Inverse statistical problems: from the inverse Ising problem to data science

H. Chau Nguyen, Riccardo Zecchina & Johannes Berg

Taylor & Francis
Taylor & Francis Group

Check for updates

REVIEW ARTICLE

Inverse statistical problems: from the inverse Ising problem to data science

H. Chau Nguyen[a], Riccardo Zecchina[b,c] and Johannes Berg[d*]

[a]*Max-Planck-Institut für Physik komplexer Systeme, Dresden, Germany;* [b]*Department of Decision Sciences, Bocconi University, Milano, Italy;* [c]*Institute of Condensed Matter Physics and Complex Systems, Politecnico di Torino, Torino, Italy;* [d]*Institute for Theoretical Physics, University of Cologne, Cologne, Germany*

Inverse problems in statistical physics are motivated by the challenges of 'big data' in different fields, in particular high-throughput experiments in biology. In inverse problems, the usual procedure of statistical physics needs to be reversed: Instead of calculating observables on the basis of model parameters, we seek to infer parameters of a model based on observations. In this review, we focus on the inverse Ising problem and closely related problems, namely how to infer the coupling strengths between spins given observed spin correlations, magnetizations, or other data. We review applications of the inverse Ising problem, including the reconstruction of neural connections, protein structure determination, and the inference of gene regulatory networks. For the inverse Ising problem in equilibrium, a number of controlled and uncontrolled approximate solutions have been developed in the statistical mechanics community. A particularly strong method, pseudolikelihood, stems from statistics. We also review the inverse Ising problem in the non-equilibrium case, where the model parameters must be reconstructed based on non-equilibrium statistics.

Contents		PAGE

*Corresponding author. Email: berg@thp.uni-koeln.de

1. Introduction and applications

The primary goal of statistical physics is to derive observable quantities from microscopic laws governing the constituents of a system. In the example of the Ising model, the starting point is a model describing interactions between elementary magnets (spins), the goal is to derive observables like spin magnetizations and correlations.

In an inverse problem, the starting point is observations of some system whose microscopic parameters are unknown and to be discovered. In the inverse Ising problem, the interactions between spins are not known to us, but we want to learn them from measurements of magnetizations, correlations, or other observables. In general, the goal is to infer the parameters describing a system (for instance, its Hamiltonian) from extant data. To this end, the relationship between microscopic laws and observables needs to be inverted.

In the last two decades, inverse statistical problems have arisen in different contexts, sparking interest in the statistical physics community in taking the path from model parameters to observables in reverse. The areas where inverse statistical problems have arisen are characterized by (i) microscopic scales becoming experimentally accessible and (ii) sufficient data storage capabilities being available. In particular, the biological sciences have generated several inverse statistical problems, including the reconstruction of neural and gene regulatory networks and the determination of the three-dimensional structure of proteins. Technological progress is likely to

to open up further fields of research to inverse statistical analysis, a development that is currently described by the label 'big data'.

In physics, inverse statistical problems also arise when we need to design a many-body system with particular desired properties. Examples are finding the potentials that result in a particular single-particle distribution [48,122], interaction parameters in a binary alloy that yield the observed correlations [142], the potentials between atoms that lead to specific crystal lattices [252], or the parameters of a Hamiltonian that lead to a particular density matrix [47]. In the context of soft matter, a question is how to design a many-body system that will self-assemble into a particular spatial configuration or has particular bulk properties [184,226]. In biophysics, we may want to design a protein that folds into a specified three-dimensional shape [120]. For RNA, even molecules with more than one stable target structure are possible [78]. As a model of such design problems, DiStasio et al. [66] and Marcotte et al. [136] study how to find the parameters of Ising Hamiltonian with a prescribed ground state.

In all these examples, 'spin' variables describe microscopic degrees of freedom particular to a given system, for instance, the states of neurons in a neural network. The simplest description of these degrees of freedom in terms of random binary variables then leads to Ising-type spins. In the simplest non-trivial scenario, correlations between the 'spins' are generated by pairwise couplings between the spins, leading to an Ising model with unknown parameters (couplings between the spins and magnetic fields acting on the spins). In many cases of interest, the couplings between spins will not all be positive, as is the case in a model of a ferromagnet. Nor will they couplings conform to a regular lattice embedded in some finite-dimensional space.

For a concrete example, we look at a system of N binary variables (Ising spins) $\{s_i\}$, $i = 1, \ldots, N$, with $s_i = \pm 1$. These spins are coupled by pairwise couplings J_{ij} and are subject to external magnetic fields h_i.

$$P(\{s_i\}) = \frac{1}{Z} \exp \left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j \right] \tag{1}$$

is the Boltzmann equilibrium distribution $P(\{s_i\}) = \mathrm{e}^{-H(\{s_i\})}/Z$, where we have subsumed temperature into the couplings and fields. (We will discuss this choice in Section 2.1.) The Hamiltonian

$$H(\{s_i\}) = - \sum_i h_i s_i - \sum_{i<j} J_{ij} s_i s_j \tag{2}$$

specifies the energy of the spin system as a function of the microscopic spin variables, local fields, and pairwise couplings. The inverse Ising problem is the determination of the couplings J_{ij} and local fields h_i, given a set of M observed spin configurations. Depending on the particular nature of the system at hand, the restriction to binary variables or pairwise interactions may need to be lifted, or the functional form of the Hamiltonian may be altogether different from the Ising Hamiltonian with pairwise interactions (2). For non-equilibrium systems, the steady state is not even described by a Boltzmann distribution with a known Hamiltonian. However, the basic idea remains the same across different types of inverse statistical problems: even when the frequencies of spin configurations may be under-sampled, the data may be sufficient to infer at least some parameters of a model.

The distribution (1) is well known not only as the equilibrium distribution of the Ising model. It is also the form of the distribution which maximizes the (Gibbs) entropy

$$S[P] = - \sum_{\{s_i\}} P(\{s_i\}) \ln P(\{s_i\}) \tag{3}$$

under the constraint that $P(\{s_i\})$ is normalized and has particular first and second moments, that is, magnetizations and correlations. We will discuss in Section 2.2.3 how this distribution emerges as the 'least biased distribution' of a binary random variable with prescribed first and second moments [107]. The practical problem is then again an inverse one: to find the couplings J_{ij} and local fields h_i such that the first and second moments observed under the Boltzmann distribution (1) match the mean values of s_i and $s_i s_j$ in data. In settings where third moments differ significantly from the prediction of (1) based on the first two moments, one may need to construct the distribution of maximum entropy given the first three moments, leading to three-spin interactions in the exponent of Equation (1).

Determining the parameters of a distribution such as Equation (1) is always a many-body problem: changing a single coupling J_{ij} generally affects correlations between many spin variables, and conversely a change in the correlation between two variables can change the values of many inferred couplings. The interplay between model parameters and observables is captured by a statistical mechanics of inverse problems, where the phase space consists of quantities normally considered as fixed model parameters (couplings, fields). The observables, such as spin correlations and magnetizations on the other hand, are taken to be fixed, as they are specified by empirical observations. Such a perspective is not new to statistical physics; the analysis of neural networks in the seventies and eighties of the last century led to a statistical mechanics of learning [72,95,237], where the phase space is defined by the set of possible rules linking the input into a machine with an output. The set of all rules compatible with a given set of input/output relations then defines a statistical ensemble. In the inverse statistical problems, however, there are generally no explicit rules linking the input and output, but data with different types of correlations or other observations, which are to be accounted for in a statistical model.

Inverse statistical problems fall into the broader realm of statistical inference [31,131], which seeks to determine the properties of a probability distribution underlying some observed data. The problem of inferring the parameters of a distribution such as (1) is known under different names in different communities; also emphasis and language differ subtly across communities.

- In statistics, an inverse problem is the inference of model parameters from data. In our case, the problem is the inference of the parameters of the Ising model from observed spin configurations. A particular subproblem is the inference of the graph formed by the non-zero couplings of the Ising model, termed graphical model selection or reconstruction. In the specific context of statistical models on graphs (graphical models), the term *inference* describes the calculation of the marginal distribution of one or several variables. (A marginal distribution describes the statistics of one or several particular variables in a many-variable distribution, for example, $P(x_1) = \sum_{x_2, x_3} P(x_1, x_2, x_3)$.)
- In machine learning, a frequent task is to train an artificial neural network with symmetric couplings such that magnetizations and correlations of the artificial neurons match the corresponding values in the data. This is a special case of what is called Boltzmann machine learning; the general case also considers so-called hidden units, whose values are unobserved [1].
- In statistical physics, much effort has been directed towards estimating the parameters of the Ising model given observed values of the magnetization and two-point correlations. As we will see in Section 2.2, this is a hard problem from an algorithmic point of view. Recently, threshold phenomena arising in inference problems have attracted much interest from the statistical physics community, due to the link between phase transitions and the boundaries separating different regimes of inference problems, for instance solvable and unsolvable problems, or easy and hard ones [144,247].

Common theme across different applications and communities is the inference of model parameters given observed data or desired properties. In this review, we will focus on a prototype inverse statistical problem: the inverse Ising problem and its close relatives. Many of the approaches developed for this problem are also readily extended to more general scenarios. We will start with a discussion of applications of the inverse Ising problem and related approaches in biology, specifically the reconstruction of neural and genetic networks, the determination of three-dimensional protein structures, the inference of fitness landscapes, the bacterial responses to combinations of antibiotics, and flocking dynamics. We will find that these applications define two distinct settings of the inverse Ising problem; equilibrium and non-equilibrium. Part 2 of this review treats the inverse Ising problem in an equilibrium setting, where the couplings between spins are symmetric. Detailed balance holds and results from equilibrium statistical physics can be used. This setting arises naturally within the context of maximum entropy models, which seek to describe the observed statistics of configurations with a simplified effective model capturing, for instance, collective effects. We introduce the basics of statistical inference and maximum entropy modelling, discuss the thermodynamics of the inverse Ising problem, and review different approaches to solve the inverse Ising problem, pointing out their connections and comparing the resulting parameter reconstructions. Part 3 of this review considers asymmetric coupling matrices, where in the absence of detailed balance couplings can be reconstructed from time series, from data on perturbations of the system, or from detailed knowledge of the non-equilibrium steady state (NESS).

We now turn to applications of the inverse Ising problem, which mostly lie in the analysis of high-throughput data from biology. One aim of inverse statistical modelling is to find the parameters of a microscopic model to describe these data. A more ambitious aim is achieved when the model is informative about the processes which produced the data, this is, when some of the mechanisms underlying the data can be inferred. The data are large-scale measurements of the degrees of freedom of some system. In the language of statistical physics, these describe the micro-states of a system: states of neurons, particular sequences of DNA or proteins, or the concentration levels of RNA. We briefly introduce some of the experimental background of these measurements, so their potential and the limitations can be appreciated. The models are simple models of the microscopic degrees of freedom. In the spirit of statistical physics, these models are simple enough so the parameters can be computed given the data, yet sufficiently complex to reproduce some of the statistical interdependences of the observed microscopic degrees of freedom. The simplest case, consisting of binary degrees of freedom with unknown pairwise couplings between them, leads to the inverse Ising problem, although we will also discuss several extensions.

1.1. *Modelling neural firing patterns and the reconstruction of neural connections*

Neurons can exchange information by generating discrete electrical pulses, termed spikes, that travel down nerve fibres. Neurons can emit these spikes at different rates, a neuron emitting spikes at a high rate is said to be 'active' or 'firing', a neuron emitting spikes at a low rate or not at all is said to be 'inactive' or 'silent'. The measurement of the activity of single neurons has a long history starting in 1953 with the development of micro-electrodes for recording [68]. Multi-electrodes were developed, allowing one to record multiple simultaneous neural signals independently over long time periods [162,208]. Such data present the intriguing possibility to see elementary brain function emerge from the interplay of a large number of neurons.

However, even when vast quantities of data are available, the different configurations of a system are still under-sampled in most cases. For instance, consider N neurons, each of which can be either active (firing) or inactive (silent). Given that the firing patterns of thousands of

neurons can be recorded simultaneously [197], the number of observations M will generally be far less than the total number of possible neural configurations, $M \ll 2^N$. For this reason, a straightforward statistical description that seeks to determine directly the frequency with which each configuration appears will likely fail.

On the other hand, a feasible starting point is a simple distribution, whose parameters can be determined from the data. For a set of N binary variables, this might be a distribution with pairwise interactions between the variables. In [195], Bialek and collaborators applied such a statistical model to neural recordings. Dividing time into small intervals of duration $\Delta \tau = 20$ ms induces a binary representation of neural data, where each neuron i either spikes during a given interval ($s_i = 1$) or it does not ($s_i = 0$). The joint statistics observed in 40 neurons in the retina of a salamander was modelled by an Ising model (1) with magnetic fields and pairwise symmetric couplings. Rather than describing the dynamics of neural spikes, this model describes the correlated firing of different neurons over the course of the experiment. The symmetric couplings J_{ij} in Equation (1) describe statistical dependencies, not physical connections. The synaptic connections between neurons, on the other hand, are generally not symmetric.

In this context, the distribution (1) can be viewed as the form of the maximum entropy distribution over neural states, given the observed one- and two-point correlations [195]. In [52,203], a good match was found between the statistics of three neurons predicted by Equation (1) and the firing patterns of the same neurons in the data. This means that the model with pairwise couplings provides a valid statistical description of the empirical data, one that can even be used to make predictions. Similar results were obtained also from cortical cells in cell cultures [216].

