affected by some level of noise: indeed, test accuracy is not always 100% effective, and consequently false positive or false negative results might be detected. We assume for simplicity that the observation of a recovered individual is not affected by test inaccuracy, so that test noise will modify the probability of being Suscebtile/Infected. .

A non-zero false positive rate will put a small additional fraction of individuals in isolation, but it does not lead to deterioration of the epidemic control. For this reason, we only focus on the influence of false negatives and check how the performance depends on their rate, denoted in the following with ν_n: this represents a worse scenario as some individuals will continue to spread the disease, despite having been tested. The likelihood of observing a Susceptible or Infected node is modified as follows:

$$p_{O,i}\left[\left(x_i^t\right)^{obs} = S \mid x_i^t\right] = \mathbb{I}\left[x_i^t = S\right] + \nu_n \mathbb{I}\left[x_i^t = I\right] \tag{7.18}$$

$$p_{O,i}\left[\left(x_i^t\right)^{obs} = I \mid x_i^t\right] = (1-\nu_n)\,\mathbb{I}\left[x_i^t = I\right] \tag{7.19}$$

In principle, observations might include additional information, for instance about the individual's viral charge or the level of symptoms, but this is the simplest setting. The BP approach allows to easily included the effect of noisy observations, ad previously discussed; on the contrary, this information is not included in the MF scheme, to keep it as simple as possible and test its robustness. Figure 7.8 shows the results of the mitigation protocol for different values of FNR ($\nu_n \in [0.09{,}0.4]$), compatible with the observed accuracies of current clinical tests. Results are shown for a fixed number of daily observations for scenarios A and B, keeping the households quarantination protocol. CT controls the spreading up to $\nu_n = 0.19$ in scenario A, MF up $\nu_n = 0.19$ in both scenarios. BP is able to suppress the epidemy up to $\nu_n = 0.31$, its performance being remarkably robust in a wide range of FNR and almost unchanged w.r.t. the noiseless case.

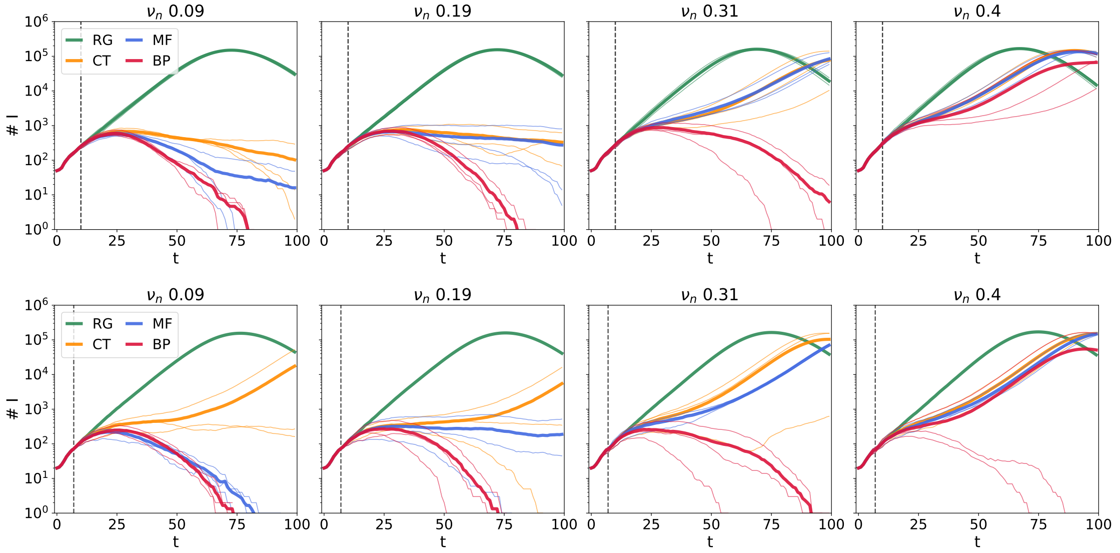

Figure 7.8: Effect of the mitigation strategies on the ABM dynamics in the presence of noisy observations. The top panel shows the scenario A ($n_0 = 50$, $t_{start} = 10$) with $n_{obs} = 2500$. The bottom panel refers to scenario B ($n_0 = 20$, $t_{start} = 7$) with $n_{obs} = 1000$. In both cases, the household quarantine protocol is implemented. Each plot shows the number of infected individuals w.r.t. t, for different values of the false negative rate $\nu_n \in \{0.09{,}0.19{,}0.31{,}0.4\}$. The thin lines represent a single instance, thick lines represent their average.

7.7.3 Effect of a partial app adoption

Finally, we test the robustness of the mitigation protocol when the contact network is not fully available to the inference methods. Indeed, we reasonably expect that a significant portion of the population will not use the contact tracing application, so that their contacts cannot be registered. In the simulations, this is obtained by hiding the daily contacts of a fraction of individuals, so that they are unknown to the inference algorithms. Let us denote with ρ^{AF} the fraction of individuals who use the app and whose contacts are tracked. Since in this way it is potentially impossible to reconstruct some infection cascades, in the BP scheme we introduce a small probability of self-infections at times $t > 0$, in order to avoid a plain incompatibility between the inference model and observations due to undetected transmissions. In all the simulations, we used $\gamma_i^t = \gamma = 10^{-4}$. Figure (7.9) shows a set of results of the intervention protocol for the two scenarios A and B, obtained for different values of the adoption fraction ranging between 0.6 and 0.9. In both cases, we adopted the households quarantine protocol and the maximum number of daily observations of Figure 7.5-7.6 (resp. $n^A_{obs} = 5000$, $n^B_{obs} = 1000$). Although the performance is severely affected for all the methods, one observes that even at AF $\in [0.6{,}0.7]$ inference algorithms allows to delay the spreading of the epidemic and to flatten the peak of infected individuals, way more efficiently than the standard contact tracing strategy. Furthermore, it should be noted that app utilization may be positively correlated to the number of contacts of individuals: for instance, elderly individuals who tipycally have less daily interactions are less likely to use the app. Including more detailed information about mobile application utilization in an age-stratified population might reduce the impact of low adoption.

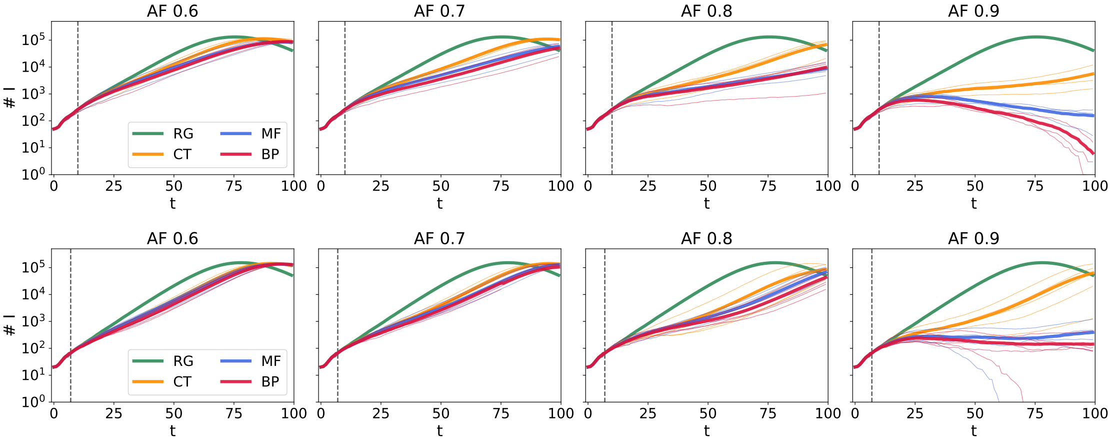

Figure 7.9: Effect of the mitigation strategies on the ABM dynamics in the presence of a partial app adoption. The top panel shows the scenario A ($n_0 = 50$, $t_{start} = 10$) with $n_{obs} = 5000$. The bottom panel refers to scenario B ($n_0 = 20$, $t_{start} = 7$) with $n_{obs} = 1000$. In both cases, the household quarantine protocol is implemented. Each plot shows the number of infected individuals w.r.t. t, for different values of the adoption fraction $\rho^{AF} \in \{0.6,0.7,0.8,0.9\}$. The thin lines represent a single instance, thick lines represent their average.

7.8 Discussion and outlook

In this Chapter we analyzed an inference-guided mitigation protocol for epidemic spreading processes from contact tracing data. Probabilistic inference allows to concatenate the information about contacts between individuals (registered through digital contact tracing applications) by using a prior description of the epidemic, additionally constrained by observations related to test results and/or symptoms appearance. The resulting posterior distribution is approximately computed using Belief Propagation or a simpler MF heuristic, so that individuals are ranked according to a certain probability of being infected at each day; quarantining the individuals with the highest risk (depending on given test capabilities) can result into a full epidemic suppression at its early stage, or eventually in a delay of the epidemic peak.

The results presented in the previous section show how these methods - in particular BP - allow to contain the disease spreading more efficiently than standard contact tracing implementations, in regimes where the epidemic is growing and the number of observations is relatively small compared to the population size, so to be compatible with real test capacity of many countries.

Belief Propagation has the best performances, because i) the approximation is more accurate than MF - and more suited on sparse topologies - in inferring the level of risk, and ii) the observations are correctly included in the posterior distribution. The robustness of BP against high levels of false negative rates (compatible with real tests capabilities) is one of the main advantages of such a Bayesian-inference guided mitigation. Another advantage of BP is that it can deal with asyncronous contact events, without a prescribed window, and with a non-markovian dynamics; the feature is essential to capture some specific properties observed in the SARS-CoV-2 transmission. On the other hand, Mean-Field heuristics still provides reasonably good risk prediction in several regimes (typically for larger number of daily observations) but it is less robust against test noise. The BP method further allows in principle to learn or adjust online the parameters used for the SIR inference (i.e. the infection probability and its time dependency and the recovery time), through an approximate maximum likelihood procedure [5]: we leave this point for future investigations.

An overall advantage of dealing with inference techniques is that each individual is given a probability of being infected in time: in principle, the information about its absolute value can be used to suggest individual-based actions, including reduction of contacts, self-isolation and testing, where each suggested action can be implemented if the infection risk exceeds some specific thresholds.

On a more computational standpoint, we remark that the volume of daily exchanged messages per pair of individuals in the two proposed methods is constant with respect to both the population size and time: in particular, for the BP approach this is achieved by defining a fixed time-window over which the inference step is carried out, as discussed in Section 7.5. A rough estimation of this volume gives about 1kB for MF and 1MB for BP per individual on each day (assuming ~ 10 daily contacts): this volume is negligible when compared with normal data usage, so that the computational load over the CPU's phone would remain reasonably small . With regard to privacy, it is worth emphasizing that the proposed inference methods are in principle more protective than the manual tracing. On one hand, both can be implemented in a fully distributed way using point-to-point cryptography without fully centralized processing and storage of information on infections or contacts. On the other hand, by identifying individuals who have the largest probability of being infected through a cumulative process by which information is integrated, the direct attribution of potential infection events to a given individual is made much harder. Details of such fully privacy preserving implementation, along the lines of [132], are left for future work.

One of the novelties introduced in this work is the mismatch between the model used to simulate the epidemic spreading of SARS-CoV-2 (the ABM [49]) and the one used for the inference. Therefore, even if the true disease dynamics is not known our results suggest that the epidemic spreading can be controlled in several regimes using a simpler parametrization of the dynamics, in terms of a SIR model.

The main drawback of the present approach resides on the knowledge of contact network: indeed, when a fraction of the population does not adopt the tracing application, none of their contacts can be detected, the risk inference becomes less effective and so does the mitigation performances (see Figure 7.9). Despite a delay of the epidemic peak can be observed even at smaller values of the adoption fraction, currently deployed applications are still far from reaching a significant level of adoption for these methods to be effective in containing the epidemic spreading.