The mapping from the neural data to a spin model rests on dividing time into discrete bins of duration $\Delta \tau$. A different choice of this interval would lead to different spin configurations; in particular, changing $\Delta \tau$ affects the magnetization of all spins by altering the number of intervals in which a neuron fires. In [190], Roudi, Nirenberg, and Latham show that the pairwise model (1) provides a good description of the underlying spin statistics (generated by neural spike trains), provided $N \Delta \tau \nu \ll 1$, where ν is the average firing rate of neurons. Increasing the bin size beyond this regime leads to an increase in bins where multiple neurons fire, as a result couplings beyond the pairwise couplings in Equation (1) can become important.

As a minimal model of neural correlations, the statistics (1) has been extended in several ways. Tkačik et al. [224] and Granot-Atedgi et al. [88] consider stimulus-dependent magnetic fields, that is, fields which depend on the stimulus presented to the experimental subject at a particular time of the experiment. Ohiorhenuan et al. [163] looks at stimulus-dependent couplings. When the number of neurons increases to ~ 100, limitations of the pairwise model (1) become apparent, which has be addressed by adding additional terms coupling triplets, etc. of spins in the exponent of the Boltzmann measure (1) [82].

The statistics (1) serves as a description of the empirical data: the couplings between spins in the Hamiltonian (2) do not describe physical connections between the neurons. The determination of the network of neural connections from the observed neural activities is thus a different question. Simoncelli and collaborators [168,178] and Cocco et al. [52] use an integrate-and-fire model [40] to infer how the neurons are interconnected on the basis of time series of spikes in all neurons. In such a model, the membrane potential of neuron i obeys the dynamics

$$C \frac{dV_i}{dt} = \sum_{j \neq i} J_{ij} \sum_l K(t - t_{jl}) + I_i - g V_i + \xi_i(t), \qquad (4)$$

where the first term on the right-hand side encodes the synaptic connections J_{ij} and a memory kernel K; t_{jl} specifies the time at which neuron j emitted its lth spike. The remaining terms describe a background current, voltage leakage, and white noise. Finding the synaptic

connections J_{ij} that best describe a large set of neural spike trains is a computational challenge; Paninski [167] and Cocco et al. [52] develop an approximation based on maximum likelihood (ML), see Section 2.2. A related approach based on point processes and generalized linear models is presented in [228]. We will discuss this problem of inferring the network of connections the context of the non-equilibrium models in Section 3.

Neural recordings give the firing patterns of several neurons over time. These neurons may have connections between them, but they also receive signals from neural cells whose activity is not recorded [133]. In [229], the effect of connections between neurons is disentangled from correlations arising from shared non-stationary input. This raises the possibility that the correlations described by the pairwise model (1) in [195] and related works originate from a confounding factor (connections to a neuron other than those whose signal is measured), rather than from connections between recorded neurons [121].

1.2. *Reconstruction of gene regulatory networks*

The central dogma of molecular biology is this: Proteins are macromolecules consisting of long chains of amino acids. The particular sequence of a protein is encoded in DNA, a double-stranded helix of complementary nucleotides. Specific parts of DNA, the genes, are transcribed by polymerases, producing a single-stranded copy called m(essenger)RNAs, which are translated by ribosomes, usually multiple times, to produce proteins.

The process of producing protein molecules from the DNA template by transcription and translation is called gene expression. The expression of a gene is tightly controlled to ensure that the right amounts of proteins are produced at the right time. One important control mechanism is transcription factors, proteins which affect the expression of a gene (or several) by binding to DNA near the transcription start site of that gene. This part of DNA is called the regulatory region of a gene. A target gene of a transcription factor may in turn encode another transcription factor, leading to a cascade of regulatory events. To add further complications, the binding of multiple transcription factors in the regulatory region of a gene leads to combinatorial control exerted by several transcription factors on the expression of a gene [37]. Can the regulatory connections between genes be inferred from data on gene expression, that is, can we learn the identity of transcription factors and their targets?

Over the last decades, the simultaneous measurement of expression levels of all genes have become routine. At the centre of this development are two distinct technological advances to measure mRNA levels. The first is microarrays, consisting of thousands of short DNA sequences, called probes, grafted to the surface of a small chip. After converting the mRNA in a sample to DNA by reverse transcription, cleaving that DNA into short segments, and fluorescently labelling the resulting DNA segments, fluorescent DNA can bind to its complementary sequence on the chip. (Reverse transcription converts mRNA to DNA, a process which requires a so-called reverse transcriptase as an enzyme.) The amount of fluorescent DNA bound to a particular probe depends on the amount of mRNA originally present in the sample. The relative amount of mRNA from a particular gene can then be inferred from the fluorescence signal at the corresponding probes [98]. A limitation of microarrays is the large amount of mRNA required: The mRNA sample is taken from a population of cells. As a result, cell-to-cell fluctuations of mRNA concentrations are averaged over. To obtain time series, populations of cells synchronized to approximately the same stage in the cell cycle are used [84].

The second way to measure gene expression levels is also based on reverse transcription of mRNA, followed by high-throughput sequencing of the resulting DNA segments. Then, the relative mRNA levels follow directly from counts of sequence reads [236]. Recently, expression levels even in single cells have been measured in this way [241]. In combination with barcoding

(adding short DNA markers to identify individual cells), 10^4 cells can have their expression profiled individually in a single sequencing run [134]. Such data may allow, for instance, the analysis of the response of target genes to fluctuations in the concentration of transcription factors. However, due to the destructive nature of single-cell sequencing, there may never be single-cell data that give time series of genome-wide expression levels.

Unfortunately, cellular concentrations of proteins are much harder to measure than mRNA levels. As a result, much of the literature focuses on mRNA levels, neglecting the regulation of translation. Advances in protein mass-spectrometry [177] may lead to data on both mRNA and protein concentrations. These data would pose the additional challenge of inferring two separate levels of gene regulation: gene transcription from DNA to mRNA and translation from mRNA to proteins.

As in the case of neural data discussed in the preceding section, gene expression data presents two distinct challenges: (i) finding a statistical description of the data in terms of suitable observables and (ii) inferring the underlying regulatory connections. Both these problems have been addressed extensively in the machine learning and quantitative biology communities. Clustering of gene expression data to detect sets of genes with correlated expression levels has been used to detect regulatory relationships. A model-based approach to the reconstruction of regulatory connections is Boolean networks. Boolean networks assign binary states to each gene (gene expression on/off), and the state of a gene at a given time depends on the state of all genes at a previous time through a set of logical functions assigned to each gene. See [64] for a review of clustering and Boolean network inference and [96] for a review of Boolean network inference.

A statistical description that has also yielded insight into regulatory connections is Bayesian networks. A Bayesian network is a probabilistic model describing a set of random variables (expression levels) through conditional dependencies described by a directed acyclic graph. Learning both the structure of the graph and the statistical dependencies is a hard computational problem, but can capture strong signals in the data that are often associated with a regulatory connection. In principle, causal relationships (like the regulatory connections) can be inferred, in particular if the regulatory network contains no cycles. For reviews, see [79,116]. Both Boolean or Bayesian networks have been applied to measurements of the response of expression levels to external perturbations of the regulatory network or of expression levels, see [103,173]. A full review of these methods is beyond the scope of this article, instead we focus on approaches related to the inverse Ising problem.

For a statistical description of gene expression levels, Lezon et al. [126] applied a model with pairwise couplings

$$P(\{x_i\}) = \exp\left[\sum_{i \leq j} J_{ij}x_i x_j + \sum_i h_i x_i\right]/Z, \tag{5}$$

fitted to gene expression levels. The standard definition of expression levels $\{x_i\}$ is $\log_2$-values of fluorescence signals with the mean value for each gene subtracted. Since Equation (5) is a multi-variate Gaussian distribution, the matrix of couplings J_{ij} must be negative definite. These couplings can be inferred simply by inverting the matrix of variances and covariances of expression levels. In [126], the resulting couplings J_{ij} were then used to identify hub genes which regulate many targets. The same approach was used in [128] to analyse the cellular signalling networks mediated by the phosphorylation of specific sites on different proteins. Again, the distribution (5) can be viewed as a maximum entropy distribution for *continuous* variables with prescribed first and second moments. This approach is also linked to the concept of partial correlations in statistics [9,119].

Again, the maximum entropy distribution (5) has symmetric couplings between expression levels, whereas the network of regulatory interactions is intrinsically asymmetric. One way to infer the regulatory connections is time series [204]. Bailly-Bechet et al. [12] use expression levels measured at different times to infer the regulatory connections, based on a minimal model of expression dynamics with asymmetric regulatory connections between pairs of genes. In this model, expression levels x_i^t at successive time intervals t obey

$$\text{sign}(x_i^{t+1}) = \begin{cases} 1 & \text{if } \sum_j J_{ij}x_j^t > \kappa, \\ -1 & \text{if } \sum_j J_{ij}x_j^t < \kappa, \end{cases} \tag{6}$$

where κ is a threshold. The regulatory connections J_{ij} are taken to be discrete, with the values $-1, 1, 0$ denoting repression, activation, and no regulation of gene i by the product of gene j. The matrix of connections is then inferred based on Bayes' theorem (see Section 2.2) and an iterative algorithm for estimating marginal probabilities (message passing, see Section 2.2.11).

A second line of approach that can provide information on regulatory connections is perturbations [217]. An example is gene knockdowns, where the expression of one or more genes is reduced, by introducing small interfering RNA (siRNA) molecules into the cell [67] or by other techniques. siRNA molecules can be introduced into cells from the outside; after various processing steps they lead to the cleavage of mRNA with a complementary sequence, which is then no longer available for translation. If that mRNA translates to a transcription factor, all targets of that transcription factor will be upregulated or downregulated (depending on whether the transcription factor acted as a repressor or an activator, respectively). Knowing the responses of gene expression levels to a sufficient number of such perturbations allows the inference of regulatory connections. Molinelli et al. [148] consider a model of gene expression dynamics based on continuous variables x_i evolving deterministically as $\partial_t x_i = a_i \tanh(\sum_j J_{ij}x_j) - c_i x_i$. The first term describes how the expression level of gene j affects the rate of gene expression of gene i via the regulatory connection J_{ij}, the second term describes mRNA degradation. The stationary points of this model shift in response to perturbations of expression levels of particular genes (for instance through knockdowns), and these changes depend on regulatory connections. In [148], the regulatory connections are inferred from perturbation data, again using belief propagation.

1.3. *Protein structure determination*

Tremendous efforts have been made to determine the three-dimensional structure of proteins. A linear amino acid chain folds into a convoluted shape, the folded protein, thus bringing amino acids into close physical proximity that are separated by a long distance along the linear sequence.

Due to the number of proteins (several thousand per organism) and the length of individual proteins (hundreds of amino acid residues), protein structure determination is a vast undertaking. However, the rewards are also substantial. The three-dimensional structure of a protein determines its physical and chemical properties, and how it interacts with other cellular components: broadly, the shape of a protein determines many aspects of its function. Protein structure determination relies on crystallizing proteins and analysing the X-ray diffraction pattern of the resulting solid. Given the experimental effort required, the determination of a protein's structure from its sequence alone has been a key challenge to computational biology for several decades [60,65]. The computational approach models the forces between amino acids in order to find the low-energy structure a protein in solution will fold into. Depending on the level of detail, this approach requires extensive computational resources (Figure 1).
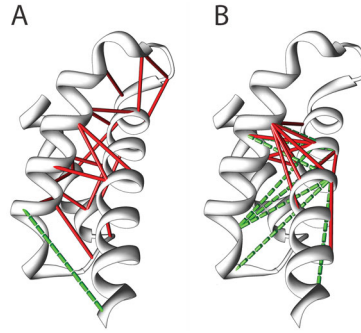
Figure 1. Correlations and couplings in protein structure determination. Both figures show the three-dimensional structure of a particular part (region 2) of the protein SigmaE of *E. coli*, as determined by X-ray diffraction. This protein, or rather a protein of similar sequence and presumably similar structure, occurs in many other bacterial species as well. In (B), lines indicate pairs of sequence positions whose amino acids are highly correlated across different bacteria: for each pair of sequence positions at least 5 amino acids apart, the mutual information of pairwise frequency counts of amino acids was calculated, and the 20 most correlated pairs are shown here. Such pairs that also turn out to be close in the three-dimensional structure are shown with full lines, those whose distance exceeds 8Å are shown as dashed lines. We see about as many highly correlated sequence pairs that are proximal to one another as correlated pairs that are further apart. By contrast, in (A), lines show sequence pairs that are strongly coupled in the Potts model (7), whose model parameters are inferred from the correlations. The fraction of false contact predictions (dashed lines) is reduced considerably. The figures are adapted from [154].

An attractive alternative enlists evolutionary information: Suppose that we have at our disposal amino acid sequences of a protein as it appears in different related species (so-called orthologs). While the sequences are not identical across species, they preserve to some degree the three-dimensional shape of the protein. Suppose a specific pair of amino acids that interact strongly with each other and bring together parts of the protein that are distal on the linear sequence. Replacing this pair with another, equally strongly interacting pair of amino acids would change the sequence, but leave the structure unchanged. For this reason, we expect sequence differences across species to reflect the structure of the protein. Specifically, we expect correlations of amino acids in positions that are proximal to each other in the three-dimensional structure. In turn, the correlations observed between amino acids at different positions might allow to infer which pairs of amino acids are proximal to each other in three dimensions (the so-called contact map, see Figure 2). The use of such genomic information has recently lead to predictions of the three-dimensional structure of many protein families inaccessible to other methods [166], for a review see [51].

Early work looked at the correlations as a measure of proximity [87,92,129,206]. However, correlations are transitive; if amino acids at sequence sites i and j are correlated due to proximity in the folded state, and j and k are correlated for some reason, i and k will also exhibit correlations, which need not stem from proximity. This problem is addressed by an inverse approach aimed at finding the set of pairwise couplings that lead to the observed correlations or sequences [39,55,58,71,99,138,154,212,238]. Since each sequence position can be taken up by one of 20 amino acids or a gap in the sequence alignment, there are 21^2 correlations at each pair of sequence positions. In [154,238], a statistical model with pairwise interactions is formulated, based on the Hamiltonian

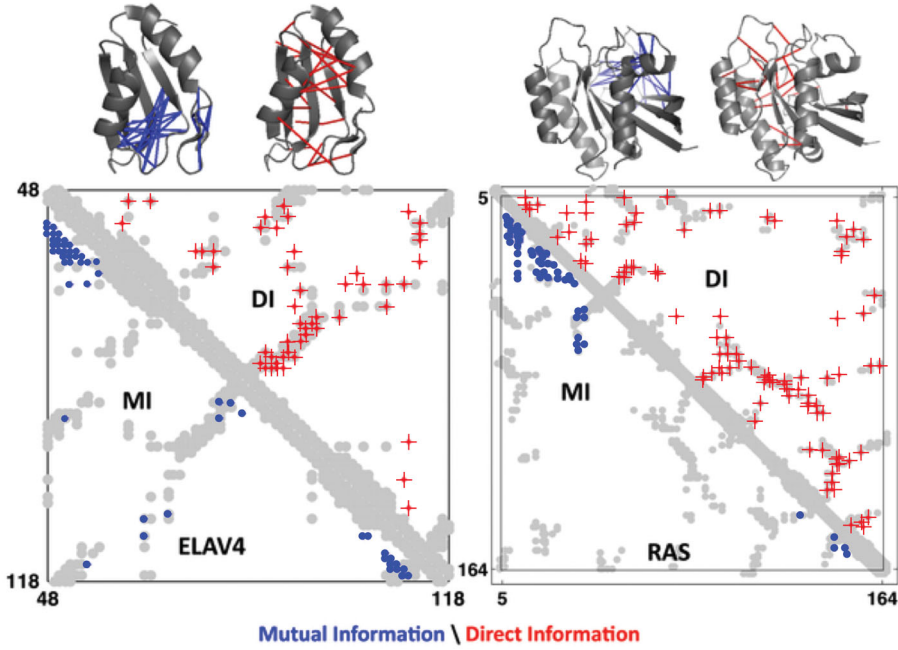$$H = -\sum_{i<j} J_{ij}(s_i, s_j) - \sum_i h_i(s_i). \tag{7}$$

Figure 2. Protein contact maps predicted from evolutionary correlations. The two figures show contact maps for the ELAV4 protein (left) and the RAS protein (right). *x*- and *y*-axes correspond to sequence positions along the linear chain of amino acids. Pairs of sequence positions whose amino acids are in close proximity in the folded protein are indicated by the shaded area (experimental data). Pairs of sequence positions with highly correlated amino acids are shown by small dots (mutual information, bottom triangle). Pairs of sequence positions with high direct information (8) calculated from Equation (7) are shown with large crosses. The coincidence of crosses and shaded area shows excellent agreement between predictions from direct information with the experimentally determined structure of the protein. The figure is adapted from [138].

This Hamiltonian depends on spin variables s_i, one for each sequence position $i = 1, \ldots, N$. Each spin variable can take on one of 21 values, describing the 20 possible amino acids at that sequence position as well as the possibility of a gap (corresponding to an extra amino acid inserted in a particular position in the sequences of other organisms). Each pair of amino acids s_i, s_j in sequence position i,j contributes $J_{ij}(s_i, s_j)$ to the energy. The inverse problem is to find the couplings $J_{ij}(A, B)$ for each pair of sequence positions i,j and pair of amino acids A,B, as well as field $h_i(A)$, such that the amino acid frequencies and correlations observed across species are reproduced. The sequence positions with strong pairwise couplings are then predicted to be proximal in the protein structure. A simple measure of the coupling between sequence positions is the matrix norm (Frobenius norm) $\sum_{s_i, s_j} (J_{ij}(s_i, s_j))^2$. The so-called direct information [238] is an alternative measure based on information theory. A two-site model is defined with $p_{ij}(s_i, s_j) = \exp\{J_{ij}(s_i, s_j) + h_i(s_i) + h_j(s_j)\}/Z_{ij}$. Direct information is the mutual information between the two-site model and a model without correlations between the amino acids

$$DI_{ij} = \sum_{s_i, s_j} p_{ij}(s_i, s_j) \ln \left(\frac{p_{ij}(s_i, s_j)}{p_i(s_i) p_j(s_j)} \right). \tag{8}$$

The Boltzmann distribution resulting from Equation (7) can be viewed as the maximum entropy distribution with one- and two-point correlations between amino acids in different sequence

positions determined by the data. There is no reason to exclude higher order terms in the Hamiltonian (7) describing interactions between triplets of sequence positions, although the introduction of such terms may lead to overfitting. Also, fitting the Boltzmann distribution (7) to sequence data uses no prior information on protein structures; for this reason it is called an *unsupervised method*. Recently, neural network models trained on sequence data *and* protein structures (*supervised learning*) have been very successful in predicting new structures [109,235].

The maximum entropy approach to structure analysis is not limited to evolutionary data. In [251], Zhang and Wolynes analyse chromosome conformation capture experiments and use the observed frequency of contacts between different parts of a chromosome in a maximum entropy approach to predict the structure and topology of the chromosomes.

1.4. *Fitness landscape inference*

The concept of fitness lies at the core of evolutionary biology. Fitness quantifies the average reproductive success (number of offspring) of an organism with a particular genotype, i.e. a particular DNA sequence. The dependence of fitness on the genotype can be visualized as a fitness landscape in a high-dimensional space, where fitness specifies the height of the landscape. As the number of possible sequences grows exponentially with their length, the fitness landscape requires in principle an exponentially large number of parameters to specify, and in turn those parameters need an exponentially growing amount of data to infer.

A suitable model system for the inference of a fitness landscape is HIV proteins, due to the large number of sequences stored in clinical databases and the relative ease of generating mutants and measuring the resulting fitness. In a series of papers, Chakraborty and co-workers proposed a fitness model the so-called Gag protein family (group-specific antigen) of the HIV virus [59,74,135,201]. The model is based on pairwise interactions between amino acids. Retaining only the information whether the amino acid at sequence position i was mutated ($s_i = 1$) with respect to a reference sequence or not ($s_i = 0$), Chakraborty and co-workers suggest a minimal model for the fitness landscape given by the Ising Hamiltonian (2). Again, one can view the landscape (2) as generating the maximum entropy distribution constrained by the observed one- and two-point correlations.

Adding a constant to Equation (2) in order to make fitness (expected number of offspring) non-negative does not alter the resulting statistics. The inverse problem is to infer the couplings J_{ij} and fields h_i from frequencies of amino acids and pairs of amino acids in particular sequence positions observed in HIV sequence data. Of course, it is not clear from the outset that a model with only pairwise interactions can describe the empirical fitness landscape. As a test of this approach, Ferguson et al. [74] compare the prediction of Equation (2) for specific mutants to the results of independent experimental measurements of fitness.

Statistical models of sequences described by pairwise interactions may be useful to model a wide range of protein families with different functions [100], and have been used in other contexts as well. Santolini et al. [194] model the statistics of sequences binding transcription factors using Equation (7), with each spin taking one of four states to characterize the nucleotides A, C, G, T. A similar model is used in [153] to model the sequence diversity of the so-called IgM protein, an antibody which plays a key role in the early immune response. The model with pairwise interactions predicts non-trivial three-point correlations which compare well with those found in the data, see Figure 3.

1.5. *Combinatorial antibiotic treatment*

Antibiotics are chemical compounds which kill specific bacteria or inhibit their growth [115,233]. Mutations in the bacterial DNA can lead to resistance against a particular antibiotic, which is a

$$\frac{P^3(\sigma_i, \sigma_{i+1}, \sigma_{i+2})}{P^1(\sigma_i)P^1(\sigma_{i+1})P^1(\sigma_{i+2})}$$
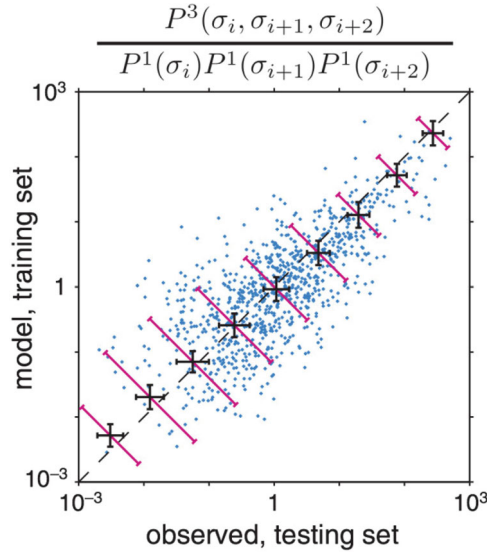

Figure 3. Three-point correlations in an amino acid sequences and their prediction from a model with pairwise interactions. Mora et al. look at the so-called D-region in the IgM protein (maximum length $N = 8$). The D-region plays an important role in immune response. The frequencies at which given triplets of consecutive amino acids occur were compiled (*x*-axis, normalized with respect to the prediction of a model with independent sites). The results are compared to the prediction from a model with pairwise interactions like Equation (2) on the *y*-axis. The figure is taken from [153].

major hazard to public health [127,242]. One strategy to slow down or eliminate the emergence of resistance is to use a combination of antibiotics either simultaneously or in rotation [115,233]. The key problem of this approach is to find combinations of compounds which are particularly effective against a particular strain of bacteria. Trying out all combinations experimentally is prohibitively expensive. Wood et al. [243] use an inverse statistical approach to predict the effect of combinations of several antibiotics from data on the effect of pairs of antibiotics. The available antibiotics are labelled $i = 1, \ldots, N$; in [243], a distribution over continuous variables x_i is constructed, such that $\langle x_i \rangle$ gives the bacterial growth rate when antibiotic i is administered, $\langle x_i x_j \rangle$ gives the growth rate with both i and j are given, etc. for higher moments. Choosing this distribution to be a multi-variate Gaussian $P(\{x_i\}) = \exp[\sum_{i \leq j} J_{ij} x_i x_j + \sum_i h_i x_i]/Z$ results in simple relationships between the different moments, which lead to predictions of the response to drug combinations that are borne out well by experiment [243].

1.6. *Interactions between species and between individuals*

Species exist in various ecological relationships. For instance, individuals of one species hunt and eat individuals of another species. Another example is microorganisms whose growth can be influenced, both positively and negatively, by the metabolic output of other microorganisms. Such relationships form a dense web of ecological interactions between species. Co-culturing and perturbation experiments (for instance species removal) lead to data which may allow the inference of these networks [73,94].

Interactions between organisms exist also at the level of individuals, for instance when birds form a flock, or fish form a school. This emergent collective behaviour is thought to have evolved to minimize the exposure of individuals to predators. In [28,29], a model with pairwise

interactions between the velocities of birds in a flock is constructed. Individual birds labelled $i = 1, \ldots, N$ move with velocity $\vec{v}_i$ in a direction specified by $\vec{s}_i = \vec{v}_i / |\vec{v}_i|$. The statistics of the these directions is modelled by a distribution

$$P(\{\vec{s}_i\}) = \exp \left[\sum_{i,j} J_{ij} \vec{s}_i \cdot \vec{s}_j \right] / Z, \tag{9}$$

where the couplings between the spins $\vec{s}_i$ need to be inferred from the experimentally observed correlations between normalized velocities. This model can be viewed as the maximum entropy distribution constrained by pairwise correlations between normalized velocities. From the point of view of statistical physics, it describes a disordered system of Heisenberg spins. As birds frequently change their neighbours in flight, the couplings are not constant in time and it makes sense to consider couplings that depend on the distance between two individuals [29]. An alternative is to apply the maximum entropy principle to entire trajectories [46].

1.7. *Financial markets*

Market participants exchange commodities, shares in companies, currencies, or other goods and services, usually for money. The change in prices of such goods are often correlated, as the demand for different goods can be influenced by the same events. In [41,42], Bury uses a spin model with pairwise interactions to analyse stock market data. Shares in N different companies are described by binary spin variables, where spin $s_i = 1$ indicates 'bullish' conditions for shares in company i with prices going up at a particular time, and $s_i = -1$ implies 'bearish' conditions with decreasing prices. Couplings J_{ij} describe how price changes in shares i affect changes in the price of j, or how prices are affected jointly by external events. Bury fit stock marked data to this spin model, and found clusters in the resulting matrix of couplings [41]. These clusters correspond to different industries whose companies are traded on the market. In [34], a similar analysis finds that heavy tails in the distribution of inferred couplings are linked to such clusters. Slonim et al. [205] identified clusters in stocks using an information-based metric of stock prices.

2. Equilibrium reconstruction

The applications discussed above can be classified according to the symmetry of pairwise couplings: In network reconstruction, couplings between spins are generally asymmetric, in maximum entropy models they are symmetric. A stochastic dynamics based on symmetric couplings entails detailed balance, leading to a steady state described by the Boltzmann distribution [118], whereas asymmetric couplings lead to an NESS. This distinction shapes the structure of this review: In this section, we discuss the inverse Ising problem in equilibrium, in Section 3 we turn to non-equilibrium scenarios.