Finally, we remark that despite the agent-based model used to simulate the epidemic spreading is very detailed from the point of view of the infection dynamics, it exploits a synthetic-generated contact network. This choice can be still considered reasonably realistic since the multi-layer structure allows to distinguish between different interaction domains (households, workplace contacts, and random events). Another potential drawback is that contact duration and distance between individuals are not taken into account within the ABM, as interactions are considered instantaneous: this assumption is consistent with an extremely simplified (and thus, more privacy preserving) contact tracing implementation, where the information acquired via Bluetooth is binary in nature: namely, two individuals have been in contact or not. However, since contact tracing protocols currently allow to register also the contact duration and, indirectly, the distance between the two individuals (that depends on the Bluetooth signal strength [28]), such additional information could in principle be exploited by the inference algorithms, e.g. in order to better determine the transmission probability.

As an example, the infection model developed by Hinch et al. [114] simulates the mobility of an age-stratified population on a closed region (e.g. a city), where each individual can interact with others in different aggregation points (e.g. schools, offices, supermarkets, other social places): the mobility simulation is based on real geographic data that take into account the distribution of both the population and aggregation points inside a city. For the above reasons, future investigation is needed and will be carried out in order to validate such inference-guided intervention strategies on different models that do take into account the duration of the contacts and exploit more realistic contact networks, as well as on real data.

Appendix A

Mixed tilted moments

In this appendix, we discuss a simple procedure to compute the "mixed" tilted moments needed to compute the stationary points of the DC free energy in Section 3.4.2. The same reasoning can be also applied to the EP free energy, as discussed in Section 2.4.2.

A.1 General approach

Suppose to have a multivariate probability distribution of N variables, expressed as the product of a Gaussian density - denoted here with $\rho\left(\boldsymbol{x}\right)$ - times a non-negative function f_a, depending only on a subset $\boldsymbol{x}_a = \left\{x_k, k \in \partial a\right\}$ of variables:

$$\zeta^{(a)}\left(\boldsymbol{x}\right) = \frac{1}{Z_a}\rho\left(\boldsymbol{x}\right) f_a\left(\boldsymbol{x}_a\right) \tag{A.1}$$

The above expression generalizes both the tilted distributions defined in Eq. (3.12) when $f_a = \Psi_a$, $\rho = g^{\setminus a}$ and the Gaussian cavities (3.16) defined in (3.16) for $\rho = q$, $f_a = \left(\phi_a\right)^{-1}$. We are interested in computing the moments of an arbitrary variable x_i not directely connected to factor node a:

$$\langle x_i^{\alpha} \rangle_{\zeta^{(a)}} \qquad \alpha = 1{,}2;\ \forall i \notin \partial a \tag{A.2}$$

In the following, we assume to know the moments of $\rho\left(\boldsymbol{x}\right)$, given by

$$\langle \boldsymbol{x} \rangle_{\rho} = \boldsymbol{\nu} \tag{A.3}$$

$$\langle \boldsymbol{x}\boldsymbol{x}^t \rangle_{\rho} = \boldsymbol{\Sigma} + \boldsymbol{\nu}\boldsymbol{\nu}^t \tag{A.4}$$

as well as the first moments of all the neighbours of a w.r.t. $\zeta^{(a)}$, namely $\left\{\langle x_j \rangle_{\zeta^{(a)}}\right\}_{j \in \partial a}$. Since node i appears in (A.1) only inside a Gaussian term, we can exploit Gaussian integration properties to write (A.2) as a function of the latter two aforementioned quantities:

$$\langle x_i^{\alpha} \rangle_{\zeta^{(a)}} = \Xi^{(\alpha)}\left(\langle \boldsymbol{x} \rangle_{\rho}, \langle \boldsymbol{x}\boldsymbol{x}^t \rangle_{\rho}, \left\{\langle x_j \rangle_{\zeta^{(a)}}\right\}_{j \in \partial a}\right) \qquad \alpha = 1{,}2 \tag{A.5}$$

The knowlegde of (A.3)-(A.4) allows to marginalize (A.1) with respect to all the variables but $x_i \cup \boldsymbol{x}_a$:

$$\zeta^{(a)}\left(\boldsymbol{x}_a, x_i\right) \propto \int d\boldsymbol{x}_{\setminus(i \cup \partial a)} \zeta^{(a)}\left(\boldsymbol{x}\right) \propto \rho\left(\boldsymbol{x}_a, x_i\right) f_a\left(\boldsymbol{x}_a\right) \tag{A.6}$$

In particular, the first and second moments of $\rho\left(\boldsymbol{x}_a, x_i\right)$ are given by the blocks of (A.3)- (A.4) on the $[i \cup \partial a]$ indeces, respectively. Let us define the matrix $\mathcal{S}^{(\partial a \cup i)} \hat{=} \left(\boldsymbol{\Sigma}_{[\partial a \cup i, \partial a \cup i]}\right)^{-1} \in \mathbb{R}^{(|\partial a|+1)\times(|\partial a|+1)}$

and isolate all the terms that depend on x_i:

$$\mathcal{S}^{(\partial a \cup i)} = \left(\boldsymbol{\Sigma}_{[\partial a \cup i, \partial a \cup i]}\right)^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{[\partial a, \partial a]} & \boldsymbol{\Sigma}_{[\partial a, i]} \\ \boldsymbol{\Sigma}_{[i, \partial a]} & \Sigma_{ii} \end{bmatrix}^{-1} = \begin{bmatrix} S_{[\partial a, \partial a]} & S_{[\partial a, i]} \\ S_{[i, \partial a]} & S_{ii} \end{bmatrix} \tag{A.7}$$

where, at the right hand sides the superscript $(\partial a \cup i)$ is dropped for simplicity. By construction, $S_{[i,\partial a]} = \{S_{ik}\}_{k \in \partial a}$ and $S_{[\partial a, i]} = S^t_{[i,\partial a]}$. After some manipulations, we rewrite the Gaussian $\rho\left(\boldsymbol{x}_a, x_i\right)$ as

$$\rho\left(\boldsymbol{x}_a, x_i\right) = \theta\left(\boldsymbol{x}_a\right) \eta\left(\boldsymbol{x}_a, x_i\right) \tag{A.8}$$

where

$$\theta\left(\boldsymbol{x}_a\right) = \exp\left[-\frac{1}{2}\left(\boldsymbol{x}_{[\partial a]} - \boldsymbol{\nu}_{[\partial a]}\right)^t S_{[\partial a, \partial a]}\left(\boldsymbol{x}_{[\partial a]} - \boldsymbol{\nu}_{[\partial a]}\right) - \frac{1}{2} S_{ii} \nu_i^2 + \nu_i \sum_{k \in \partial a} S_{ik}\left(x_k - \nu_k\right)\right] \tag{A.9}$$

$$\eta\left(\boldsymbol{x}_a, x_i\right) = \exp\left\{-\frac{1}{2} S_{ii} x_i^2 + x_i\left[S_{ii}\nu_i - \sum_{k \in \partial a} S_{ik}\left(x_k - \nu_k\right)\right]\right\} \tag{A.10}$$

The two moments we are interested in (A.2) can be computed by first integrating over x_i and then w.r.t. the other variables $\boldsymbol{x}_a$:

$$\langle x_i^\alpha \rangle_{\zeta^{(a)}} = \frac{\int d\boldsymbol{x}_a \theta\left(\boldsymbol{x}_a\right) f_a\left(\boldsymbol{x}_a\right) \int dx_i \eta\left(x_i, \boldsymbol{x}_a\right) x_i^\alpha}{\int d\boldsymbol{x}_a \theta\left(\boldsymbol{x}_a\right) f_a\left(\boldsymbol{x}_a\right) \int dx_i \eta\left(x_i, \boldsymbol{x}_a\right)} \qquad \alpha = 1{,}2 \tag{A.11}$$

The inner integral over x_i can be computed using $1-$dimensional Gaussian integrals, leading to

$$\int dx_i \eta\left(x_i, \boldsymbol{x}_a\right) x_i^\alpha \hat{=} \mathcal{I}_i\left(\boldsymbol{x}_a\right) \times \xi_i^{(\alpha)}\left(\boldsymbol{x}_a\right) \qquad \alpha = 0{,}1{,}2 \tag{A.12}$$

with

$$\mathcal{I}_i\left(\boldsymbol{x}_a\right) = \sqrt{\frac{\pi}{2} S_{ii}} \exp\left\{\frac{1}{2S_{ii}}\left[S_{ii}\nu_i - \sum_{k \in \partial a} S_{ik}\left(x_k - \nu_k\right)\right]\right\} \tag{A.13}$$

$$\xi_i^{(1)}\left(\boldsymbol{x}_a\right) = 1 \tag{A.14}$$

$$\xi_i^{(1)}\left(\boldsymbol{x}_a\right) = \nu_i - \sum_{k \in \partial a} \frac{S_{ik}}{S_{ii}}\left(x_k - \nu_k\right) \tag{A.15}$$

$$\xi_i^{(2)}\left(\boldsymbol{x}_a\right) = \frac{1}{S_{ii}} + \frac{\left[S_{ii}\nu_i - \sum_{k \in \partial a} S_{ik}\left(x_k - \nu_k\right)\right]^2}{S_{ii}^2} \tag{A.16}$$

Notice now that the product $\theta\left(\boldsymbol{x}_a\right) f_a\left(\boldsymbol{x}_a\right) \mathcal{I}\left(\boldsymbol{x}_a\right)$ is nothing but the marginal of $\zeta^{(a)}$ over the $\boldsymbol{x}_a$, apart from a normalization constant that gets cancelled out with the denominator in (A.11). Therefore, the moments of i can be expressed as combination the moments w.r.t. ρ and moments of $\boldsymbol{x}_a$ under the same distribution $\zeta^{(a)}$:

$$\langle x_i^\alpha \rangle_{\zeta^{(a)}} = \langle \xi_i^{(\alpha)}\left(\boldsymbol{x}_a\right) \rangle_{\zeta^{(a)}} \quad \alpha = 1{,}2 \tag{A.17}$$

A same reasoning can be applied to compute mixed moments of *two* nodes i, j: when either $i \in \partial a$ or $j \in \partial a$ the computation of $\langle x_i x_j \rangle_{\xi^{(a)}}$ is equivalent to what discussed so far; on the other hand, when both $i, j \notin \partial a$ one should extend the above reasoning by using 2-dimensional Gaussian integrals.

A.2 Tilted moments

We apply now the above formula to compute the moments of x_i for a tilted distribution $q^{(a)}$, defined over a factor node a such that $i \notin \partial a$, whose density reads (see Eq. (3.14)):

$$q^{(a)}\left(\boldsymbol{x}\right) \propto g^{\backslash a}\left(\boldsymbol{x}\right)\Psi_a\left(\boldsymbol{x}_a\right) \tag{A.18}$$

Using (A.17) with $\rho = g^{\backslash a}$ and $f_a = \Psi_a$, we can write the moments of node $i \notin \partial a$ as:

$$\langle x_i^{\alpha}\rangle_{q^{(a)}} = \langle \xi_i^{(\alpha)}\left(\boldsymbol{x}_a\right)\rangle_{q^{(a)}} \quad \alpha = 1{,}2 \tag{A.19}$$

where

$$\xi_i^{(1)}\left(\boldsymbol{x}_a\right) = \mu_i^{\backslash a} - \sum_{k\in\partial a}\frac{S_{ik}^{\backslash a}}{S_{ii}^{\backslash a}}\left(x_k - \mu_k^{\backslash a}\right) \tag{A.20}$$

$$\xi_i^{(2)}\left(\boldsymbol{x}_a\right) = \frac{1}{S_{ii}^{\backslash a}} + \frac{\left[S_{ii}^{\backslash a}\mu_i^{\backslash a} - \sum_{k\in\partial a}S_{ik}^{\backslash a}\left(x_k - \mu_k^{\backslash a}\right)\right]^2}{\left(S_{ii}^{\backslash a}\right)^2} \tag{A.21}$$

In this notation, $\mu_k^{\backslash a} = \langle x_k\rangle_{g^{\backslash a}}$ and the quantities $\left\{S_{kl}^{\backslash a}\right\}$ are the elements of the inverse sub-block of the *cavity* correlation matrix, i.e $\left[\mathcal{S}^{\backslash a}\right]^{(\partial a\cup i)} = \left(\boldsymbol{\Sigma}^{\backslash a}_{[\partial a\cup i,\partial a\cup i]}\right)^{-1}$. Therefore, the above expressions depend on the cavity moments computed on variables that do not belong to the neighborhood of a, so that we can further simplify (A.19) by applying the same reasoning to the cavity moments $\mu_i^{\backslash a}$, as shown in the next section.