2.1. *Definition of the problem*

We consider the Ising model with N binary spin variables $\{s_i = \pm 1\}$, $i = 1, \ldots, N$. Pairwise couplings (or coupling strengths) J_{ij} encode pairwise interactions between the spin variables, and local magnetic fields h_i act on individual spins. The energy of a spin configuration $\boldsymbol{s} \equiv \{s_i\}$ is specified by the Hamiltonian

$$H_{\boldsymbol{J},\boldsymbol{h}}(\boldsymbol{s}) = - \sum_{i<j} J_{ij} s_i s_j - \sum_i h_i s_i. \tag{10}$$

The equilibrium statistics of the Ising model is described by the Boltzmann distribution

$$p(\boldsymbol{s}) = \frac{1}{Z} e^{-H_{J,h}(s)}, \tag{11}$$

where we have subsumed the temperature into couplings and fields such that $k_B T = 1$: The statistics of spins under the Boltzmann distribution $\exp\{-\beta H\}/Z$ depends on couplings, magnetic fields, and temperature only through the products βJ_{ij} and βh_i. As a result, only the products βJ_{ij} and βh_i can be inferred and we set β to 1 without loss of generality. The energy specified by the Hamiltonian (2) or its generalization (7) is thus a dimensionless quantity.

Z denotes the partition function

$$Z(\boldsymbol{J}, \boldsymbol{h}) = \sum_{\boldsymbol{s}} e^{-H_{J,h}(s)}. \tag{12}$$

In such a statistical description of the Ising model, each spin is represented by a random variable. Throughout, we denote a random spin variable by σ, and a particular realization of that random variable by s. This distinction will become particularly useful in the context of non-equilibrium reconstruction in Section 3. The expectation values of spin variables and their functions then are denoted

$$\langle Q(\boldsymbol{\sigma}) \rangle \equiv \sum_{\boldsymbol{s}} p(\boldsymbol{s}) Q(\boldsymbol{s}), \tag{13}$$

where $Q(\boldsymbol{s})$ is some function mapping a spin configuration to a number. Examples are the equilibrium magnetizations $m_i \equiv \langle \sigma_i \rangle = \sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i$ or the pair correlations $\chi_{ij} \equiv \langle \sigma_i \sigma_j \rangle = \sum_{\boldsymbol{s}} p(\boldsymbol{s})(s_i s_j)$. In statistics, the latter observable is called the pair average. We are also interested in the connected correlation $C_{ij} = \chi_{ij} - m_i m_j$, which in statistics is known as the covariance.

The equilibrium statistics of the Ising problem (11) is fully determined by the couplings between spins and the magnetic fields acting on the spins. Collectively, couplings and magnetic fields are the *parameters* of the Ising problem. The *forward Ising problem* is to compute statistical observables such as the magnetizations and correlations under the Boltzmann distribution (11); the couplings and fields are taken as given. The *inverse Ising problem* works in the reverse direction: The couplings and fields are unknown and are to be determined from observations of the spins. The *equilibrium inverse Ising problem* is to infer these parameters from spin configurations sampled independently from the Boltzmann distribution. We denote such a data set of M samples by $D = \{s^\mu\}$ for $\mu = 1, 2, \ldots, M$. (This usage of the term 'sample' appears to differ from how it is used in the statistical mechanics of disordered systems, where a sample often refers to a random choice of the model parameters, not spin configurations. However, it is in line with the inverse nature of the problem: From the point of view of the statistical mechanics of disordered systems in an inverse statistical problem, the 'phase space variables' are couplings and magnetic fields to be inferred, the 'quenched disorder' is spin configurations sampled from the Boltzmann distribution.)

Generally, neither the values of couplings nor the graph structure formed by non-zero couplings is known. Unlike in many instances of the forward problem, the couplings often do not conform to a regular, finite-dimensional lattice; there is no sense of spatial distance between spins. Instead, the couplings might be described by a fully connected graph, with all pairs of spins coupling to each other, generally all with different values of the J_{ij}. Alternatively, most of the couplings might be zero, and the non-zero entries of the coupling matrix might define a structure that is (at least locally) treelike. The graph formed by the couplings might also be highly heterogeneous with few highly connected nodes with many non-zero couplings and many spins coupling only to a few other spins. These distinctions can affect how well specific inference

methods perform, a point we will revisit in Section 2.4, which compares the quality of different methods in different situations.

2.2. *Maximum likelihood*

The inverse Ising problem is a problem of statistical inference [31,131]. At the heart of many methods to reconstruct the parameters of the Ising model is the maximum likelihood framework, which we discuss here.

Suppose a set of observations $x^1, x^2, \ldots, x^M$ drawn from a statistical model $p(x^1, x^2, \ldots, x^M \mid \theta)$. In the case of the Ising model, each observation would be a spin configuration $\boldsymbol{s}$. While the functional form of this model may be known a priori, the parameter θ is unknown to us and needs to be inferred from the observed data. Of course, with a finite amount of data, one cannot hope to determine the parameter θ exactly. The so-called *maximum likelihood (ML) estimator*

$$\theta^{\mathrm{ML}} = \underset{\theta}{\mathrm{argmax}}\, p(x_1, x_2, \ldots, x_M \mid \theta) \tag{14}$$

has a number of attractive properties [57]: In the limit of a large number of samples, θ^{ML} converges in probability to the value θ being estimated. This property is termed consistency. Also for large sample sizes, there is no consistent estimator with a smaller mean-squared error. For a finite number of samples, the ML estimator may, however, be biased, that is, the mean of θ^{ML} over many realizations of the samples does not equal θ (although the difference vanishes with the sample size). The term *likelihood* refers to $p(x_1, x_2, \ldots, x_M \mid \theta)$ viewed as a function of the parameter θ at constant values of the data $x_1, x_2, \ldots, x_M$. The same function at constant θ gives the *probability* of observing the data $x_1, x_2, \ldots, x_M$.

The ML estimator (14) can also be derived using Bayes' theorem [31,131]. In Bayesian inference, one introduces a probability distribution $p(\theta)$ over the unknown parameter θ. This *prior distribution* describes our knowledge prior to receiving the data. Upon accounting for additional information from the data, our knowledge is described by the *posterior* distribution given by Bayes' theorem

$$\begin{aligned}
p(\theta \mid x_1, x_2, \ldots, x_M) &= \frac{p(\theta, x_1, x_2, \ldots, x_M)}{p(x_1, x_2, \ldots, x_M)} \\
&= \frac{p(x_1, x_2, \ldots, x_M \mid \theta) p(\theta)}{p(x_1, x_2, \ldots, x_M)}.
\end{aligned} \tag{15}$$

For the case where θ is a priori uniformly distributed (describing a scenario where we have no prior knowledge of the parameter value), the posterior probability distribution of the parameter conditioned on the observations $p(\theta | x_1, x_2, \ldots, x_M)$ is proportional to $p(x_1, x_2, \ldots, x_M \mid \theta)$. (This line of argument only works if the parameter space is bounded, so a uniform prior can be defined). Then, the parameter value maximizing the probability density $p(\theta \mid x_1, x_2, \ldots, x_M)$ is given by the ML estimator (14). Maximizing the logarithm of the likelihood function, termed the log-likelihood function, leads to the same parameter estimate, because the logarithm is a strictly monotonic function. As the likelihood scales exponentially with the number of samples, the log-likelihood is more convenient to use. (This is simply the convenience of not having to deal with very small numbers: the logarithm is not linked to the quenched average considered in the statistical mechanics of disordered systems; there is no average involved and the likelihood depends on both the model parameters and the data.)

We now apply the principle of ML to the inverse Ising problem. Assuming that the configurations in the dataset were sampled independently from the Boltzmann distribution (1), the

log-likelihood of the model parameters given the observed configurations $D = \{s^\mu\}$ is derived easily:

$$
\begin{aligned}
L_D(\boldsymbol{J}, \boldsymbol{h}) &= \frac{1}{M} \ln p(D|\boldsymbol{J}, \boldsymbol{h}) \\
&= \sum_{i<j} J_{ij} \frac{1}{M} \sum_\mu s_i^\mu s_j^\mu + \sum_i h_i \frac{1}{M} \sum_i s_i^\mu - \ln Z(\boldsymbol{J}, \boldsymbol{h}) \\
&= \sum_{i<j} J_{ij} \langle \sigma_i \sigma_j \rangle^D + \sum_i h_i \langle \sigma_i \rangle^D - \ln Z(\boldsymbol{J}, \boldsymbol{h})
\end{aligned}
\tag{16}
$$

gives the log-likelihood per sample, a quantity of order zero in M since the likelihood scales exponentially with the number of samples. The sample averages of spin variables and their functions are defined by

$$
\langle Q \rangle^D = \frac{1}{M} \sum_\mu Q(s^\mu).
\tag{17}
$$

Beyond the parameters of the Ising model, the log-likelihood (16) depends only on the correlations between pairs of spins observed in the data $\langle \sigma_i \sigma_j \rangle^D$ and the magnetizations $\langle \sigma_i \rangle^D$. To determine the ML estimates of the model parameters, we thus only need the pair correlations and magnetizations observed in the sample (sample averages); at least in principle, further observables are superfluous. In the language of statistics, these sets of sample averages provide *sufficient statistics* to determine the model parameters.

The log-likelihood (16) has a physical interpretation: The first two terms are the sample average of the (negative of the) energy, the second term adds the free energy. Thus, the log-likelihood is the (negative of the) entropy of the Ising system, based on the sample estimate of the energy. We will further discuss this connection in Section 2.2.5.

A second interpretation of the log-likelihood is based on the difference between the Boltzmann distribution (11) and the empirical distribution of the data in the sample D, denoted $p^D(s) \equiv (1/M) \sum_\mu \delta_{s^\mu, s}$. The difference between two probability distributions $p(s)$ and $q(s)$ can be quantified by the Kullback–Leibler (KL) divergence

$$
KL(p|q) = \sum_s p(s) \ln \frac{p(s)}{q(s)},
\tag{18}
$$

which is non-negative and reaches zero only when the two distributions are identical [56]. The KL divergence between the empirical distribution and the Boltzmann distribution is

$$
\begin{aligned}
KL(p^D|p) &= \sum_s p^D(s) \ln \frac{p^D(s)}{p(s)} \\
&= -L_D(\boldsymbol{J}, \boldsymbol{h}) + \sum_s p^D(s) \ln p^D(s).
\end{aligned}
\tag{19}
$$

The second term (the negative empirical entropy) is independent of the model parameters; the best match between the Boltzmann distribution and the empirical distribution (minimal KL divergence) is thus achieved when the likelihood (16) is maximal.

Above, we derived the principle of ML (14) from Bayes' theorem under the assumption that the model parameter θ is sampled from a uniform prior distribution. Suppose that we had the prior information that the parameter θ was taken from some non-uniform distribution, the

posterior distribution would then acquire an additional dependence on the parameter. In the case of the inverse Ising problem, prior information might, for example, describe the sparsity of the coupling matrix, with a suitable prior $p_J \sim \exp[-\gamma \sum_{i<j} |J_{ij}|]$ that assigns small probabilities to large entries in the coupling matrix. The resulting (log) posterior is

$$\ln p(\boldsymbol{J}, \boldsymbol{h} \mid \mathrm{D}) = ML_{\mathrm{D}}(\boldsymbol{J}, \boldsymbol{h}) - \gamma \sum_{i<j} |J_{ij}| \tag{20}$$

up to terms that do not depend on the model parameters. The maximum of the posterior is now no longer achieved by maximizing the likelihood, but involves a second term that penalizes coupling matrices with large entries. Maximizing the posterior with respect to the parameters no longer makes the Boltzmann distribution as similar to the empirical distribution as possible, but strikes a balance between making these distributions similar while avoiding large values of the couplings. In the context of inference, the second term is called a regularization term. Different regularization terms have been used, including the absolute-value term in Equation (20) as well as a penalty on the square values of couplings $\sum_{i<j} J_{ij}^2$ (called ℓ_1- and ℓ_2-regularizers, respectively). One standard way to determine the value of the regularization coefficient γ is to cross-validate with a part of the data that is initially withheld, that is to probe (as a function of γ) how well the model can predict aspects of the data not yet used to infer the model parameters [93].

2.2.1. *Exact maximization of the likelihood*

The ML estimate of couplings and magnetic fields

$$\{\boldsymbol{J}^{\mathrm{ML}}, \boldsymbol{h}^{\mathrm{ML}}\} = \mathrm{argmax} L_{\mathrm{D}}(\boldsymbol{J}, \boldsymbol{h}) \tag{21}$$

has a simple interpretation. Since $\ln Z(\boldsymbol{J}, \boldsymbol{h})$ serves as a generating function for expectation values under the Boltzmann distribution, we have

$$\begin{aligned}
\frac{\partial L_{\mathrm{D}}}{\partial h_i}(\boldsymbol{J}, \boldsymbol{h}) &= \langle \sigma_i \rangle^{\mathrm{D}} - \langle \sigma_i \rangle, \\
\frac{\partial L_{\mathrm{D}}}{\partial J_{ij}}(\boldsymbol{J}, \boldsymbol{h}) &= \langle \sigma_i \sigma_j \rangle^{\mathrm{D}} - \langle \sigma_i \sigma_j \rangle.
\end{aligned} \tag{22}$$

At the maximum of the log-likelihood, these derivatives are zero; the ML estimate of the parameters is reached when the expectation values of pair correlations and magnetizations under the Boltzmann statistics match their sample averages

$$\begin{aligned}
\langle \sigma_i \rangle &= \langle \sigma_i \rangle^{\mathrm{D}}, \\
\langle \sigma_i \sigma_j \rangle &= \langle \sigma_i \sigma_j \rangle^{\mathrm{D}}.
\end{aligned} \tag{23}$$

The log-likelihood (16) turns out to be a concave function of the model parameters, see Section 2.2.2. Thus, in principle, it can be maximized by a convex optimization algorithm. One particular way to reach the maximum of the likelihood is a gradient-descent algorithm called Boltzmann machine learning [1]. At each step of the algorithm fields and couplings are updated

according to

$$h_i^{n+1} = h_i^n + \eta \frac{\partial L_D}{\partial h_i}(\boldsymbol{J}^n, \boldsymbol{h}^n), \tag{24}$$

$$J_{ij}^{n+1} = J_{ij}^n + \eta \frac{\partial L_D}{\partial J_{ij}}(\boldsymbol{J}^n, \boldsymbol{h}^n). \tag{25}$$

The parameter η is the learning rate of the algorithm, which has Equation (23) as its fixed point.

In order to calculate the expectation values $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$ on the left-hand side of these equations, one needs to perform thermal averages of the form (13) over all 2^N configurations, which is generally infeasible for all but the smallest system sizes. Analogously, when maximizing the log-likelihood (16) directly, the partition function is the sum over 2^N terms. Moreover, the expectation values or the partition function needs to be evaluated many times during an iterative search for the solution of Equation (23) or the maximum of the likelihood. As a result, also numerical sampling techniques such as Monte Carlo sampling are cumbersome, but have been used for moderate system sizes [36]. Habeck [91] proposes a Monte Carlo sampler that draws model parameters from the posterior distribution. A recent algorithm uses information contained in the shape of the likelihood maximum to speed up the convergence [75]. An important development in machine learning has led to the so-called restricted Boltzmann machines, where couplings form a symmetric and bipartite graph. Variables fall into two classes, termed 'visible' and 'hidden', with couplings never linking variables of the same class. This allows fast learning algorithms [76] at the expense of additional hidden variables.

We stress that the difficulty of maximizing the likelihood is associated with the restriction of our input to the first two moments (magnetizations and correlations) of the data. On the one hand, this restriction is natural, as the likelihood only depends on these two moments. On the other hand, *computationally efficient methods* have been developed that effectively use correlations in the data beyond the first two moments. An important example is pseudolikelihood, which we will discuss in Section 2.3. Other learning techniques that sidestep the computation of the partition function include score matching [101] and minimum probability flow [207]. Also, when the number of samples is small (compared to the number of spins), the likelihood need no longer be the best quantity to optimize.

2.2.2. *Uniqueness of the solution*

We will show that the log-likelihood (16) is a strictly concave function of the model parameters (couplings and magnetic fields). As the space of parameters is convex, the maximum of the log-likelihood is unique.

We use the shorthands $\boldsymbol{\lambda} = \{\boldsymbol{J}, \boldsymbol{h}\}$ and $\{Q_k(\boldsymbol{s})\} = \{s_i s_j, \ s_i\}$ for the model parameters and the functions coupling to them and write the Boltzmann distribution as

$$p(\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{\lambda})} e^{\sum_k \lambda_k Q_k(\boldsymbol{s})}. \tag{26}$$

For such a general class of exponential distributions [93], the second derivatives of the log-likelihood L_D with respect to a parameters obey

$$-\frac{\partial^2 L_D}{\partial \lambda_i \partial \lambda_j}(\boldsymbol{\lambda}) = \langle Q_i Q_j \rangle - \langle Q_i \rangle \langle Q_j \rangle. \tag{27}$$

This matrix of second derivatives is non-negative (has no negative eigenvalues) since $\sum_{ij} (\langle Q_i Q_j \rangle - \langle Q_i \rangle \langle Q_j \rangle) x_i x_j = \langle [\sum_k (x_k Q_k - \langle x_k Q_k \rangle)]^2 \rangle \geq 0$ for all x_i. If no non-trivial linear

combination of the observables Q_k has vanishing fluctuations, the Hessian matrix is even positive definite. For the inverse Ising problem, there are indeed no non-trivial linear combinations of the spin variables σ_i and pairs of spins variables $\sigma_i \sigma_j$ that do not fluctuate under the Boltzmann measure, unless some of the couplings or fields are infinite. As a result, the maximum of the likelihood, if it exists, is unique. However, it can happen that the maximum lies at infinite values of some of the parameters (for instance, when the samples contain only positive values of a particular spin, the ML value of the corresponding magnetic field is infinite). These divergences can be avoided with the introduction of a regularization term, see Section 2.2.

2.2.3. *Maximum entropy modelling*

The Boltzmann distribution in general and the Ising Hamiltonian (2) in particular can be derived from information theory and the principle of maximum entropy. This principle has been invoked in neural modelling [195], protein structure determination [238], and DNA sequence analysis [153]. In this section, we discuss the statistical basis of Shannon's entropy, the principle of maximum entropy, and their application to inverse statistical modelling.

Consider M distinguishable balls, each to be placed in a box with R compartments. The number of ways of placing the balls such that n_r balls are in the rth compartment ($r \in 1, \ldots, R$) is

$$W = \frac{M!}{\prod_{r=1}^{R} n_r!}, \tag{28}$$

with $\sum_{r=1}^{R} n_r = M$. For large M, we write $n_r = M q_r$ and exploit Stirling's formula $n_r! \approx e^{n_r} n_r^{n_r}$, yielding the Gibbs entropy

$$\ln W / M \approx - \sum_{r=1}^{R} q_r \ln q_r. \tag{29}$$

This combinatorial result forms the basis of equilibrium statistical physics in the classic treatment due to Gibbs and can be found in standard textbooks. In the context of statistical physics, each of the R compartments corresponds to a microstate of a system, and each microstate r is associated with energy E_r. The M balls in the compartments describe a set of copies of the system, or a so-called *ensemble of replicas*. The replicas may exchange energy with each other, while the ensemble of replicas itself is isolated and has a fixed total energy ME (and possibly other conserved quantities). In this way, the replicas can be thought of as providing a heat-bath for each other. If we assume that each state of the ensemble of replicas with a given total energy is equally likely, the statistics of q_r is dominated by a sharp maximum of W as a function of the q_r, subject to the constraint $\sum_r q_r = 1$ and $\sum_r E_r q_r = E$. Using Lagrange multipliers to maximize Equation (29) subject to these constraints yields the Boltzmann distribution [107].

This seminal line of argument can also be used to derive Shannon's information entropy (3). The argument is due to Wallis and is recounted in [108]. Suppose that we want to find a probability distribution p_r compatible with a certain constraint, for instance a specific expectation value $\sum_r p_r E_r = E$ to within some small margin of error. Consider M independent casts of a fair die with R faces. We denote the number of times outcome r is realized in these throws as n_r. The probability of a particular set $\{n_r\}$ is

$$\frac{M!}{\prod_{r=1}^{R} n_r!} \prod_{r=1}^{R} (1/R)^{n_r}. \tag{30}$$

In the limit of large M, the logarithm of this probability is $-M \sum_r q_r \ln q_r - M \ln R$ with $q_r = n_r / M$.

Each set of M casts defines one instance of the $\{n_r\}$. In most instances, the constraint will not be realized. For those (potentially rare) instances obeying the constraint, we can ask what are the most likely values of n_r, and correspondingly q_r. Maximising Shannon's information entropy $-\sum_r q_r \ln q_r$ subject to the constraint and the normalization $\sum_r q_r = 1$ gives the so-called *maximum entropy estimate* of p_r. If the underlying set of probabilities (the die with R faces) differs from the uniform distribution, so outcome r occurs with probability q_r^0, it is not the entropy but the relative entropy $-\sum_r q_r \ln(q_r/q_r^0)$ that is to be maximized. Up to a sign, this is the KL divergence (18) between q_r and q_r^0.

The maximum entropy estimate can be used to approximate an unknown probability distribution q_r that is under-sampled. Suppose that data are sampled several times from some unknown probability distribution. With a sufficient number of samples M, the distribution q_r can be easily determined from frequency counts $q_r = n_r/M$. Often this is not feasible; if $q_r M \ll 1$, n_r fluctuates strongly from one set of samples to the next. This situation appears naturally when the number of possible outcomes R grows exponentially with the size of the system, see e.g. Section 1.1. Nevertheless, the data may be sufficient to pin down one or several expectation values. The maximum entropy estimate has been proposed as the most unbiased estimate of the unknown probability distribution compatible with the observed expectation values [108]. For a discussion of the different ways to justify the maximum entropy principle, and derivations based on robust estimates, see [219].

Many applications of the maximum entropy estimate are in image analysis and spectral analysis [90], for reviews in physics and biology see [14,30,151], and for critical discussion see [5,230].

The connection between maximum entropy and the inverse Ising problem is simple: For a set of N binary variables, the distribution with given first and second moments maximizing the information entropy is the Boltzmann distribution (1) with the Ising Hamiltonian (2). We use Lagrange multipliers to maximize the information entropy (3) subject to the normalization condition and the constraints on the first and second moments (magnetizations and pair correlations) of $p(s)$ to be $\boldsymbol{m}$ and $\boldsymbol{\chi}$. Setting the derivatives of

$$\sum_s -p(s)\ln p(s) + \eta\left[1 - \sum_s p(s)\right] + \sum_i h_i\left[m_i - \sum_s p(s)s_i\right] + \sum_{i<j} J_{ij}\left[\chi_{ij} - \sum_s p(s)s_i s_j\right]$$

(31)

with respect to $p(s)$ to zero yields the Ising model (1). The Lagrange multipliers $\boldsymbol{h}$ and $\boldsymbol{J}$ need to be chosen to reproduce the first and second moments (magnetizations and correlations) of the data and can be interpreted as magnetic fields and couplings between spins.

While this principle appears to provide a statistical foundation to the model (1), there is no a priori reason to disregard empirical data beyond the first two moments. Instead, the pairwise couplings result from the particular choice of making the probability distribution match the first two moments of the data. The reasons for this step may be different in different applications.

- Moments beyond the first and second may be poorly determined by the data. Conversely, with an increasing number of samples, the determination of higher order correlations and hence interactions between triplets of spin variables *etc.* becomes viable.
- The data may actually be generated by an equilibrium model with (at most) pairwise interactions between spin variables. This need not be obvious from observed correlations of any order, but can be tested by comparing three-point correlations predicted by a model with pairwise couplings to the corresponding correlations in the data. Examples are found in sequence analysis, where population dynamics leads to an equilibrium steady state [24,199]

and the energy can often be approximated by pairwise couplings [153,194]. For a review, see [210]. Surprisingly, also in neural data (not generated by an equilibrium model), three-point correlations are predicted well by a model with pairwise interactions [221,225].

- A model of binary variables interacting via high-order coupling terms $J_{ijk...}s_i s_j s_k \ldots$ can sometimes be approximated surprisingly well by pairwise interactions. This seems to be the case when the couplings are dense, so that each variable appears in several coupling terms [143].

- Often one seeks to describe a subset of n variables $s_1, s_2, \ldots, s_n$ from a larger set of N variables, for instance when only the variables in the subset can be observed. The subset of variables is characterized by effective interactions which stem from interactions between variables in the subset, and from interactions with the other variables. If the subset is sufficiently small, the resulting statistics is often described by a model with pairwise couplings [190].

- The true probability distribution underlying some data may be too complicated to calculate in practice. A more modest goal then is to describe the data using an effective statistical model such as Equation (1), which is tractable and allows the derivation of bounds on the entropy or the free energy. Examples are the description of neural data and gene expression data using the Ising model with symmetric couplings (see Sections 1.1 and 1.2).

- There are also useful models which are computationally tractable but do not maximize the entropy. An example is Gaussian models to generate artificial spike trains with prescribed pair correlations [3,132].

2.2.4. *Information theoretic bounds on graphical model reconstruction*

A particular facet of the inverse Ising problem is graphical model selection. Consider the Ising problem on a graph. A graph is a set of nodes and edges connecting these nodes, and nodes associated with spin variables. Couplings between node pairs connected by an edge are non-zero, couplings between unconnected node pairs are zero. The graphical model selection problem is to recover the underlying graph (and usually also the values of the couplings) from data sampled independently from the Boltzmann distribution. Given a particular number of samples, one can ask with which probability a given method can reconstruct the graph correctly (the reconstruction fluctuates between different realizations of the samples). Notably, there are also *universal* limits on graphical model selection that are independent of a particular method.

In [193], Santhanam and Wainwright derive information-theoretic limits to graphical model selection. Key result is the dependence of the required number of samples on the smallest and on the largest coupling

$$\alpha = \min_{i<j} |J_{ij}|, \quad \beta = \max_{i<j} |J_{ij}| \tag{32}$$

and on the maximum node connectivity (number of neighbours on the graph) d. Reconstruction of the graph, by any method, is impossible if fewer than

$$\max \left\{ \frac{\ln N}{2\alpha \tanh \alpha}, \frac{e^{\beta d} \ln(Nd/4 - 1)}{4d\alpha \, e^{\alpha}}, \frac{d}{8} \ln \left(\frac{N}{8d} \right) \right\} \tag{33}$$

samples are available (the precise statement is of a probabilistic nature, see [193]). If the maximum connectivity d grows with the system size, this result implies that at least $c \max\{d^2, \alpha^{-2}\} \ln N$ samples are required (with some constant c) [193]. The derivation of this and other results is based on Fano's inequality (Fano's lemma) [56], which gives a lower bound for the probability of error of a classification function (such as the mapping from samples to the graph underlying these samples).

2.2.5. *Thermodynamics of the inverse Ising problem*

Calculations in statistical physics are greatly simplified by introducing thermodynamic potentials. In this section, we will discuss the method of thermodynamic potentials for the inverse Ising problem. It turns out that the ML estimation of the fields and couplings is simply a transformation of the thermodynamic potentials.