A.2.1 Cavity moments

Let us recall the definition of the cavity distribution $g^{\backslash a}\left(\boldsymbol{x}\right)$ as in Eq. (3.16):

$$g^{\backslash a}\left(\boldsymbol{x}\right) \propto q\left(\boldsymbol{x}\right)\frac{1}{\phi_a\left(\boldsymbol{x}_a\right)}$$

Applying the same reasoning as before with $\rho = q$ and $f_a = \left(\phi_a\right)^{-1}$, we get:

$$\langle x_i\rangle_{g^{\backslash a}} \hat{=} \mu_i^{\backslash a} = \mu_i - \sum_{k\in\partial a}\frac{S_{ik}}{S_{ii}}\left(\mu_k^{\backslash a} - \mu_k\right) \tag{A.22}$$

where now $\mu_k = \langle x_k\rangle_q$ and the elements S_{ii},$\{S_{ik}\}_{k\in\partial a}$ corresponds to the inverse sub-block of the full Gaussian measure q, namely $\mathcal{S}^{(\partial a\cup i)} = \left[\boldsymbol{\Sigma}_{[\partial a\cup i,\partial a\cup i]}\right]^{-1}$. Notice now that, by construction of the cavity distribution (3.16):

$$\left[\mathcal{S}^{\backslash a}\right]^{(\partial a\cup i)} = \left(\boldsymbol{\Sigma}^{\backslash a}_{[\partial a\cup i,\partial a\cup i]}\right)^{-1} = S^{(\partial a\cup i)} - \begin{bmatrix}\boldsymbol{\Gamma}^{(a)} & 0\\ 0 & 0\end{bmatrix} \tag{A.23}$$

where $\boldsymbol{\Gamma}^{(a)}$ is the inverse covariance matrix of the approximate factor ϕ_a defined in Eq. 3.3. As a consequence, all the elements of the type $\{S_{ik}\}_{k\in\partial a}$ - i.e. the only ones appearing in (A.22)-(A.19) - are equal in the two matrices. Using this result and combining (A.22) with (A.19), we get the final expression for the first "mixed" tilted moment $\langle x_i \rangle_{q^{(a)}}$:

$$\begin{aligned}
\langle x_i \rangle_{q^{(a)}} &= \mu_i^{\setminus a} - \sum_{k \in \partial a} \frac{S_{ik}^{\setminus a}}{S_{ii}^{\setminus a}} \left(\langle x_k \rangle_{q^{(a)}} - \mu_k^{\setminus a} \right) \\
&= \mu_i^{\setminus a} - \sum_{k \in \partial a} \frac{S_{ik}}{S_{ii}} \left(\langle x_k \rangle_{q^{(a)}} - \mu_k^{\setminus a} \right) \\
&= \mu_i - \sum_{k \in \partial a} \frac{S_{ik}}{S_{ii}} \left(\mu_k^{\setminus a} - \mu_k \right) - \sum_{k \in \partial a} \frac{S_{ik}}{S_{ii}} \left[\langle x_k \rangle_{q^{(a)}} - \mu_k^{\setminus a} \right] \\
&= \mu_i - \sum_{k \in \partial a} \frac{S_{ik}}{S_{ii}} \left(\langle x_k \rangle_{q^{(a)}} - \mu_k \right) \qquad \text{(A.24)}
\end{aligned}$$

which is equal to (3.87). The formula (3.88) in section 3.4.2 is derived in the same way using the above reasoning with $\rho = q$ and $f_a \equiv f_i = \Delta_i$. We also report the final expression for the second moment $\langle x_i^2 \rangle_{q^{(a)}}$, that will be used in the following section:

$$\langle x_i^2 \rangle_{q^{(a)}} = \frac{1}{S_{ii}} + \left\langle \left[\mu_i - \sum_{k \in \partial a} \frac{S_{ik}}{S_{ii}} \left(x_k - \mu_k \right) \right]^2 \right\rangle_{q^{(a)}} \qquad \text{(A.25)}$$

A.3 Stationary points of (univariate) EP free energy

In Section (2.4.2) we have seen how to derive an expression for the EP free energy. Its stationary points (2.77)-(2.78) depend on the mixed tilted moments $\langle x_i^\alpha \rangle_{q^{(j)}}$, i.e. the expectation values of node i w.r.t. a tilted distribution defined on $j \neq i$. Therefore, we can apply the procedure discussed so far to explicitly compute them. In particular, starting from (A.24)-(A.25) and using $a \equiv j$, the mixed tilted moments are given by:

$$\langle x_i \rangle_{q^{(j)}} = \mu_i + \frac{\Sigma_{ij}}{\Sigma_{jj}} \left[\langle x_j \rangle_{q^{(j)}} - \mu_j \right] \qquad \text{(A.26)}$$

$$\langle x_i^2 \rangle_{q^{(j)}} = \Sigma_{ii} + \mu_i^2 + 2\mu_i \frac{\Sigma_{ij}}{\Sigma_{jj}} \left[\langle x_j \rangle_{q^{(j)}} - \mu_j \right] + \frac{\Sigma_{ij}^2}{\Sigma_{jj}^2} \left[\langle x_j^2 \rangle_{q^{(j)}} - 2 \langle x_j \rangle_{q^{(j)}} \mu_j + \mu_j^2 - \Sigma_{jj} \right] \qquad \text{(A.27)}$$

where the summation $\sum_{k \in \partial a}$ here reduces to a single contribution for $k \equiv j$, and we also used the explicit expression for the 2×2 matrix $\mathcal{S}^{(i \cup j)}$:

$$\mathcal{S}^{[j \cup i]} = \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ij} & \Sigma_{jj} \end{bmatrix}^{-1} = \frac{1}{\Sigma_{ii}\Sigma_{jj} - \Sigma_{ij}^2} \begin{bmatrix} \Sigma_{jj} & -\Sigma_{ij} \\ -\Sigma_{ij} & \Sigma_{ii} \end{bmatrix}. \qquad \text{(A.28)}$$

By inserting (A.26)-(A.27) into the stationary conditions for the free energy (2.77)-(2.78) and simplifying, we finally get:

$$0 = \sum_{j \neq i} \frac{\Sigma_{ij}}{\Sigma_{jj}} \left[\langle x_j \rangle_{q^{(j)}} - \mu_j \right] \qquad \text{(A.29)}$$

$$0 = 2\mu_i \sum_{j \neq i} \frac{\Sigma_{ij}}{\Sigma_{jj}} \left[\langle x_j \rangle_{q^{(j)}} - \mu_j \right] + \sum_{j \neq i} \frac{\Sigma_{ij}^2}{\Sigma_{jj}^2} \left[\langle x_j^2 \rangle_{q^{(j)}} - 2 \langle x_j \rangle_{q^{(j)}} \mu_j + \mu_j^2 - \Sigma_{jj} \right] \qquad \text{(A.30)}$$

which is satisfied by the moment matching conditions

$$\langle x_i \rangle_{q^{(i)}} = \mu_i \qquad \forall i \qquad \text{(A.31)}$$

$$\langle x_i^2 \rangle_{q^{(i)}} = \Sigma_{ii} + \mu_i^2 \qquad \forall i \qquad \text{(A.32)}$$

Appendix B

Diagonalization of the Adjacency Matrix of hypercubic lattices

From a graph-theoretical perspective, a hypercubic lattice in $d > 1$ dimensions can be regarded as the cartesian product d linear-chains, one for each dimension [120]. As a consequence, the adiacency matrix of the d-dimensional lattice (denoted in Chapter 4 with $\mathcal{A}^{(d)}$) can be easily expressed in terms of the adiacency matrices of the single linear chains, by means of the Kronecker product (denoted with $\otimes$) [72]:

$$\mathcal{A}^{(d)} = \sum_{k=1}^{d} \mathcal{P}^{(k)} \left[\mathcal{A}^{(1)} \otimes \underbrace{\mathbb{I}_L \otimes ... \otimes \mathbb{I}_L}_{d-1} \right] \tag{B.1}$$

where $\mathcal{P}^{(k)}$ is a permutation operator and $\mathbb{I}_L$ is the identity matrix of size L. Alternatively, the above expression can be re-phrased as a recursive relation for $d > 1$:

$$\mathcal{A}^{(d)} = \mathcal{A}^{(d-1)} \otimes \mathbb{I}_L + \mathbb{I}_{L^{d-1}} \otimes \mathcal{A}^{(1)} \qquad d > 1 \tag{B.2}$$

In Eqs. (B.1) and (B.2) the quantity $\mathcal{A}^{(1)}$ is the adiacency matrix of a closed linear chain of size L:

$$\mathcal{A}^{(1)} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \tag{B.3}$$

Eqs. (B.1) or (B.2) allow to compute the spectral decomposition of $\mathcal{A}^{(d)}$ by knowing the spectrum of the adiacency matrix of the linear chain. The matrix $\mathcal{A}^{(1)}$ is a simple case of a circulant matrix [60] and it can be diagonalized exactly. Its eigenvalues and eigenvectors are shown below:

$$\lambda_x^{(1)} = 2\cos\left(\frac{2\pi}{L}x\right) \tag{B.4}$$

$$\boldsymbol{\nu}_x^{(1)} = \frac{1}{\sqrt{L}}\left(1, w_x, w_x^2, \ldots, w_x^{L-1}\right)^t \qquad w_x = e^{i\frac{2\pi}{L}x} \tag{B.5}$$

where i is the imaginary unit and $x \in \{0, ..., L-1\}$ is an integer index referring to the eigenvalue (or eigenvector), that can be thought as a site coordinate on the linear chain.
Exploiting the properties of the Kronecker product [72], it is possible to compute the eigenspectrum of $\mathcal{A}^{(d)}$. In the following, we denote respectively with $\lambda^{(d)}$ and $\boldsymbol{\nu}^{(d)}$ the eigenvalues and the eigenvectors of $\mathcal{A}^{(d)}$, given by:

$$\lambda^{(d)}_{(x_1,\ldots,x_d)} = \sum_{k=1}^{d} \lambda^{(1)}_{x_k} = 2\sum_{k=1}^{d} \cos\left(\frac{2\pi}{L} x_k\right) \tag{B.6}$$

$$\boldsymbol{v}^{(d)}_{(x_1,\ldots x_d)} = \otimes_{k=1}^{d} \boldsymbol{v}^{(1)}_{x_k} \tag{B.7}$$