Recall that the thermodynamic potential most useful for the forward problem, where couplings and magnetic fields are given, is the Helmholtz free energy $F(\boldsymbol{J}, \boldsymbol{h}) = -\ln Z(\boldsymbol{J}, \boldsymbol{h})$. Derivatives of this free energy give the magnetizations, correlations, and other observables. The thermodynamic potential most useful for the inverse problem, where the pair correlations $\boldsymbol{\chi}$ and magnetizations $\boldsymbol{m}$ are given, is the Legendre transform of the Helmholtz free energy with respect to both couplings and fields [53,54,200]

$$S(\boldsymbol{\chi}, \boldsymbol{m}) = \min_{\boldsymbol{J}, \boldsymbol{h}} \left[-\sum_i h_i m_i - \sum_{i<j} J_{ij} \chi_{ij} - F(\boldsymbol{J}, \boldsymbol{h}) \right]. \tag{34}$$

This thermodynamic potential is readily recognized as the entropy function; up to a sign, it gives the ML (16) of the model parameters. The transformation (34) thus provides a link between the inference via ML and the statistical physics of the Ising model as described by its Helmholtz free energy. The couplings and the fields are found by differentiation

$$
\begin{aligned}
J_{ij} &= -\frac{\partial S}{\partial \chi_{ij}}(\boldsymbol{\chi}, \boldsymbol{m}), \\
h_i &= -\frac{\partial S}{\partial m_i}(\boldsymbol{\chi}, \boldsymbol{m}),
\end{aligned}
\tag{35}
$$

where the derivatives are evaluated at the sample correlations and magnetizations. These relationships follow from the inverse transformation of (34)

$$F(\boldsymbol{J}, \boldsymbol{h}) = \min_{\boldsymbol{\chi}, \boldsymbol{m}} \left[-\sum_i h_i m_i - \sum_{i<j} J_{ij} \chi_{ij} - S(\boldsymbol{\chi}, \boldsymbol{m}) \right], \tag{36}$$

by setting derivatives of the term in square brackets with respect to $\boldsymbol{\chi}$ and $\boldsymbol{m}$ to zero.

In practice, performing the Legendre transformation of both $\boldsymbol{h}$ and $\boldsymbol{J}$ is often not necessary; derivatives of the Helmholtz free energy $F(\boldsymbol{J}, \boldsymbol{h})$ with respect to $\boldsymbol{J}$ can also be generated by differentiating with respect to $\boldsymbol{h}$, e.g.

$$\frac{\partial F}{\partial J_{ij}}(\boldsymbol{J}, \boldsymbol{h}) = \frac{\partial^2 F}{\partial h_i \partial h_j}(\boldsymbol{J}, \boldsymbol{h}) - \frac{\partial F}{\partial h_i}(\boldsymbol{J}, \boldsymbol{h}) \frac{\partial F}{\partial h_j}(\boldsymbol{J}, \boldsymbol{h}). \tag{37}$$

The thermodynamics of the inverse problem can thus be reduced to a single Legendre transform of the Helmholtz free energy, yielding the Gibbs free energy

$$G(\boldsymbol{J}, \boldsymbol{m}) = \max_{\boldsymbol{h}} \left[\sum_i h_i m_i + F(\boldsymbol{J}, \boldsymbol{h}) \right]. \tag{38}$$

The magnetic fields are given by the first derivative of the Gibbs free energy

$$h_i = \frac{\partial G}{\partial m_i}(\boldsymbol{J}, \boldsymbol{m}). \tag{39}$$

To infer the couplings, we consider the second derivatives of Gibbs potential, which give

$$\frac{\partial^2 G}{\partial m_j \partial m_i}(\boldsymbol{J}, \boldsymbol{m}) = (\boldsymbol{C}^{-1})_{ij}, \tag{40}$$

where $\boldsymbol{C}$ is the matrix of connected correlations $C_{ij} \equiv \chi_{ij} - m_i m_j$. Equation (40) follows from the inverse function theorem,

$$\left[\frac{\partial(h_1, \ldots, h_N)}{\partial(m_1, \ldots, m_N)}\right]_{ij} = \left[\left(\frac{\partial(m_1, \ldots, m_N)}{\partial(h_1, \ldots, h_N)}\right)^{-1}\right]_{ij}, \tag{41}$$

and linear response theory

$$C_{ij} = \frac{\partial m_j}{\partial h_i}(\boldsymbol{J}, \boldsymbol{h}) = -\frac{\partial^2 F}{\partial h_i \partial h_j}(\boldsymbol{J}, \boldsymbol{h}), \tag{42}$$

which links the susceptibility of the magnetization to a small change in the magnetic field with the connected correlation [209].

The result (40) turns out to be central to many methods for the inverse Ising problem. The left-hand side of this expression is a function of J_{ij}. If the Gibbs free energy (38) can be evaluated or approximated, Equation (40) can be solved to yield the couplings. Similarly, Equation (39) with the estimated couplings and the sample magnetizations gives the magnetic fields, completing the reconstruction of the parameters of the Ising model.

2.2.6. *Variational principles*

For most systems, neither the free energy $F(\boldsymbol{J}, \boldsymbol{h})$ nor other thermodynamic potentials can be evaluated. However, there are many approximation schemes for $F(\boldsymbol{J}, \boldsymbol{h})$ [164], which lead to approximations for the entropy and the Gibbs free energy. Direct approximation schemes for $S(\chi, \boldsymbol{m})$ and $G(\boldsymbol{J}, \boldsymbol{m})$ have also be formulated within the context of the inverse Ising problem. The key idea behind most of these approximations is the variational principle.

The variational principle for the free energy is

$$F(\boldsymbol{J}, \boldsymbol{h}) = \min_q\{U[q] - S[q]\} \equiv \min_q F[q], \tag{43}$$

where q denotes a probability distribution over spin configurations. $U[q] \equiv \langle H \rangle_q$ and $S[q] \equiv \langle \ln q \rangle_q$ and the minimization is taken over all distributions q. This principle finds its origin in information theory. Take an arbitrary trial distribution $q(\boldsymbol{s})$, the KL divergence (18) quantifies the difference between q and the Boltzmann distribution p is positive and vanishes if and only if $q = p$ [56]. One then arrives directly at Equation (43) when rewriting $\mathrm{KL}(q|p) = U[q] - S[q] - F(\boldsymbol{J}, \boldsymbol{h})$.

We will refer to $F[q] \equiv U[q] - S[q]$ as the *functional* Helmholtz free energy, also called the non-equilibrium free energy in the context of non-equilibrium statistical physics [170]. Another term in use is 'Gibbs free energy' [164], which we have reserved for the thermodynamic potential (38).

So far, nothing has been gained as the minimum is over all possible distributions q, including the Boltzmann distribution itself. A practical approximation arises when a constraint is put on q, leading to a family of *trial distributions* q. Often the minimization can then be carried out over that family, yielding an upper bound to the Helmholtz free energy [164].

In the context of the inverse problem, it is useful to derive the variational principles for other thermodynamic potentials as well. Using the definition of Gibbs potential (38) and the variational principle for the Helmholtz potential (43), we obtain

$$G(\boldsymbol{J}, \boldsymbol{m}) = \max_{\boldsymbol{h}} \left\{ \sum_i h_i m_i + \min_q \{ U[q] - S[q] \} \right\}. \tag{44}$$

By means of Lagrange multipliers, it is easy to show that this double extremum can be obtained by a single conditional minimization,

$$G(\boldsymbol{J}, \boldsymbol{m}) = \min_{q \in \mathcal{G}} \left\{ - \sum_{i<j} J_{ij} \langle \sigma_i \sigma_j \rangle_q - S[q] \right\}, \tag{45}$$

where the set $\mathcal{G}$ denotes all distributions q with a given $\langle \sigma_i \rangle_q = m_i$ [164]. We will refer to the functional $G[q] = \{ -\sum_{i<j} J_{ij} \langle \sigma_i \sigma_j \rangle_q - S[q] \}$ as the functional Gibbs free energy defined on $\mathcal{G}$.

Similarly, the variational principle can be applied to the entropy function $S(\boldsymbol{\chi}, \boldsymbol{m})$, leading once again to a close relationship between statistical modelling and thermodynamics. The entropy (34) is found to be

$$S(\boldsymbol{\chi}, \boldsymbol{m}) = \max_{q \in \mathcal{S}} \{ S[q] \}, \tag{46}$$

where $\mathcal{S}$ denotes distributions with $\langle \sigma_i \rangle_q = m_i$ and $\langle \sigma_i \sigma_j \rangle_q = \chi_{ij}$. This is nothing but the maximum entropy principle [108]: the variational principle identifies the distribution with the maximum information entropy subject to the constraints on magnetizations and spin correlations, which are set equal to their sample averages (see the section on maximum entropy modelling above).

2.2.7. *Mean-field theory*

As a first demonstration of the variational principle, we derive the mean-field theory for the inverse Ising problem. The starting point is an ansatz for the Boltzmann distribution (11) which factorizes in the sites [164,209,215]

$$p^{\mathrm{MF}}(\boldsymbol{s}) = \prod_i \frac{1 + \tilde{m}_i s_i}{2}, \tag{47}$$

thus making the different spin variables statistically independent of one another. The parameters $\tilde{m}_i$ of this ansatz describe the spin magnetizations; each spin has a magnetization resulting from the effective magnetic field acting on that spin. This effective field arises from its local magnetic field h_i, as well as from couplings with other spin. The mean field giving its name to mean-field theory is the average over typical configurations of the effective field.

Using the mean-field ansatz, we now estimate the Gibbs free energy. Within the mean-field ansatz, the minimization of the variational Gibbs potential (45) is trivial: there is only a single mean-field distribution (47) that satisfies the constraint $\mathcal{G}$ that spins have magnetizations $\boldsymbol{m}$, namely $\tilde{\boldsymbol{m}} = \boldsymbol{m}$. We can thus directly write down the mean-field Gibbs free energy

$$\begin{aligned} G^{\mathrm{MF}}(\boldsymbol{m}, \boldsymbol{J}) = & - \sum_{i<j} J_{ij} m_i m_j \\ & + \sum_i \left[\frac{1 + m_i}{2} \ln \frac{1 + m_i}{2} + \frac{1 - m_i}{2} \ln \frac{1 - m_i}{2} \right]. \end{aligned} \tag{48}$$

The equation for the couplings $\boldsymbol{J}$ follows from the second-order derivative of $G(\boldsymbol{m}, \boldsymbol{J})$, cf. Equation (40)

$$(\boldsymbol{C}^{-1})_{ij} = -J_{ij}^{\mathrm{MF}} \quad (i \neq j). \tag{49}$$

Similarly, the reconstruction of the magnetic field follows from the derivative of $G(\boldsymbol{m}, \boldsymbol{J})$ with respect to m_i, cf. Equation (39),

$$h_i^{\mathrm{MF}} = -\sum_{j \neq i} J_{ij}^{\mathrm{MF}} m_j + \mathrm{artanh} \, m_i. \tag{50}$$

This result establishes a simple relationship between the observed connected correlations and the couplings between spins in terms of the inverse of the correlation matrix. The matrix inverse in Equation (49) is of course much simpler to compute than the maximum over the likelihood (16) and takes only a polynomial number of steps: Gauß-Jordan elimination for inverting an $N \times N$ matrix requires $\mathcal{O}(N^3)$ operations, compared to the exponentially large number of steps to compute a partition function or its derivatives.

The standard route to mean-field reconstruction proceeds somewhat differently, namely via the Helmholtz free energy rather than the Gibbs free energy. Early work to address the inverse Ising problem using the mean-field approximation was performed by Peterson and Anderson [176]. In [112], Kappen and Rodríguez construct the Helmholtz functional free energy $F_{\tilde{\boldsymbol{m}}}^{\mathrm{MF}}(\boldsymbol{J}, \boldsymbol{h})$ given by Equation (43) under the mean-field ansatz (47). $F_{\tilde{\boldsymbol{m}}}^{\mathrm{MF}}(\boldsymbol{J}, \boldsymbol{h})$ is then minimized with respect to the parameters of the mean-field ansatz $\tilde{\boldsymbol{m}}$ by setting its derivatives with respect to $\tilde{\boldsymbol{m}}$ to zero. This yields equations which determine the values m_i^{MF} of the magnetization parameters that minimize the KL divergence between the mean-field ansatz and the Boltzmann distribution, namely the well-known self-consistent equations

$$m_i^{\mathrm{MF}} = \mathrm{th} \left(h_i + \sum_{j \neq i} J_{ij} m_j^{\mathrm{MF}} \right). \tag{51}$$

Using m_i^{MF} as an approximation for the equilibrium magnetizations m_i, one can derive the so-called linear response approximation for the connected correlation function $C_{ij}^{\mathrm{MF-LR}} \equiv (\partial m_i^{\mathrm{MF}} / \partial h_j)(\boldsymbol{h})$. Taking derivatives of the self-consistent Equations (51) with respect to local fields gives

$$\sum_j \left(\frac{\delta_{ij}}{1 - m_i^2} - J_{ij} \right) C_{jk}^{\mathrm{MF-LR}} = \delta_{ik}, \tag{52}$$

where we have used the fact that diagonal terms of the coupling matrix are zero. Identifying the result for the connected correlations $C_{jk}^{\mathrm{MF-LR}}$ with the sample correlations C_{jk} leads to a system of linear equations to be solved for the couplings [112]. However, to obtain Equation (52), we have used that the diagonal elements J_{ii} are zero. With these constraints, the system of Equations (52) becomes over-determined and in general there is no solution. Although different procedures have been suggested to overcome this problem [105,112,185], there seems to be no canonical way out of this dilemma. The most common approach is to ignore the constraints on the diagonal elements altogether and invert Equation (52) to obtain

$$J_{ij}^{\mathrm{MF-LR}} = \frac{\delta_{ij}}{1 - m_i^2} - (\boldsymbol{C}^{-1})_{ij}. \tag{53}$$

This result agrees with the reconstruction via the Gibbs free energy except for the non-zero diagonal couplings, which bear no physical meaning and are to be ignored. No diagonal couplings arise

in the approach based on the Gibbs free energy since Equation (40) with $j = i$ does not involve any unknown couplings J_{ij}.

2.2.8. *The Onsager term and TAP reconstruction*

The variational estimate of the Gibbs free energy (48) can be improved further. In 1977, Thouless, Anderson, and Palmer (TAP) advocated adding a term to the Gibbs free energy

$$G^{\mathrm{TAP}}(\boldsymbol{J},\boldsymbol{m}) = G^{\mathrm{MF}}(\boldsymbol{J},\boldsymbol{m}) - \frac{1}{2}\sum_{i<j} J_{ij}^2 (1 - m_i^2)(1 - m_j^2). \tag{54}$$

This term can be interpreted as describing the effect of fluctuations of a spin variable on the magnetization of that spin via their impact on neighbouring spins [218]. It is called the Onsager term, which we will derive in Section 2.2.13 in the context of a systematic expansion around the mean-field ansatz. For the forward problem, adding this term modifies the self-consistent Equation (51) to the so-called TAP equation

$$m_i^{\mathrm{TAP}} = \mathrm{th}\left(h_i + \sum_{j\neq i} J_{ij} m_j^{\mathrm{TAP}} - m_i^{\mathrm{TAP}} \sum_j J_{ij}^2 (1 - (m_j^{\mathrm{TAP}})^2)\right). \tag{55}$$

In the inverse problem, the TAP free energy (54) gives the equation for the couplings based on Equation (40)

$$(\boldsymbol{C}^{-1})_{ij} = -J_{ij}^{\mathrm{TAP}} - 2(J_{ij}^{\mathrm{TAP}})^2 m_i m_j. \tag{56}$$

Solving this quadratic equation gives the TAP reconstruction [112,214]

$$J_{ij}^{\mathrm{TAP}} = \frac{-2(\boldsymbol{C}^{-1})_{ij}}{1 + \sqrt{1 - 8(\boldsymbol{C}^{-1})_{ij} m_i m_j}}, \tag{57}$$

where we have chosen the solution that coincides with the mean-field reconstruction when the magnetizations are zero. The magnetic fields can again be found by differentiating the Gibbs free energy

$$h_i = \mathrm{artanh}(m_i) - \sum_{j\neq i} J_{ij}^{\mathrm{TAP}} m_j + m_i \sum_{j\neq i} (J_{ij}^{\mathrm{TAP}})^2 (1 - m_j^2). \tag{58}$$

2.2.9. *Couplings without a loop: mapping to the minimum spanning tree problem*

The computational hardness of implementing Boltzmann machine learning (24) comes from the difficulty of computing correlations under the Boltzmann measure, which can require a computational time that scales exponentially with the system size. This scaling originates from the presence of loops in the graph of couplings between the spins. Graphs for which correlations can be computed efficiently are the acyclic graphs or trees, so it comes as no surprise that the first efficient method to solve the inverse Ising problem was developed for trees already in 1968. This was done by Chow and Liu [50] in the context of a product approximation to a multi-variate probability distribution. While the method itself can be used as a crude approximation for models with loops or as reference point for more advanced methods, the exact result by Chow and Liu is of basic interest in itself as it provides a mapping of the inverse Ising problem for couplings forming a tree onto a minimum spanning tree (MST) problem. MST is a core problem in computational complexity theory, for which there are many efficient algorithms. This section on Chow

and Liu's result also provides some of the background needed in Section 2.2.10 on the Bethe ansatz and Section 2.2.11 on belief propagation.

We consider an Ising model whose pairwise couplings form a tree. The graph of couplings may consist of several parts that are not connected to each other (in any number of steps along connected node pairs), or it may form one single connected tree, but it contains no loops. We denote the set of nodes (vertices) associated with a spin of the tree T by V_T and the set of edges (couplings between nodes) by E_T. It is straightforward to show that in this case, the Boltzmann distribution for the Ising model can be written in a pairwise factorized form

$$p_T(\mathbf{s}) = \prod_{i \in V_T} p_i(s_i) \prod_{(i,j) \in E_T} \frac{p_{ij}(s_i, s_j)}{p_i(s_i)p_j(s_j)}$$

$$= \prod_{(ij) \in E_T} p_{ij}(s_i, s_j) \prod_{i \in V_T} p_i(s_i)^{1 - |\partial i|}. \tag{59}$$

∂i denotes the set of neighbours of node i, so $|\partial i|$ is the number of nodes i couples to. The distributions p_i and p_{ij} denote the one-point and two-point marginals of p_T.

The KL divergence (19) between the empirical distribution $p^D(\mathbf{s})$ and p_T is given by

$$\mathrm{KL}(p^D | p_T) = \sum_{\mathbf{s}} p^D(\mathbf{s}) \ln \frac{p^D(\mathbf{s})}{\prod_{i \in V_T} p_i(s_i) \prod_{(i,j) \in E_T} \frac{p_{ij}(s_i, s_j)}{p_i(s_i)p_j(s_j)}}. \tag{60}$$

For a given tree, it is straightforward to show that the KL divergence is minimized when the marginals p_i and p_{ij} equal the empirical marginals p_i^D and p_{ij}^D. This gives

$$\min_{\{p_i, p_{ij}\}} \mathrm{KL}(p^D | p_T) = -H + \sum_{i \in V_T} H_i - \sum_{(ij) \in E_T} I_{ij}, \tag{61}$$

where $H = -\sum_{\{\mathbf{s}\}} p^D(\mathbf{s}) \ln p^D(\mathbf{s})$ is the entropy of the empirical distribution p^D, $H_i = -\sum_{s_i} p_i^D(s_i) \ln p_i^D(s_i)$ is the single site entropy and I_{ij} is the mutual information between a pair of spins

$$I_{ij} = \sum_{s_i, s_j} p_{ij}^D(s_i, s_j) \ln \frac{p_{ij}^D(s_i, s_j)}{p_i^D(s_i)p_j^D(s_j)}. \tag{62}$$

Assuming that the graph of couplings is an (unknown) tree and that empirical estimates of all pairs of mutual informations are available, the inverse Ising problem can then be solved by minimizing the KL divergence (60) over all possible N^{N-2} trees T. The first two terms in Equation (61) do not depend on the choice of the tree, so only the last term needs to be minimized over, which is a sum over local terms on the graph. The optimal tree topology T_{opt} is thus given by

$$E_{T_{\mathrm{opt}}} = \underset{\{E_T\}}{\mathrm{argmin}} \left(-\sum_{(ij) \in E_T} I_{ij} \right).$$

This is where the mapping onto MST problem emerges: T_{opt} connects all vertices of the graph and its edges are such that the total sum of their weights is minimal. In our case, each edge weight is the (negative) pairwise mutual information I_{ij} between spin variables. Finding the MST does not require an infeasible exploration of the space of all possible trees. On the contrary, it can be found in a number of steps bounded by $\mathcal{O}(|V|^2 \ln |V|)$ by greedy iterative procedures

which identify the optimal edges to be added at each step (V is the set of nodes in the data). The most famous algorithms for the MST problem date back to the 1950s and are known under the names of Prim's algorithm, Kruskal's algorithms, and Boruvka's algorithm (see, eg. [150]). In practice, one has to compute the empirical estimates of the mutual information from samples and then proceed with one of the above algorithms. An interesting observation which makes the Chow–Liu approach even easier to apply is that one may use as edge weights also the connected correlations between spins [50].

Once the optimal tree T_{opt} has been identified, we still need to find the optimal values of the couplings J_{ij} of the Ising model. This is, however, an easy task: the factorized form of the probability measure over the tree (59) allows one to compute the couplings using the independent pair approximation, see Section 2.2.12.

2.2.10. *The Bethe–Peierls ansatz*

The factorizing probability distribution (59) can also be used as an ansatz in situations where the graph of couplings is not a tree. In this context,

$$p^{BP}(\{s_i\}) = \prod_i p_i(s_i) \prod_{(i,j) \in E} \frac{p_{ij}(s_i, s_j)}{p_i(s_i)p_j(s_j)} \tag{63}$$

is called the Bethe–Peierls ansatz [27,174]. $(i,j) \in E$ runs over pairs of interacting spins, or equivalently over edges in the graph of couplings. One can parameterize the marginal distribution p_i and p_{ij} using the magnetization parameters $\tilde{m}_i$ and the (connected) correlation parameters $\tilde{C}_{ij}$,

$$p_i(s_i) = \frac{1 + \tilde{m}_i s_i}{2},$$

$$p_{ij}(s_i, s_j) = \frac{(1 + \tilde{m}_i s_i)(1 + \tilde{m}_j s_j) + \tilde{C}_{ij} s_i s_j}{4} \tag{64}$$

subject to the constraints

$$-1 \le \tilde{m}_i \le 1,$$

$$-1 + |\tilde{m}_i + \tilde{m}_i| \le \tilde{C}_{ij} + \tilde{m}_i \tilde{m}_j \le +1 - |\tilde{m}_i - \tilde{m}_j|. \tag{65}$$

The Bethe–Peierls ansatz can be compared to the mean-field ansatz (47), which assigns a magnetization (or an effective field) to each spin. The Bethe–Peierls ansatz goes one step further; it assigns to each *coupled pair of spins* correlations as well as magnetizations. These correlations and magnetizations are then determined self-consistently. An important feature of the Bethe–Peierls ansatz (63) is that its Shannon entropy (3) can be decomposed into spin pairs

$$S[p^{BP}] = \sum_i S[p_i] + \sum_{(i,j)} (S[p_{ij}] - S[p_i] - S[p_j]). \tag{66}$$

For graphs containing loops, the entropy generally contains terms involving larger sets of spins than pairs, a situation we will discuss in Section 2.2.12.

The Bethe–Peierls ansatz is well defined, and indeed exact, when the couplings form a tree. When the graph of couplings contains loops, the probability distribution (63) is not normalized and the ansatz is not well defined. In that case, the Bethe–Peierls ansatz is an uncontrolled approximation, although recently progress has been made in the control of the resulting error [49,169]. We start with the assumption that there are no loops.

To address the inverse Ising problem, we use the Bethe–Peierls ansatz (63) as a variational ansatz to minimize the functional Gibbs free energy (45). Again the constraint $\mathcal{G}$ in (45) implies $\tilde{\boldsymbol{m}} = \boldsymbol{m}$. The remaining minimization is over the correlation parameters $\tilde{\boldsymbol{C}}$,

$$G^{\text{BP}}(\boldsymbol{J}, \boldsymbol{m}) = \min_{\tilde{C}} \left\{ -\sum_{(i,j)} J_{ij} \langle \sigma_i \sigma_j \rangle_{p^{\text{BP}}} - S[p^{\text{BP}}] \right\} \tag{67}$$

and yields

$$\begin{aligned}
G^{\text{BP}}(\boldsymbol{J}, \boldsymbol{m}) = &-\sum_{(ij)} J_{ij}(C_{ij}^{\text{BP}} + m_i m_j) \\
&+ \sum_i (1 - z_i) \sum_{s_i} \frac{1 + m_i s_i}{2} \ln \frac{1 + m_i s_i}{2} \\
&+ \sum_{(i,j)} \sum_{s_i, s_j} \frac{(1 + m_i s_i)(1 + m_j s_j) + C_{ij}^{\text{BP}} s_i s_j}{4} \\
&\times \ln \frac{(1 + m_i s_i)(1 + m_j s_j) + C_{ij}^{\text{BP}} s_i s_j}{4},
\end{aligned} \tag{68}$$

where $\boldsymbol{C}^{\text{BP}}$ ist the optimal value of $\tilde{\boldsymbol{C}}$ and satisfies

$$J_{ij} = \sum_{s_i, s_j} \frac{s_i s_j}{4} \ln \frac{(1 + m_i s_i)(1 + m_j s_j) + C_{ij}^{\text{BP}} s_i s_j}{4}. \tag{69}$$

z_i denotes the number of neighbours (interaction partners with non-zero couplings) of node i. From Equation (40), the equation for the couplings can be found again by equating the second derivative of the Gibbs free energy with the inverse of the correlation matrix, that is,

$$(\boldsymbol{C}^{-1})_{ij} = \frac{C_{ij}^{\text{BP}}}{(C_{ij}^{\text{BP}})^2 - (1 - m_i^2)(1 - m_j^2)} \quad (j \neq i). \tag{70}$$

This quadratic equation can be solved for C_{ij}^{BP}; inserting the solution

$$C_{ij}^{\text{BP}} = \frac{1}{2} \left\{ \frac{1}{(\boldsymbol{C}^{-1})_{ij}} - \sqrt{\frac{1}{(\boldsymbol{C}^{-1})_{ij}^2} - 4(1 - m_i^2)^2(1 - m_j^2)^2} \right\} \tag{71}$$

for $(\boldsymbol{C}^{-1})_{ij} \neq 0$ and $C_{ij}^{\text{BP}} = 0$ for $(\boldsymbol{C}^{-1})_{ij} = 0$ into Equation (69) gives the couplings of the Bethe–Peierls reconstruction [160,185]. For the special case $m_i = 0$, one obtains a particularly simple result

$$J_{ij}^{\text{BP}} = -\tfrac{1}{2} \operatorname{arsinh}[2(\boldsymbol{C}^{-1})_{ij}] \quad (j \neq i). \tag{72}$$

In graph theory, this formula can be related to the expression for the distance in a tree whose links carry weights specified by pair correlations between spin pairs [15]. Correspondingly, the

magnetic fields follow from the first derivative of the Gibbs free energy as in Equation (39) giving

$$h_i^{\text{BP}} = (1 - z_i)\text{artanh}\, m_i - \sum_{j \in \partial i} J_{ij}^{\text{BP}} m_j$$

$$+ \sum_{j \in \partial i} \sum_{s_i, s_j} \frac{s_i + m_j s_i s_j}{4} \ln \frac{(1 + m_i s_i)(1 + m_j s_j) + C_{ij}^{\text{BP}} s_i s_j}{4}. \tag{73}$$

In Section 2.2.13, we will show that the Bethe–Peierls ansatz, and hence the resulting reconstruction, is exact for couplings forming a tree. However, the reconstruction of couplings and magnetic fields based on the Bethe–Peierls ansatz can also be applied to cases where the couplings do not form a tree. Although the results then arise from an uncontrolled approximation, the quality of the reconstruction can still be rather good. For a comparison of the different approaches, see Section 2.4.