In this notation, the vector $\vec{x} = (x_1, \ldots, x_d)$ can be considered as a vectorial index that identifies a certain site in the hypercubic lattice with coordinate $\vec{x}$. The precision matrix of the full Gaussian measure (4.37), denoted in Chapter (4) with $\mathcal{K}^{(d)}$ and defined by (4.38), is nothing but a linear combination between the identity matrix and $\mathcal{A}^{(d)}$. Therefore, the expression of its eigenvalues follows directly from (B.6):

$$\lambda_{\vec{x}} = 2d\Gamma_0 + 2\Gamma_1 \sum_{k=1}^{d} \cos\left(\frac{2\pi}{L} x_k\right) \tag{B.8}$$

where the parameters Γ_0, Γ_1 are defined in (4.36). From now on, we refer to $\lambda_{\vec{x}}$ as the eigenvalues of the precision matrix $\mathcal{K}^{(d)}$; the set of its eigenvectors is the same as (B.7) since an identity matrix always commutes with $\mathcal{A}^{(d)}$. In this way, we can exploit the known eigenspectrum to compute any element of the inverse matrix of $\mathcal{K}^{(d)}$, that holds at any size L. In particular, by denoting with $\mathbb{U}$ the matrix obtained by stacking column-wise all the eigenvectors, and using the vectorial notation of indeces $(\vec{x}, \vec{x}')$ to indicate a generic element of the matrix, we get:

$$\begin{aligned}
\left[\left(\mathcal{K}^{(d)}\right)^{-1}\right]_{(\vec{x},\vec{x}')} &= \sum_{\vec{y},\vec{y}' \in \{0,\ldots,L-1\}^d} \mathbb{U}_{(\vec{x},\vec{y})} \left[\Lambda^{-1}\right]_{(\vec{y},\vec{y}')} \mathbb{U}^{\dagger}_{(\vec{y}',\vec{x}')} \\
&= \sum_{\vec{y},\vec{y}' \in \{0,\ldots,L-1\}^d} \mathbb{U}_{(\vec{x},\vec{y})} \frac{\delta_{(\vec{y},\vec{y}')}}{\lambda_{\vec{y}}} \mathbb{U}^{\dagger}_{(\vec{y}',\vec{x}')} \\
&= \frac{1}{L^d} \sum_{\vec{y} \in \{0,\ldots,L-1\}^d} \frac{\exp\left[i\frac{2\pi}{L}\sum_{k=1}^{d} y_k \left(x_k - x'_k\right)\right]}{\lambda_{\vec{y}}}
\end{aligned} \tag{B.9}$$

where Λ is the diagonal matrix of the eigenvalues, i.e. $\Lambda_{(\vec{x},\vec{y})} = \delta_{(\vec{x},\vec{y})}\lambda_{\vec{x}}$, and in the last line we use the explicit expression of the elements of $\mathbb{U}$ that follows from Eqs. (B.7)-(B.5):

$$\mathbb{U}_{(\vec{x},\vec{y})} = \frac{1}{L^{d/2}} \exp\left[i\frac{2\pi}{L}\sum_{k=1}^{d} x_k y_k\right] \tag{B.10}$$

As discussed in Chapter 4, in order to apply Density Consistency to the homogeneous Ising model, we are interested in computing only two types of elements: the diagonal terms (where $\vec{x} = \vec{x}'$), denoted with Σ_0, and the ones corresponding to nearest neighbours spins, denoted with Σ_1. The latters correspond to indeces $\vec{x}, \vec{x}'$ that differ only on one direction, namely $(\vec{x} - \vec{x}')_k = \pm\delta_{kk^*}$ for a certain $k^* \in \{1, \ldots, d\}$. At fixed size L, their expressions read:

$$\Sigma_0 = \left[\left(\mathcal{K}^{(d)}\right)^{-1}\right]_{(\vec{x},\vec{x})} = \frac{1}{L^d} \sum_{\vec{y} \in \{0,\ldots,L-1\}^d} \frac{1}{\lambda_{\vec{y}}} \tag{B.11}$$

$$\Sigma_1 = \left[\left(\mathcal{K}^{(d)}\right)^{-1}\right]_{(\vec{x},\vec{x}')} = \frac{1}{L^d} \sum_{\vec{y} \in \{0,\ldots,L-1\}^d} \frac{1}{\lambda_{\vec{y}}} \exp\left[\pm i\frac{2\pi}{L} y_{k^*}\right] \qquad k^* \in \{1, \ldots, d\} \tag{B.12}$$

Eqs (B.11)-(B.12) can be used to apply Density Consistency for a homogenous Ising model at finite size L as in Section 4.3. However, we are mainly interested in taking the thermodynamic limit $L \to \infty$: in this case, by defining a continuous set of coordinates $s_k = \frac{2\pi x_k}{L} \in [0{,}2\pi)\,, k \in \{1, \ldots, d\}$ and replacing the summations with a multidimensional integral, we get:

$$\Sigma_0 = \int_0^{2\pi} \prod_{k=1}^{d} \frac{ds_k}{2\pi} \frac{1}{2d\Gamma_0 + 2\Gamma_1 \sum_{k=1}^{d} \cos\, s_k} \tag{B.13}$$

$$\Sigma_1 = \int_0^{2\pi} \prod_{k=1}^{d} \frac{ds_k}{2\pi} \frac{\exp\left[\pm s_{k^*}\right]}{2d\Gamma_0 + 2\Gamma_1 \sum_{k=1}^{d} \cos\, s_k} \tag{B.14}$$

The integrals (B.13)-(B.14) can be easily performed by using the Laplace representation of the Heaviside step function $1/x = \int_0^\infty e^{-tx} dt$. We report below their final expression, valid for any $d \geq 1$:

$$\Sigma_0 = \frac{1}{\Gamma_0} R_d\,(r) \tag{B.15}$$

$$\Sigma_1 = \frac{1}{r\Gamma_0} \left[\frac{1}{2d} - R_d\,(r)\right] \tag{B.16}$$

where $r = \Gamma_1/\Gamma_0$ and the function $R_d\,(r)$ is given by:

$$R_d\,(r) = \frac{1}{2} \int_0^\infty dt \left[e^{-t} \mathcal{I}_0\,(rt)\right]^d \tag{B.17}$$

with $\mathcal{I}_0$ being the modified Bessel function of the first kind of order 0:

$$\mathcal{I}_0\,(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{x \cos\theta} d\theta \tag{B.18}$$

The function $R_d\,(r)$ is strictly related to the Lattice Green Functions (LGFs) [61], i.e. the probability generating function of a random walk on the lattice. In particular, by a simple inspection of $R_d\,(r)$, we find that the quantity $1 - 1/\left[2dR_d\,(-1)\right]$ is nothing but the return probability of a symmetric random walk on the infinite d-dimensional lattice. The integral (B.17) can be analytically computed for $d \leq 3$ (in particular, we refer to [69] for the $d = 3$ case) and we report their expressions below:

$$R_1\,(r) = \frac{1}{2\sqrt{1 - r^2}} \tag{B.19}$$

$$R_2\,(r) = \frac{K\left(r^2\right)}{2\pi} \tag{B.20}$$

$$R_3\,(r) = \frac{2}{3\pi^2} \sqrt{\left(1 - \frac{3}{4}a\right) \frac{K\left(l_+^2\right) K\left(l_-^2\right)}{1 - a}} \tag{B.21}$$

where

$$l_\pm^2 = \frac{1}{2} \pm \frac{1}{4} q\sqrt{4 - q} - \frac{2 - \eta}{4} \sqrt{1 - q}$$
$$q = a/\left(a - 1\right)$$
$$a\,(r) = \frac{1}{2} + \frac{r^2}{6} - \frac{1}{2}\sqrt{1 - r^2}\sqrt{1 - \frac{r^2}{9}}$$

and $K(m)$ denotes the complete elliptic integral of the first kind with modulus m [59].
Finally, we report below the series expansion of $R_d(-1)$ in the infinite dimensional limit $d \to \infty$, that is used in Section 4.3.6 to derive the corresponding expansion for the Density Consistency's critical temperature:

$$R_d(-1) = \frac{1}{2}d^{-1} + \frac{1}{4}d^{-2} + \frac{3}{8}d^{-3} + \frac{3}{4}d^{-4} + \frac{15}{8}d^{-5} + \frac{355}{64}d^{-6} + \frac{595}{32}d^{-7} + O\left(d^{-8}\right) \tag{B.22}$$

Appendix C

Computation of β_m and stability of fixed points

In this appendix, we brefly discuss how to compute the minimum of the ferromagnetic solution to the homogeneous Ising model within DC approximation, presented in Sec. (4.3), denoted with β_m. The fixed points equations (given by (4.52)-(4.53)) are solved at fixed r with respect to the magnetization m and the inverse temperature β. Therefore, (4.52)-(4.53) implicitly define a function $m(r)$ such that $M(m(r), r) = m$ and $\beta(r) = B(m(r), r)$. The quantity β_m can be computed by finding the point $m^* = m(r^*)$ such that $\frac{d\beta}{dr}(m(r^*), r^*)\Big|_{r^*} = 0$, so that $\beta_m = B(m^*, r^*)$. Taking the total derivative of $B(m(r), r)$ we get the following equation to be solved:

$$0 = \frac{d\beta}{dr} = \frac{\partial B}{\partial r} + \frac{\partial B}{\partial m}\frac{dm}{dr} \tag{C.1}$$

To compute $\frac{dm}{dr}$ we use its implicit definition given by M:

$$\begin{aligned} 0 &= \frac{d}{dr}\left[M(m(r), r) - m(r)\right] \\ &= \frac{\partial M}{\partial m}(m(r), r) + \frac{\partial M}{\partial r}(m(r), r) - \frac{dm}{dr} \\ &= \left[\frac{\partial M}{\partial m}(m(r), r) - 1\right]\frac{dm}{dr} + \frac{\partial M}{\partial r}(m(r), r) \end{aligned} \tag{C.2}$$

which gives

$$\frac{dm}{dr} = \frac{\frac{\partial M}{\partial r}(m(r), r)}{1 - \frac{\partial M}{\partial m}(m(r), r)} \tag{C.3}$$

By plugging (C.3) into (C.1) and using the fixed point equation for the magnetization, we get a 2×2 system to be solved w.r.t. m, r:

$$M(m, r) - m = 0 \tag{C.4}$$

$$\frac{\partial B}{\partial r}(m, r)\left(1 - \frac{\partial M}{\partial m}(m, r)\right) + \frac{\partial M}{\partial r}(m, r)\frac{\partial B}{\partial m}(m, r) = 0 \tag{C.5}$$

C.1 Stability

The stability of a fixed point $m^* = m\left(r^*\right)$ can be analyzed by computing $\left.\frac{dM}{dm}\right|_{m^*}$. In particular, starting from the system (4.52)-(4.53), the instability occurs when $\left.\frac{dM}{dm}\right|_{m^*} = 1$. Writing the original system where r is implicitly defined as $r = \mathcal{R}\left(\beta, m\right)$, we get $m = M\left(m, \mathcal{R}\left(\beta, m\right)\right)$ and $\beta = B\left(m, \mathcal{R}\left(\beta, m\right)\right)$. The equation we want to solve is

$$1 = \frac{dM}{dm} = \frac{\partial M}{\partial m} + \frac{\partial M}{\partial r}\frac{\partial \mathcal{R}}{\partial m} \tag{C.6}$$

To compute $\frac{\partial \mathcal{R}}{\partial m}$ we can use again its implicit definition. Starting from the fixed point equation for the temperature, $\beta = B\left(m, \mathcal{R}\left(\beta, m\right)\right)$, we get:

$$\begin{aligned} 0 = \frac{dB}{dm} &= \frac{\partial B}{\partial m} + \frac{\partial B}{\partial r}\frac{\partial \mathcal{R}}{\partial m} \\ \frac{\partial \mathcal{R}}{\partial m} &= -\frac{\partial B}{\partial m}\Big/\frac{\partial B}{\partial r} \end{aligned} \tag{C.7}$$

Putting together (C.7) and the fixed point equation for the magnetization we get the following system of equations to be solved w.r.t r, m:

$$M\left(m, r\right) - m = 0 \tag{C.8}$$

$$\frac{\partial M}{\partial m} - \frac{\partial M}{\partial r}\frac{\frac{\partial B}{\partial m}}{\frac{\partial B}{\partial r}} = 1 \tag{C.9}$$

For $d \geq 3$, the paramagnetic solution is stable in the full range $r \in [-1,0]$, i.e $\beta \in [0, \beta_p]$, while the ferromagnetic solution becomes unstable exactly at the point (r_m, β_m) computed through C.4-C.5.

Appendix D

Mean-Field heuristiscs on the SIR model

The Mean-field method for the inference in the SIR model introduced in Chapter 7 can be derived as a limiting case of the dynamical message passing (DMP) framework, developed in [85, 86], in the limit where the transmission probabilities are small. Moreover, by construction it can be only applied for a Markovian evolution. The first assumption is therefore $p\left(x_i^{t+1} \mid \boldsymbol{x}^0, \ldots, \boldsymbol{x}^t\right) = p\left(x_i^{t+1} \mid \boldsymbol{x}^t\right)$ in the SIR dynamics (7.2), that is valid when the recovery rate and the transmission probabilities do not depend on the time since infection. MF assumes that the joint probability distribution at each time can be factorized over nodes, namely $p\left(\mathbf{x}^t\right) \approx \prod_i p\left(x_i^t\right) \forall t \in [0, T]$. This allows to rewrite the marginal probability of the individual i's state at time t as:

$$p\left(x_i^{t+1}\right) \approx \sum_{x_i^t, \boldsymbol{x}_{\partial i}^t} p\left(x_i^t \mid \boldsymbol{x}_{\partial i}^t, x_i^t\right) \prod_{j \in \{\partial i \cup i\}} p\left(x_j^t\right) \tag{D.1}$$

where $\partial i\left(t\right)$ denotes the set of node i's neighbours at time t, and $x_{\partial i}^t = \left\{x_j^t\right\}_{j \in \partial i(t)}$. The transition probabilities $p\left(x_i^{t+1} \mid \boldsymbol{x}_{\partial i}^t, x_i^t\right)$ are computed starting from Eqs. (7.1) in the limit where the infection probabilities $\{\lambda_{k \to i}\}$ are small. This allows to approximate the products $\prod_{k \in \partial i}\left(1 - b_{k \to i}\right) \approx 1 - \sum_{k \in \partial i} b_{k \to i}$, which in turn gives the following expressions for the SIR dynamics:

$$p\left(x_i^{t+1} = S \mid x_{\partial i}^t, x_i^t\right) = \mathbb{I}\left[x_i^t = S\right]\left(1 - \sum_{k \neq i} \lambda_{k \to i} \mathbb{I}\left[x_k^t = I\right]\right) \tag{D.2a}$$

$$p\left(x_i^{t+1} = I \mid x_{\partial i}^t, x_i^t\right) = \left(1 - \mu_i\right) \mathbb{I}\left[x_i^t = I\right] + \mathbb{I}\left[x_i^t = S\right] \sum_{k \neq i} \lambda_{k \to i} \mathbb{I}\left[x_k^t = I\right] \tag{D.2b}$$

$$p\left(x_i^{t+1} = R \mid x_{\partial i}^t, x_i^t\right) = \mu_i \mathbb{I}\left[x_i^t = I\right] + \mathbb{I}\left[x_i^t = R\right] \tag{D.2c}$$

Then, combining Eqs. (D.2) with (D.1) one gets the update equations for the individual's marginal probabilities at each time:

$$p\left(x_i^{t+1} = S\right) = p\left(x_i^t = S\right)\left(1 - \sum_{k \neq i} \lambda_{k \to i}\left(t\right) p\left(x_j^t = I\right)\right) \tag{D.3a}$$

$$p\left(x_i^{t+1} = I\right) = \left(1 - \mu_i\right) p\left(x_i^t = I\right) + p\left(x_i^t = S\right) \sum_{k \neq i} \lambda_{k \to i}\left(t\right) p\left(x_j^t = I\right) \tag{D.3b}$$

$$p\left(x_i^{t+1} = R\right) = \mu_i p\left(x_i^t = I\right) + p\left(x_i^t = R\right) \tag{D.3c}$$

Equations (D.3) determine a Markovian evolution of the single node probabilities from a suitable initial condition $\left\{p\left(x_i^0=\alpha\right)\right\}_{i=1,\dots,N}^{\alpha\in\{S,I,R\}}$. The effect of observations can be included by using a simple heuristic: given a certain observation of node i at time t_{obs}, the Mean-field equations are propagated back in time in order to update the risk estimate for individual i to be infected at time $t > t_{obs}$. Under the assumption that observations are noiseless, at each time t the Mean-Field equations are run starting at a time $t - \tau_{MF}$ such that $t_{obs} \in [t - \tau_{MF}, t]$ by imposing the following conditions:

$$\text{if } x_i^{t_{obs}} = S \to P\left(x_i^{t'} = S\right) = 1 \quad \text{for } t' \in [t - \tau_{MF}, t_{obs}] \tag{D.4a}$$

$$\text{if } x_i^{t_{obs}} = I \to P\left(x_i^{t'} = I\right) = 1 \quad \text{for } t' \in [t_{obs} - \delta_{MF}, t_{obs}] \tag{D.4b}$$

$$\text{if } x_i^{t_{obs}} = R \to P\left(x_i^{t'} = R\right) = 1 \quad \text{for } t' \in [t_{obs}, T] \tag{D.4c}$$

where T is the simulation's time window. These equations have a simple interpretation: if an individual is observed S at time t_{obs}, it has been susceptible at all previous times; if it is observed as Recovered, it will continue to be as such for all future times; finally, if it is tested Infected at time t_{obs}, we assume that it has been infected in a time window between $t_{obs} - \delta_{MF}$ and t_{obs}, where δ_{MF} is a measure of the typical time between infection and the test, consecutive to symptoms appearance. Therefore, the Mean Field algorithm is parametrized by two quantities, τ_{MF} (the window time used to back-propagate the equations) and δ_{MF} just defined. In the MF approach marginal probabilities can be estimated in a fully distributed way on the individual's cell phones, by exchanging at the marginal probabilities its contacts at each day, without a central storage system.

MF parameters from ABM

As previously discussed, the MF inference can only deal with markovian dynamics, so that the time-dependency of the infective rate cannot be taken into account. We therefore implement an extremely simplified hypothesis of a constant infection probability $\lambda_{ij}^{MF}(t) = \lambda_0^{MF} \times \mathcal{G}_{ij}$, where, again, only the information about intra-household contacts is encoded by $\mathcal{G}_{ij}$ (defined in section 7.4). By averaging over a typical realization of the ABM we find $\lambda_0^{MF} = 0.02$. Notice that this parameter is reasonably small to justify the approximation presented at the beginning of this section, valid for small transmission probabilities.

The recovery time is parametrized by a memory-less exponential distribution, with a recovery rate $\mu^{MF} = 1/\left(12 \text{ days}\right)$. As a final remark, notice that the MF approach presented here is not - strictly speaking - a Bayesian inference method, since the effect of the observation is included only heuristically to propagate back in time the MF equations. In principle, a better MF approach could be designed by using the factorization assumption directly on the posterior distribution (7.3) rather than on the prior. This issue is left for future investigations.

Appendix E

BP equations for the SIR model

The message passing equations are derived following the approach discussed in Section 2.2.1 on the factor graph model (7.11). Before writing them, notice that the distribution of recovery times $R_i(r_i - t_i)$ needs to be discretized. All the terms in (7.10)-(7.11) except R_i are constant functions w.r.t. r_i in any interval $(\hat{r}_i, \hat{r}_i')$ of consecutive times in $\mathcal{T}_i$: in other words, besides the natural discretization imposed by the transmission times s_{ij}, the only thing that can happen in between is the recovery of a certain individual. For this reason, for each node we define a new discrete variable $\hat{r}_i$ so that $r_i = \hat{r}_i + u_i$, where $\hat{r}_i = \max\{r \in \mathcal{T}_i : r < r_i\}$. Then, by integrating out the u_is we obtain a fully discretized model on the variables $\boldsymbol{t}, \boldsymbol{s}, \hat{\boldsymbol{r}}$, whose functional form is identical to (7.11), with the following replacements:

$$
\begin{aligned}
r_i &\to \hat{r}_i \\
R_i(r_i - t_i) &\to \hat{R}_i(\hat{r}_i - t_i) = \int_{\hat{r}_i - t_i}^{\hat{r}_i' - t_i} R(u)\, du.
\end{aligned}
$$

In the following, the symbols $\hat{}$ are dropped to simplify the notation. The update equations for the factor-to-node message $m_{\psi_i \to (ij)}(s_{ij}, s_{ji})$ is given by:

$$
\begin{aligned}
m_{\psi_i \to (ij)}(s_{ij}, s_{ji}) \quad \propto \quad & \sum_{t_i} \sum_{r_i} p_{O,i}(\mathcal{O}_i \mid t_i, r_i)\, A_i(\boldsymbol{s}_{i^*})\, R_i(r_i - t_i)\, S_{ij}(s_{ij} \mid t_i, r_i) \times \\
& \times \sum_{\{s_{ki}\}} \sum_{\{s_{ik}\}} \delta\big(t_i, \min_{k \in \partial^* i} s_{ki}\big) \prod_{k \in \partial^* i \setminus j} S_{ik}(s_{ik} \mid t_i, r_i)\, m_{\psi_k \to (ik)}(s_{ki}, s_{ik})
\end{aligned}
\tag{E.1}
$$

Similarly, the marginals for t_i and r_i, denoted with $b_i(t_i)$ and $b_i(r_i)$ respectively read:

$$
b_i(t_i) \quad \propto \quad \sum_{r_i} p_{O,i}(\mathcal{O}_i \mid t_i, r_i)\, A_i(\boldsymbol{s}_{i^*})\, R_i(r_i - t_i)\, S_{ij}(s_{ij} \mid t_i, r_i) \times \tag{E.2}
$$

$$
\times \sum_{\{s_{ki}\}} \sum_{\{s_{ik}\}} \delta\big(t_i, \min_{k \in \partial^* i} s_{ki}\big) \prod_{k \in \partial^* i} S_{ik}(s_{ik} \mid t_i, r_i)\, m_{\psi_k \to (ik)}(s_{ki}, s_{ik}) \tag{E.3}
$$

$$
b_i(r_i) \quad \propto \quad \sum_{t_i} p_{O,i}(\mathcal{O}_i \mid t_i, r_i)\, A_i(\boldsymbol{s}_{i^*})\, R_i(r_i - t_i)\, S_{ij}(s_{ij} \mid t_i, r_i) \times \tag{E.4}
$$

$$
\times \sum_{\{s_{ki}\}} \sum_{\{s_{ik}\}} \delta\big(t_i, \min_{k \in \partial^* i} s_{ki}\big) \prod_{k \in \partial^* i} S_{ik}(s_{ik} \mid t_i, r_i)\, m_{\psi_k \to (ik)}(s_{ki}, s_{ik}) \tag{E.5}
$$

In the above equations the computational cost scales exponentially with the number of neighbours of node i. A more efficient computation can be achieved by decoupling the summation over s_{ki},

s_{ik}. Using the identity

$$\delta\left(t_i, \min_{k\in\partial^* i} s_{ki}\right) = \prod_{j\in\partial^* i} \mathbb{I}\left[s_{ki} \geq t_i\right] - \prod_{j\in\partial^* i} \mathbb{I}\left[s_{ki} > t_i\right] \tag{E.6}$$

and defining

$$G_k^0\left(t_i, r_i\right) = \sum_{\substack{s_{ki}\geq t_i \\ s_{ik} > t_i}} S_{ik}\left(s_{ik} \mid t_i, r_i\right) m_{\psi_k \to (ik)}\left(s_{ki}, s_{ik}\right) \tag{E.7}$$

$$G_k^1\left(t_i, r_i\right) = \sum_{\substack{s_{ki} > t_i \\ s_{ik} > t_i}} S_{ik}\left(s_{ik} \mid t_i, r_i\right) m_{\psi_k \to (ik)}\left(s_{ki}, s_{ik}\right) \tag{E.8}$$

The above BP equations (E.1)-(E.2)-(E.4) can be rewritten as:

$$\begin{aligned} m_{\psi_i \to (ij)}\left(s_{ij}, s_{ji}\right) \propto & \sum_{t_i}\sum_{r_i} p_{O,i}\left(\mathcal{O}_i \mid t_i, r_i\right) A_i\left(\boldsymbol{s}_{i^*}\right) R_i\left(r_i - t_i\right) S_{ij}\left(s_{ij} \mid t_i, r_i\right) \times & \text{(E.9)} \\ & \times \sum_{\{s_{ki}\}} \left(\prod_{l\in\partial^* i} \mathbb{I}\left[s_{li} \geq t_i\right] - \prod_{l\in\partial^* i} \mathbb{I}\left[s_{li} > t_i\right]\right) \prod_{k\in\partial^* i\setminus j} S_{ik}\left(s_{ik} \mid t_i, r_i\right) m_{ki}\left(s_{ki}, s_{ik}\right) \\ \propto & \sum_{t_i < s_{ji}} \sum_{r_i \geq t_i} p_{O,i}\left(\mathcal{O}_i \mid t_i, r_i\right) R_i\left(r_i - t_i\right) S_{ij}\left(s_{ij} \mid t_i, r_i\right) \times & \text{(E.10)} \\ & \times \prod_{t<t_i}\left(1-\gamma_i^t\right)\left\{\prod_{k\in\partial i\setminus j} G_k^0\left(t_i, r_i\right) - \left(1-\gamma_i^{t_i}\right)\prod_{k\in\partial i\setminus j} G_k^1\left(t_i, r_i\right)\right\} + \\ & + \sum_{r_i \geq s_{ji}} p_{O,i}\left(\mathcal{O}_i \mid s_{ji}, r_i\right) R_i\left(r_i - s_{ji}\right) \prod_{k\in\partial i\setminus j} G_k^0\left(s_{ji}, r_i\right) \end{aligned}$$

where in addition, we explicitly wrote the messages from the "virtual" neighbours $\tilde{i}^t$ (that accounts for the auto-infections) as

$$m_{i^{*t}i}\left(s_{i^{*t}i}, s_{ii^{*t}}\right) = \begin{cases} \gamma_i^t & s_{i^{*t}i} = t, s_{ii^{*t}} = \infty \\ 1-\gamma_i^t & s_{i^{*t}i} = \infty, s_{ii^{*t}} = \infty \end{cases} \tag{E.11}$$

Notice that the summation over r_i has non-zero contributions only for $r_i \geq t_i$ (by construction, recovery can occur only if node i is already infected) and the summation over t_i runs only for values $t_i \leq s_{ji}$, by virtue of Eq. (7.8). Analougsly, for the beliefs, we get:

$$b_i\left(t_i\right) \propto \sum_{r_i} p_{R,i}\left(r_i - t_i\right) p_{O,i}\left(\mathcal{O}_i \mid t_i, r_i\right) \times \tag{E.12}$$

$$\times \prod_{t<t_i}\left(1-\gamma_i^t\right)\left\{\prod_{k\in\partial i} G_k^0\left(t_i, r_i\right) - \left(1-\gamma_i^{t_i}\right)\prod_{k\in\partial i} G_k^1\left(t_i, r_i\right)\right\} \tag{E.13}$$

$$b_i\left(r_i\right) \propto \sum_{t_i} p_{R,i}\left(r_i - t_i\right) p_{O,i}\left(\mathcal{O}_i \mid t_i, r_i\right) \times \tag{E.14}$$

$$\times \prod_{t<t_i}\left(1-\gamma_i^t\right)\left\{\prod_{k\in\partial i} G_k^0\left(t_i, r_i\right) - \left(1-\gamma_i^{t_i}\right)\prod_{k\in\partial i} G_k^1\left(t_i, r_i\right)\right\} \tag{E.15}$$

The above implementation of the update equations for of all the messages in factor ψ_i has a computational cost $O\left(|\mathcal{T}_i| \sum_{j\in\partial i}\left(|\mathcal{T}_i| + |\mathcal{T}_{ij}|^2\right)\right)$.

In practice, the simulations are run for a fixed number of iterations, starting from the beliefs computed at the previous time step.
Once the beliefs are found, the marginal probabilities to be in one of the SIR states at any time t can be computed as follows:

$$p\left(x_i^t = S\right) = \sum_{\substack{t' \in \mathcal{T}_i \\ t' \geq t}} b_i^t\left(t'\right) \tag{E.16}$$

$$p\left(x_i^t = R\right) = \sum_{\substack{t' \in \mathcal{T}_i \\ r' \leq t}} b_i\left(r'\right) \tag{E.17}$$

$$p\left(x_i^t = I\right) = 1 - \sum_{\substack{t' \in \mathcal{T}_i \\ t' \geq t}} b_i^t\left(t'\right) - \sum_{\substack{t' \in \mathcal{T}_i \\ r' \leq t}} b_i\left(r'\right) \tag{E.18}$$

The simplest estimation for the individual's infection risk would be equal to its probability of being infected at the current time, namely $p\left(x_i^t = I\right)$, given by (E.18). However, given the infective features of SARS-CoV-2 included in the ABM we found more convenient to adopt a criterion that gives higher priority to most recent infections, i.e those that are more likely to be relevant for the infection dynamics. As a consequence, the infection risk within the BP scheme is computed as the sum of the beliefs in the window $\left[t - \delta^{BP}, t\right]$:

$$q_i^t = \sum_{t' \in \left[t - \delta^{BP}, t\right] \subset \mathcal{T}_i} b_i\left(t'\right) \tag{E.19}$$

In all the simulations, we used $\delta^{BP} = 10$ days.

Bibliography

1. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for boltzmann machines. eng. *Cognitive science* **9.** Place: Norwood, NJ Publisher: Elsevier Inc, 147–169 (1985) (cited on p. 95).

2. Altarelli, F., Braunstein, A., Dall'Asta, L. & Zecchina, R. Optimizing spread dynamics on graphs by message passing. *Journal of Statistical Mechanics: Theory and Experiment* **2013,** P09011 (2013) (cited on p. 123).

3. Altarelli, F., Braunstein, A., Dall'Asta, L., Wakeling, J. & Zecchina, R. Containing Epidemic Outbreaks by Message-Passing Techniques. en. *Physical Review X* **4,** 021024 (2014) (cited on p. 123).

4. Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A. & Zecchina, R. Bayesian inference of epidemics on networks via Belief Propagation. en. *Physical Review Letters* **112.** arXiv: 1307.6786, 118701 (2014) (cited on p. 123).

5. Altarelli, F., Braunstein, A., Dall'Asta, L., Ingrosso, A. & Zecchina, R. The patient-zero problem with noisy observations. en. *Journal of Statistical Mechanics: Theory and Experiment* **2014,** P10016 (2014) (cited on pp. 123, 138).

6. Aurell, E. & Ekeberg, M. Inverse Ising inference using all the data. en. *Physical Review Letters* **108.** arXiv: 1107.3536 (2012) (cited on pp. 92, 102).

7. Bailey, N. *The Mathematical Theory of Infectious Diseases and its Applications* (Griffin, London, 1975) (cited on p. 120).

8. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. en. *Science* **286.** Publisher: American Association for the Advancement of Science Section: Report, 509–512 (1999) (cited on pp. 17, 72, 111, 112).

9. Barber, M. N., Pearson, R. B., Toussaint, D. & Richardson, J. L. Finite-size scaling in the three-dimensional Ising model. *Physical Review B* **32.** Publisher: American Physical Society, 1720–1730 (1985) (cited on p. 87).

10. Barton, J. P., Cocco, S., De Leonardis, E. & Monasson, R. Large Pseudo-Counts and L_2-Norm Penalties Are Necessary for the Mean-Field Inference of Ising and Potts Models. en. *Physical Review E* **90,** 012132 (2014) (cited on p. 110).

11. Bay, J. *et al.* BlueTrace: A privacy-preserving protocol for community-driven contact tracing across borders. *Government Technology Agency-Singapore, Tech. Rep* **" "** (2020) (cited on p. 119).

12. Bentout, S., Chekroun, A. & Kuniya, T. Parameter estimation and prediction for coronavirus disease outbreak 2019 (COVID-19) in Algeria. **7,** 306–318 (2020) (cited on p. 121).

13. Besag, J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36,** 192–236 (1974) (cited on p. 102).

14. Bethe, H. A. & Bragg, W. L. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* **150.** Publisher: Royal Society, 552–575 (1935) (cited on p. 26).
15. Biggs, N. *Algebraic Graph Theory* 2nd ed. (Cambridge University Press, Cambridge, 1974) (cited on p. 83).
16. Blume, M., Emery, V. J. & Griffiths, R. B. Ising Model for the λ Transition and Phase Separation in He^3-He^4 Mixtures. *Physical Review A* **4.** Publisher: American Physical Society, 1071–1077 (1971) (cited on pp. 61, 115).
17. Borysov, S. S., Roudi, Y. & Balatsky, A. V. U.S. stock market interaction network as learned by the Boltzmann machine. en. *The European Physical Journal B* **88,** 321 (2015) (cited on p. 92).
18. Braunstein, A., Mézard, M. & Zecchina, R. Survey propagation: An algorithm for satisfiability. en. *Random Structures & Algorithms* **27,** 201–226 (2005) (cited on p. 80).
19. Braunstein, A. & Ingrosso, A. Inference of causality in epidemics on temporal contact networks. ENG. *Sci Rep* **6,** 27538 (2016) (cited on p. 123).
20. Braunstein, A. & Zecchina, R. Learning by Message Passing in Networks of Discrete Synapses. *Physical Review Letters* **96.** Publisher: American Physical Society, 030201 (2006) (cited on p. 30).
21. Braunstein, A., Kayhan, F. & Zecchina, R. Efficient data compression from statistical physics of codes over finite fields. en. *Physical Review E* **84.** arXiv: 1108.6239, 051111 (2011) (cited on p. 30).
22. Braunstein, A., Muntoni, A. P. & Pagnani, A. An analytic approximation of the feasible space of metabolic networks. en. *Nature Communications* **8.** Number: 1 Publisher: Nature Publishing Group, 14915 (2017) (cited on p. 37).
23. Braunstein, A., Muntoni, A. P., Pagnani, A. & Pieropan, M. Compressed sensing reconstruction using expectation propagation. en. *Journal of Physics A: Mathematical and Theoretical* **53.** Publisher: IOP Publishing, 184001 (2020) (cited on pp. 36, 38, 39).
24. Bury, T. Market structure explained by pairwise interactions. *Physica A: Statistical Mechanics and its Applications* **392,** 1375 –1385 (2013) (cited on p. 92).
25. Callen, H. B. A note on Green functions and the Ising model. en. *Physics Letters* **4,** 161 (1963) (cited on p. 102).
26. Cantwell, G. T. & Newman, M. E. J. Message passing on networks with loops. *Proceedings of the National Academy of Sciences* **116,** 23398–23403 (2019) (cited on p. 34).
27. Catania, G. *Density Consistency* `https://github.com/giovact/DensityConsistency`. 2021 (cited on p. 59).
28. Cencetti, G. *et al.* Digital proximity tracing on empirical contact networks for pandemic control. en. *Nature Communications* **12.** Number: 1 Publisher: Nature Publishing Group, 1655 (2021) (cited on p. 138).
29. Chan, J. *et al.* Pact: Privacy sensitive protocols and mechanisms for mobile contact tracing. *arXiv preprint arXiv:2004.03544* (2020) (cited on p. 119).
30. Chernyak, V. Y. & Chertkov, M. *Loop Calculus and Belief Propagation for q-ary Alphabet: Loop Tower* in *2007 IEEE International Symposium on Information Theory* ISSN: 2157-8117 (2007), 316–320 (cited on p. 32).
31. Chertkov, M. & Chernyak, V. Y. Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment* **2006.** Publisher: IOP Publishing, P06009 (2006) (cited on pp. 32, 114).

32. Cho, H., Ippolito, D. & Yu, Y. W. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511* (2020) (cited on p. 119).

33. Cocco, S. & Monasson, R. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. *Physical Review Letters* **106.** Publisher: American Physical Society, 090601 (2011) (cited on p. 99).

34. Cocco, S & Monasson, R. Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests. *Journal of Statistical Physics* **147,** 252–314 (2012) (cited on p. 99).

35. Cocco, S., Leibler, S. & Monasson, R. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences* **106,** 14058–14062 (2009) (cited on p. 92).

36. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics* **81.** Publisher: IOP Publishing, 032601 (2018) (cited on pp. 62, 63, 92, 115).

37. Coppersmith, D. & Winograd, S. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation* **9,** 251 –280 (1990) (cited on p. 58).

38. Cover, T. M. & Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, USA, 2006) (cited on pp. 12, 13, 24).

39. Davie, A. M. & Stothers, A. J. Improved bound for complexity of matrix multiplication. en. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* **143.** Publisher: Royal Society of Edinburgh Scotland Foundation, 351–369 (2013) (cited on p. 58).

40. Decelle, A. & Ricci-Tersenghi, F. Pseudolikelihood Decimation Algorithm Improving the Inference of the Interaction Network in a General Class of Ising Models. *Physical Review Letters* **112.** Publisher: American Physical Society, 070603 (2014) (cited on p. 103).

41. Derrida, B. Random-Energy Model: Limit of a Family of Disordered Models. *Physical Review Letters* **45.** Publisher: American Physical Society, 79–82 (1980) (cited on p. 14).

42. Domínguez, E., Lage-Castellanos, A., Mulet, R., Ricci-Tersenghi, F. & Rizzo, T. Characterizing and improving generalized belief propagation algorithms on the 2D Edwards–Anderson model. en. *Journal of Statistical Mechanics: Theory and Experiment* **2011.** Publisher: IOP Publishing, P12007 (2011) (cited on p. 77).

43. Domínguez, E., Lage-Castellanos, A., Mulet, R. & Ricci-Tersenghi, F. Gauge-free cluster variational method by maximal messages and moment matching. *Physical Review E* **95.** Publisher: American Physical Society, 043308 (2017) (cited on pp. 86–89).

44. Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory* **52,** 1289–1306 (2006) (cited on p. 36).

45. Durbin, R., Eddy, S., Krogh, A. S. & Mitchison, G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. English. Publisher: Cambridge University Press (1998) (cited on p. 110).

46. Erdös, P & Rényi, A. On Random Graphs I. *Publicationes Mathematicae Debrecen* **6,** 290–297 (1959) (cited on p. 16).

47. Erisman, A. M. & Tinney, W. F. On computing certain elements of the inverse of a sparse matrix. *Communications of the ACM* **18,** 177–179 (1975) (cited on p. 59).

48. Fawcett, T. An introduction to ROC analysis. en. *Pattern Recognition Letters. ROC Analysis in Pattern Recognition* **27,** 861–874 (2006) (cited on p. 111).

49. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368** (2020) (cited on pp. 125, 138).

50. Fintzi, J. *et al. Using multiple data streams to estimate and forecast SARS-CoV-2 transmission dynamics, with application to the virus spread in Orange County, California* arXiv:2009.02654 (2020) (cited on p. 121).

51. Fisher, M. E. & Gaunt, D. S. Ising Model and Self-Avoiding Walks on Hypercubical Lattices and "High-Density" Expansions. *Phys. Rev.* **133,** A224–A239 (1964) (cited on pp. 87, 89).

52. Franco, N. *Covid-19 Belgium: Extended SEIR-QD model with nursery homes and long-term scenarios-based forecasts from school opening* arXiV 2009.03450 (2020) (cited on p. 121).

53. Garey, M. R. & Johnson, D. S. *Computers and Intractability; A Guide to the Theory of NP-Completeness* (W. H. Freeman & Co., USA, 1990) (cited on p. 62).

54. Gaunt, D. S., Sykes, M. F. & McKenzie, S. Susceptibility and fourth-field derivative of the spin-1/2 Ising model for $T > T_c$ and $d = 4$. en. *Journal of Physics A: Mathematical and General* **12.** Publisher: IOP Publishing, 871–877 (1979) (cited on p. 87).

55. Geman, S. & Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6,** 721–741 (1984) (cited on pp. 70, 107, 110).

56. Georges, A & Yedidia, J. S. How to expand around mean-field theory using high-temperature expansions. en. *Journal of Physics A: Mathematical and General* **24,** 2173–2192 (1991) (cited on pp. 98, 114).

57. Giordano, R., Broderick, T. & Jordan, M. *Linear response methods for accurate covariance estimates from Mean field variational Bayes* in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (MIT Press, Cambridge, MA, USA, 2015), 1441–1449 (cited on p. 26).

58. *GitHub : ViraTrace* `github.com/ViraTrace/InfectionModel`. 2020 (cited on p. 121).

59. Gradshteyn, I. S., Ryzhik, I. M., Zwillinger, D. & Moll, V. *Table of integrals, series, and products; 8th ed.* (Academic Press, Amsterdam, 2014) (cited on p. 148).

60. Gray, R. M. *Toeplitz and Circulant Matrices: A Review* eng. Tech. rep. (2006) (cited on p. 145).

61. Guttmann, A. J. Lattice Green's functions in all dimensions. en. *Journal of Physics A: Mathematical and Theoretical* **43.** Publisher: IOP Publishing, 305205 (2010) (cited on pp. 83, 147).

62. Gómez, V., Mooij, J. M. & Kappen, H. J. Truncating the loop series expansion for belief propagation. *Journal of Machine Learning Research* **8,** 1987–2016 (2007) (cited on p. 32).

63. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* 2nd ed. en (Springer-Verlag, New York, 2009) (cited on p. 94).

64. Herbrich, R., Rastogi, R. & Vollgraf, R. CRISP: A Probabilistic Model for Individual-Level COVID-19 Infection Risk Estimation Based on Contact Data. *arXiv:2006.04942* (2020) (cited on p. 121).

65. *Immuni app* `https://www.immuni.italia.it/`. Accessed: 2021-02-19. 2020 (cited on p. 132).

66. Ising, E. Beitrag zur Theorie des Ferromagnetismus. de. *Zeitschrift für Physik* **31,** 253–258 (1925) (cited on pp. 14, 87).

67. Jacquelin, M., Lin, L. & Yang, C. PSelInv – A distributed memory parallel algorithm for selected inversion: The non-symmetric case. en. *Parallel Computing. Parallel Matrix Algorithms and Applications (PMAA'16)* **74,** 84–98 (2018) (cited on p. 59).

68. Jaynes, E. T. Information Theory and Statistical Mechanics. *Physical Review* **106.** Publisher: American Physical Society, 620–630 (1957) (cited on pp. 19, 91, 93).

69. Joyce, G. S. On the cubic modular transformation and the cubic lattice Green functions. en. *Journal of Physics A: Mathematical and General* **31.** Publisher: IOP Publishing, 5105–5115 (1998) (cited on pp. 84, 147).

70. Kabashima, Y. & Saad, D. Belief propagation vs. TAP for decoding corrupted messages. en. *EPL (Europhysics Letters)* **44.** Publisher: IOP Publishing, 668 (1998) (cited on p. 27).

71. Kappen, H. J. & Ortiz, F. d. B. R. in *Advances in Neural Information Processing Systems 10* (eds Jordan, M. I., Kearns, M. J. & Solla, S. A.) 280–286 (MIT Press, 1998) (cited on pp. 26, 96).

72. Kaveh, A. & Rahami, H. A unified method for eigendecomposition of graph products. en. *Communications in Numerical Methods in Engineering* **21,** 377–388 (2005) (cited on pp. 145, 146).

73. Kefayati, S. *et al.* On Machine Learning-Based Short-Term Adjustment of Epidemiological Projections of COVID-19 in US. *medRxiv.* Publisher: Cold Spring Harbor Laboratory Press (2020) (cited on p. 121).

74. Kermack, W. O., McKendrick, A. G. & Walker, G. T. Contributions to the mathematical theory of epidemics. II. —The problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **138.** Publisher: Royal Society, 55–83 (1932) (cited on p. 120).

75. Kikuchi, R. A theory of cooperative phenomena. *Physical review* **81.** Publisher: APS, 988 (1951) (cited on p. 32).

76. Kirkley, A., Cantwell, G. T. & Newman, M. E. J. Belief propagation for networks with loops. en. *Science Advances* **7.** Publisher: American Association for the Advancement of Science Section: Research Article, eabf1211 (2021) (cited on p. 34).

77. Lauer, S. A. *et al.* The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine* (2020) (cited on p. 127).

78. Lauritzen, S. L. *Graphical models* (Clarendon Press, 1996) (cited on p. 17).

79. Lebowitz, J. L. Coexistence of phases in Ising ferromagnets. en. *J Stat Phys* **16,** 463–476 (1977) (cited on p. 87).

80. Lenz, W. *Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern* en. Pages: 613-615 Volume: 21. 1920 (cited on p. 14).

81. Lezon, T. R., Banavar, J. R., Cieplak, M., Maritan, A. & Fedoroff, N. V. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences* **103,** 19033–19038 (2006) (cited on p. 92).

82. Liebmann, R. Statistical Mechanics of Periodic Frustrated Ising Systems. eng. *Lecture Notes in Physics* **251** (1986) (cited on p. 83).

83. Lin, L. *et al.* SelInv—An Algorithm for Selected Inversion of a Sparse Symmetric Matrix. *ACM Transactions on Mathematical Software* **37,** 40:1–40:19 (2011) (cited on p. 59).

84. Locasale, J. W. & Wolf-Yadlin, A. Maximum Entropy Reconstructions of Dynamic Signaling Networks from Quantitative Proteomics Data. *PLOS ONE* **4.** Publisher: Public Library of Science, 1–10 (2009) (cited on p. 92).

85. Lokhov, A. Y., Mézard, M., Ohta, H. & Zdeborová, L. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **90** (2014) (cited on p. 151).

86. Lokhov, A. Y., Mézard, M. & Zdeborová, L. Dynamic message-passing equations for models with unidirectional dynamics. *Physical Review E* **91.** Publisher: American Physical Society, 012811 (2015) (cited on p. 151).

87. MacKay, D. J. C. *Information Theory, Inference & Learning Algorithms* (Cambridge University Press, USA, 2002) (cited on p. 18).

88. Merchan, L. & Nemenman, I. On the Sufficiency of Pairwise Interactions in Maximum Entropy Models of Networks. en. *Journal of Statistical Physics* **162,** 1294–1308 (2016) (cited on p. 94).

89. Mezard, M., Parisi, G. & Virasoro, M. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications* (World Scientific Publishing Company, 1987) (cited on pp. 14, 15, 31).

90. Minka, T. P. *Expectation Propagation for approximate Bayesian inference* in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001), 362–369 (cited on pp. 35, 40, 88).

91. Montanari, A. & Rizzo, T. How to compute loop corrections to the Bethe approximation. en. *J. Stat. Mech.* **2005,** P10011 (2005) (cited on pp. 32, 86, 87, 89).

92. Mooij, J., Wemmenhove, B., Kappen, B. & Rizzo, T. *Loop Corrected Belief Propagation* in *Artificial Intelligence and Statistics* (2007), 331–338 (cited on pp. 32, 74).

93. Mooij, J. M. libDAI: A Free and Open Source C++ Library for Discrete Approximate Inference in Graphical Models. *Journal of Machine Learning Research* **11,** 2169–2173 (2010) (cited on pp. 74, 75).

94. Mooij, J. M. & Kappen, H. J. *Sufficient conditions for convergence of Loopy Belief Propagation* in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Arlington, Virginia, USA, 2005), 396–403 (cited on pp. 30, 32).

95. Moore, C. & Mertens, S. *The Nature of Computation* (Oxford University Press, Inc., USA, 2011) (cited on pp. 23, 79).

96. Mulet, R., Pagnani, A., Weigt, M. & Zecchina, R. Coloring Random Graphs. *Physical Review Letters* **89.** Publisher: American Physical Society, 268701 (2002) (cited on p. 62).

97. Muntoni, A. P., Rojas, R. D. H., Braunstein, A., Pagnani, A. & Pérez Castillo, I. Nonconvex image reconstruction via expectation propagation. *Physical Review E* **100.** Publisher: American Physical Society, 032134 (2019) (cited on p. 37).

98. Mézard, M., Parisi, G. & Zecchina, R. Analytic and Algorithmic Solution of Random Satisfiability Problems. en. *Science* **297.** Publisher: American Association for the Advancement of Science Section: Research Article, 812–815 (2002) (cited on p. 80).

99. Mézard, M. & Montanari, A. *Information, physics, and computation* (Oxford University Press, 2009) (cited on pp. 17, 23, 27, 28, 30, 79, 80).

100. Mézard, M. & Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. en. *Journal of Physiology-Paris. Neuromathematics of Vision* **103,** 107–113 (2009) (cited on p. 99).

101. Nguyen, H. C., Zecchina, R. & Berg, J. Inverse statistical problems: from the inverse Ising problem to data science. en. *Advances in Physics* **66.** arXiv: 1702.01522, 197–261 (2017) (cited on pp. 92, 98, 103, 107).

102. Onsager, L. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Physical Review* **65.** Publisher: American Physical Society, 117–149 (1944) (cited on pp. 14, 79, 87).

103. Opper, M. & Winther, O. Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling. *Physical Review E* **64.** Publisher: American Physical Society, 056131 (2001) (cited on pp. 35, 88).

104. Opper, M. & Winther, O. Expectation Consistent Approximate Inference. *The Journal of Machine Learning Research* **6.** Publisher: JMLR. org, 2177–2204 (2005) (cited on p. 35).

105. Pagnani, A. *Pseudo-likelihood Maximization* `https://github.com/pagnani/PlmIsing`. 2019 (cited on p. 110).

106. Parisi, G. Infinite Number of Order Parameters for Spin-Glasses. *Physical Review Letters* **43.** Publisher: American Physical Society, 1754–1756 (1979) (cited on p. 15).

107. Parisi, G. & Ruiz-Lorenzo, J. J. Scaling above the upper critical dimension in Ising models. *Physical Review B* **54.** Publisher: American Physical Society, R3698–R3701 (1996) (cited on p. 87).

108. Pearl, J. *Reverend Bayes on inference engines: a distributed hierarchical approach* in *Proceedings of the Second AAAI Conference on Artificial Intelligence* (1982), 133–136 (cited on p. 27).

109. Peierls, R. & Bragg, W. L. Statistical theory of superlattices with unequal concentrations of the components. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* **154.** Publisher: Royal Society, 207–222 (1936) (cited on p. 26).

110. Pelizzola, A. Cluster Variation Method in Statistical Physics and Probabilistic Graphical Models (2005) (cited on pp. 33, 34).

111. PETERSON, C. A mean field theory learning algorithm for neural networks. *Complex Systems* **1,** 995–1019 (1987) (cited on p. 96).

112. Plefka, T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. en. *Journal of Physics A: Mathematical and General* **15.** Publisher: IOP Publishing, 1971–1978 (1982) (cited on pp. 97, 98, 101).

113. Potts, R. B. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society* **48.** Publisher: Cambridge University Press, 106–109 (1952) (cited on p. 61).

114. R, H. *et al. COVID-19 Agent-based model with instantaneous contact tracing* tech. rep. (2020) (cited on pp. 121, 125–127, 129, 138).

115. Raskar, R. *et al.* Apps gone rogue: Maintaining personal privacy in an epidemic. *arXiv:2003.08567* (2020) (cited on p. 119).

116. Ricci-Tersenghi, F. The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. en. *Journal of Statistical Mechanics: Theory and Experiment* **2012.** Publisher: IOP Publishing, P08015 (2012) (cited on pp. 72, 88, 98–100).

117. Rodrigues, H. S. Application of SIR epidemiological model: new trends. *International Journal of Applied Mathematics and Informatics* **10,** 92–97 (2016) (cited on p. 120).

118. Roudi, Y., Aurell, E. & Hertz, J. Statistical physics of pairwise probability models. en. *Frontiers in Computational Neuroscience* **3.** arXiv: 0905.1410 (2009) (cited on pp. 98, 101).

119. Roudi, Y., Tyrcha, J. & Hertz, J. The Ising Model for Neural Data: Model Quality and Approximate Methods for Extracting Functional Connectivity. en. *Physical Review E* **79.** arXiv: 0902.2885, 051915 (2009) (cited on pp. 98, 99).

120. Sabidussi, G. Graph multiplication. en. *Mathematische Zeitschrift* **72,** 446–457 (1959) (cited on p. 145).

121. Salathé, M. *et al.* A high-resolution human contact network for infectious disease transmission. en. *Proceedings of the National Academy of Sciences* **107.** Publisher: National Academy of Sciences Section: Social Sciences, 22020–22025 (2010) (cited on p. 126).

122. Schelling, T. C. Dynamic models of segregation. *The Journal of Mathematical Sociology* **1,** 143–186 (1971) (cited on p. 15).

123. Sessak, V. Inverse problems in spin models. fr. arXiv: 1010.1960 (2010) (cited on p. 101).

124. Sessak, V. & Monasson, R. Small-correlation expansions for the inverse Ising problem. en. *Journal of Physics A: Mathematical and Theoretical* **42.** arXiv: 0811.3574, 055001 (2009) (cited on pp. 92, 101, 106, 114).

125. Shannon, C. E. A Mathematical Theory of Communication. en. *Bell System Technical Journal* **27,** 379–423 (1948) (cited on pp. 12, 18, 19).

126. Sherrington, D. & Kirkpatrick, S. Solvable Model of a Spin-Glass. *Physical Review Letters* **35.** Publisher: American Physical Society, 1792–1796 (1975) (cited on pp. 15, 97, 107).

127. Stein, R. R., Marks, D. S. & Sander, C. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. en. *PLOS Computational Biology* **11.** Publisher: Public Library of Science, e1004182 (2015) (cited on p. 94).

128. Strecka, J. & Jascur, M. A brief account of the Ising and Ising-like models: Mean-field, effective-field and exact results. arXiv: 1511.03031 (2015) (cited on pp. 17, 84).

129. Tanaka, T. Mean-field theory of Boltzmann machine learning. *Physical Review E* **58.** Publisher: American Physical Society, 2302–2310 (1998) (cited on p. 97).

130. Thouless, D. J., Anderson, P. W. & Palmer, R. G. Solution of 'Solvable model of a spin glass'. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics* **35,** 593–601 (1977) (cited on p. 97).

131. Torquato, S. Toward an Ising model of cancer and beyond. eng. *Physical Biology* **8,** 015017 (2011) (cited on p. 15).

132. Troncoso, C. *et al.* Decentralized privacy-preserving proximity tracing. arXiv:2005.12273 (2020) (cited on pp. 119, 138).

133. Tyrcha, J., Roudi, Y., Marsili, M. & Hertz, J. The effect of nonstationarity on models inferred from neural data. en. *Journal of Statistical Mechanics: Theory and Experiment* **2013,** P03005 (2013) (cited on p. 92).

134. Wainwright, M. J. & Jordan, M. I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* **1,** 1–305 (2008) (cited on p. 17).

135. Wannier, G. H. Antiferromagnetism. The Triangular Ising Net. *Physical Review B* **7.** Publisher: American Physical Society, 5017–5017 (1973) (cited on p. 84).

136. Wartell, R. M. & Benight, A. S. Thermal denaturation of DNA molecules: A comparison of theory with experiment. en. *Physics Reports* **126,** 67–107 (1985) (cited on p. 15).

137. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. en. *Nature* **393,** 440–442 (1998) (cited on pp. 17, 126).

138. Welling, M. & Teh, Y. W. Approximate inference in Boltzmann machines. en. *Artificial Intelligence* **143,** 19–50 (2003) (cited on pp. 32, 72, 99).

139. Welling, M. & Teh, Y. W. Linear Response Algorithms for Approximate Inference in Graphical Models. *Neural Computation* **16.** Conference Name: Neural Computation, 197–221 (2004) (cited on pp. 32, 99).

140. Wheeler, J. C. Decorated Lattice-Gas Models of Critical Phenomena in Fluids and Fluid Mixtures. *Annual Review of Physical Chemistry* **28,** 411–443 (1977) (cited on p. 15).

141. Wolff, U. Collective Monte Carlo Updating for Spin Systems. *Physical Review Letters* **62.** Publisher: American Physical Society, 361–364 (1989) (cited on p. 78).

142. Wu, F. Y. The Potts model. *Reviews of Modern Physics* **54,** 235–268 (1982) (cited on pp. 61, 91).

143. Yedidia, J. S., Freeman, W. T. & Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* **51.** Conference Name: IEEE Transactions on Information Theory, 2282–2312 (2005) (cited on pp. 26–28, 30, 32, 34).

144. Yedidia, J. S. An Idiosyncratic Journey Beyond Mean Field Theory. en (2000) (cited on p. 27).

145. Yedidia, J. S., Freeman, W. T. & Weiss, Y. *Bethe free energy, Kikuchi approximations and belief propagation algorithms* (2000) (cited on pp. 26, 27).

146. Yedidia, J. S., Freeman, W. T. & Weiss, Y. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium* **8,** 236–239 (2003) (cited on pp. 26, 27, 32).

147. Yuille, A. L. CCCP algorithms to minimize the Bethe and Kikuchi free energies: convergent alternatives to belief propagation. eng. *Neural Computation* **14,** 1691–1722 (2002) (cited on pp. 30, 32).

148. Yuille, A. *A Double-Loop Algorithm to Minimize the Bethe Free Energy* en. in *Energy Minimization Methods in Computer Vision and Pattern Recognition* (eds Figueiredo, M., Zerubia, J. & Jain, A. K.) (Springer, Berlin, Heidelberg, 2001), 3–18 (cited on p. 32).

149. Zdeborová, L. & Krzakala, F. Statistical physics of inference: thresholds and algorithms. *Advances in Physics* **65,** 453–552 (2016) (cited on p. 19).

150. Zhao, H., Zhou, J., Zhang, A., Su, G. & Zhang, Y. Self-organizing Ising model of artificial financial markets with small-world network topology. en. *EPL (Europhysics Letters)* **101.** Publisher: IOP Publishing, 18001 (2013) (cited on p. 15).