2.2.11. *Belief propagation and susceptibility propagation*

Belief propagation is a distributed algorithm to compute marginal distributions of statistical models, like $p_i(s_i)$ and $p_{ij}(s_i, s_j)$ of the preceding sections. Again, it is exact on trees, but also gives a good approximation when the graph of couplings is only locally treelike (so any loops present are long). The term belief propagation is used in the machine learning [171] and statistical physics communities, in coding theory the approach is known as the sum-product algorithm [81]. Belief propagation shares a deep conceptual link with the Bethe–Peierls ansatz; Yedidia et al. [245] showed that belief propagation is a numerical scheme to efficiently compute the parameters of Bethe–Peierls ansatz.

A detailed exposition and many applications of belief propagation can be found in the textbook by Mézard and Montanari [144]. Here, we briefly introduce the basics and discuss applications to the inverse Ising problem. We start by considering the ferromagnetic Ising model in 1D, that is, a linear chain of N spins. The textbook solution of this problem considers the partitions function of the system when the last spin is constrained to take on the values $s_N = \pm 1$, $Z_N(+1)$, and $Z_N(-1)$. The corresponding partition functions for the linear chain with $N + 1$ spins are linked to the former via the so-called transfer matrix [20]. The partition function for a system of any size can be computed iteratively starting from a single spin and extending the 1D lattice with each multiplication of the transfer matrix. In fact, the transfer matrix can also be used to solve the Ising model on a tree [70]. Belief propagation is similar in spirit, but can be extended (as an approximation) also to graphs which are not trees. The best book-keeping device is again a restricted partition function, namely $Z_{i \to j}(s_i)$. It is defined as the partition function of the part of the system containing i when the coupling present between spins i and j has been deleted from the tree and spin i is constrained to take on s_i. (Deleting the edge (i,j) splits the tree containing i and j into two disconnected parts.) On a tree we obtain the recursion relation

$$Z_{i \to j}(s_i) = e^{h_i s_i} \prod_{k \in \partial i \setminus j} \left(\sum_{s_k} Z_{k \to i}(s_k)\, e^{J_{ik} s_i s_k} \right), \tag{74}$$

which can be computed recursively starting from leaves of the tree (nodes connected to a single edge only). In statistical physics, Equation (74) is called the cavity recursion for partition functions, since deleting a link can be thought of as leaving a 'cavity' in the original system. The

partition function for the entire tree with spin i constrained to s_i is then

$$Z_i(s_i) = e^{h_i s_i} \prod_{k \in \partial i} \left(\sum_{s_k} Z_{k \to i}(s_k) e^{J_{ik} s_i s_k} \right). \tag{75}$$

The marginal distribution $p_i(s_i)$ can be calculated by normalizing $p_i(s_i) \propto Z_i(s_i)$ and the marginal distribution $p_{ij}(s_i, s_j)$ can be calculated by normalizing $p_{ij}(s_i, s_j) \propto Z_{i \to j}(s_i) e^{J_{ij} s_i s_j} Z_{j \to i}(s_j)$.

The power of the cavity recursion (76) lies in its application to graphs which are not trees. It is particularly effective when the graph of couplings is at least locally treelike, so any loops present are long. To extend the cavity recursion as an approximation to graphs which are not trees, it makes sense to consider normalized quantities and define $\pi_{i \to j}(s_i) = Z_{i \to j}(s_i) / \sum_s Z_{i \to j}(s)$ with the recursion

$$\pi_{i \to j}(s_i) \propto e^{h_i s_i} \prod_{k \in \partial i} \left(\sum_{s_k} \pi_{k \to i}(s_k) e^{J_{ik} s_i s_k} \right) \tag{76}$$

leading to the estimate of the two-point marginals $\pi_{ij}(s_i, s_j) \propto e^{J_{ij} s_i s_j} \pi_{i \to j}(s_i) \pi_{j \to i}(s_j)$. (This step is necessary, as deleting the link (i, j) on a tree leads to two disjoint subtrees with separate partition functions and zero connected correlations. On a locally treelike graph, correlations between s_i and s_j can still be small once the link (i, j) has been cut, but the tree does not split into disjoint parts with separate partition functions.) Associating each edge (i, j) in the graph with particular values of $\pi_{i \to j}(s_i)$ and $\pi_{j \to i}(s_j)$, one can update these values in an iterative scheme which replaces them with the right-hand side of Equation (76) at each step. The fixed point of this procedure obeys Equation (76). In this context, $\pi_{i \to j}(s_i)$ is termed a 'message' that is being 'passed' between nodes. Belief propagation is an example of a message passing algorithm.

Belief propagation has been used to solve the inverse Ising problem in two different ways. In the first, belief propagation is used to approximately calculate the magnetizations and correlations given some parameters of the Ising model, and then fields and couplings are updated according to the Boltzmann learning rule (24). To estimate the correlations, one has to define additional messages also for the susceptibilities of each spin together with their update rules. This approach, termed susceptibility propagation, was developed by Welling and Teh [239,240] for the forward problem. Susceptibility propagation thus solves the 'forward' problem, that is, it offers a computationally efficient approximation to the correlations and magnetizations, which are then used for Boltzmann machine learning.

A sophisticated variant of susceptibility propagation was developed by Mézard and Mora [8,145] specifically for the inverse Ising problem, where also the couplings are updated at each step. This approach gives the same reconstruction as the Bethe–Peierls reconstruction from Section 2.2.10. However, the iterative equations can fail to converge, even when the analytical approach (71)–(73) gives valid approximate solutions [160].

2.2.12. *Independent pair approximation and the Cocco–Monasson adaptive-cluster expansion*

We expect the Bethe–Peierls ansatz to work well when the couplings are locally treelike and loops are long. Conversely, we expect the ansatz to break down when the couplings generate many short loops. Cocco and Monasson developed an iterative procedure to identify clusters of spins whose couplings form short loops and evaluate their contribution to the entropy (34) [53,54]. In the context of disordered systems, the expansion of the entropy in terms of clusters of connected spins is known as Kikuchi's cluster variational method [114,246].

The Cocco–Monasson adaptive-cluster expansion (ACE) directly approximates the entropy potential (34). We start by considering the statistics of a single spin variable described by its

magnetization m_i, with the entropy

$$S_{(1)}(m_i) = \sum_{s_i} \frac{1 + m_i s_i}{2} \ln \frac{1 + m_i s_i}{2}. \tag{77}$$

The simplest entropy involving coupled spins is the two-spin entropy

$$S_{(2)}(m_i, m_j, \chi_{ij}) = \sum_{s_i, s_j} \frac{1 + m_i s_i + m_j s_j + \chi_{ij} s_i s_j}{4}$$

$$\times \ln \frac{1 + m_i s_i + m_j s_j + \chi_{ij} s_i s_j}{4}. \tag{78}$$

When the two spins are statistically independent, we obtain $S_{(2)}(m_i, m_j, \chi_{ij}) = S_{(1)}(m_i) + S_{(2)}(m_j)$. Hence, the residual entropy which accounts for the correlation between the two spins is $\Delta S_2(m_i, m_j, \chi_{ij}) \equiv S_{(2)}(m_i, m_j, \chi_{ij}) - S_{(1)}(m_i) - S_{(1)}(m_j)$. To make the notation uniform, we also define $\Delta S_{(1)}(m_i) \equiv S_{(1)}(m_i)$. A very simple approximation is based on the assumption that the N-spin entropy (34) is described by pairwise terms

$$S(\boldsymbol{\chi}, \boldsymbol{m}) \approx \sum_i \Delta S_{(1)}(m_i) + \sum_{(i,j)} \Delta S_{(2)}(m_i, m_j, \chi_{ij}), \tag{79}$$

where the pair (i, j) denotes a cluster of two distinct spins. The couplings and fields can then be obtained via differentiation as in Equation (35),

$$J_{ij} = \sum_{s_i, s_j} \frac{s_i s_j}{4} \ln \frac{1 + m_i s_i + m_j s_j + \chi_{ij} s_i s_j}{4},$$

$$h_i = \frac{2 - N}{2} \ln \frac{1 + m_i}{1 - m_i}$$

$$+ \sum_{j \neq i} \sum_{s_i, s_j} \frac{s_i}{4} \ln \frac{1 + m_i s_i + m_j s_j + \chi_{ij} s_i s_j}{4}. \tag{80}$$

This result is called the independent pair approximation for the couplings, see [187,192]. (Expressions (8a) and (8b) in [187] differ slightly from Equation (80) due to a typo.) When the topology of couplings is known and forms a tree, Equation (79) gives the exact entropy of the system when the second sum is restricted to pairs of interacting spins (see Sections 2.2.9 and 2.2.10). In this case, Equation (80) gives the exact couplings.

However, in most cases the topology is not known, so second sum in Equation (79) runs over all pairs of spins. In this case, Equation (79) is only a (rather bad) approximation to the entropy. However, the independent pair approximation (79) can serve as starting point for an expansion that includes clusters of spins of increasing size

$$S(\boldsymbol{\chi}, \boldsymbol{m}) = \sum_i \Delta S_{(1)}(m_i)$$

$$+ \sum_{(i,j)} \Delta S_{(2)}(\chi_{ij}, m_i, m_j)$$

$$+ \sum_{(i,j,k)} \Delta S_{(3)}(\chi_{ij}, \chi_{jk}, \chi_{kl}, m_i, m_j, m_k) + \cdots. \tag{81}$$

In this expansion, the contribution from clusters consisting of three spins is

$$\begin{aligned}
\Delta S_{(3)} = {} & S_{(3)}(\chi_{ij}, \chi_{jk}, \chi_{kl}, m_i, m_j, m_k) \\
& - S_{(2)}(\chi_{ij}, m_i, m_j) - S_{(2)}(\chi_{jk}, m_j, m_k) \\
& - S_{(2)}(\chi_{ki}, m_k, m_i) + S_{(1)}(m_i) \\
& + S_{(1)}(m_j) + S_{(1)}(m_k),
\end{aligned} \tag{82}$$

where we have dropped the argument of $\Delta S_{(3)}$ to simplify the notation. The residual entropies of higher clusters are defined analogously.

The evaluation of $S_{(3)}$ is not as straightforward as $S_{(2)}$ and $S_{(1)}$, but still tractable. In general, the evaluation of $\Delta S_{(k)}$ requires the computation of order 2^k steps and therefore becomes quickly intractable. Cocco and Monasson argue that the contribution of ΔS_k decreases with k and can be neglected from a certain order on [53]. This inspires an adaptive procedure to build up the library of all clusters which contribute significantly to the total entropy of the system. Starting from 1-clusters, one constructs all 2-clusters by merging pairs of 1-clusters. If the entropy contribution of the new cluster is larger than some threshold, the new cluster is added to a list of clusters. One then constructs the 3-clusters, 4-clusters in the same way until no new cluster gives a contribution to the entropy exceeding the threshold. The total entropy and the reconstructed couplings and fields are updated each time a new cluster is added. This procedure fares well in situations when there are many short loops, like in a regular lattice in more than one dimension or when the graph of couplings is highly heterogeneous and some subsets of spins are strongly connected among each other.

The approaches introduced so far are each built on an ansatz for the Boltzmann distribution such as the mean-field ansatz (47) or the Bethe–Peierls ansatz (63): They are all uncontrolled approximations. In the next sections, we discuss controlled approximations based on either an expansion in small couplings or small correlations.

2.2.13. *The Plefka expansion*

In 1982, Plefka gave a systematic expansion of the Gibbs free energy of the Sherrington–Kirkpatrick (SK) model [202] in the couplings between spins [179]. Already 8 years earlier, Bogolyubov Jr. et al. had used similar ideas in the context of the ferromagnetic Ising model on a lattice [33]. The resulting estimate of the partition function can also be used to derive new solutions of the inverse Ising problem. Zeroth- and first-order terms of Plefka's expansion turn out to yield mean-field theory (48), the second-order term gives the TAP free energy (54), and correspondingly the reconstructions of couplings (49) and (57).

Plefka's expansion is a Legendre transformation of the cumulant expansion of the Helmholtz free energy $F(\boldsymbol{J}, \boldsymbol{h})$. Plefka introduced a perturbation parameter λ to the interacting part of the Hamiltonian

$$H(\boldsymbol{s}) = H_0(\boldsymbol{s}) + \lambda V(\boldsymbol{s}), \tag{83}$$

where $H_0(\boldsymbol{s}) = -\sum_i h_i s_i$ and $V(\boldsymbol{s}) = -\sum_{i<j} J_{ij} s_i s_j$. The perturbation parameter λ serves to distinguish the different orders in the strength of couplings and will be set to one at the end of the calculation. The standard cumulant expansion of the Helmholtz free energy $F^\lambda(\boldsymbol{J}, \boldsymbol{h})$ then reads

$$F^\lambda(\boldsymbol{J}, \boldsymbol{h}) = F^{(0)}(\boldsymbol{h}) - \lambda \langle V \rangle_0 + \frac{\lambda^2}{2}[\langle V^2 \rangle_0 - \langle V \rangle_0^2] + \cdots, \tag{84}$$

where $F^{(0)}(\boldsymbol{h}) = \sum_i 2\cosh h_i$ and $\langle \rangle_0$ denotes the average with respect to the Boltzmann distribution corresponding to the non-interacting part H_0 of the Hamiltonian. Next, we perform the

Legendre transformation of $F^{\lambda}(\boldsymbol{J}, \boldsymbol{h})$ with respect to $\boldsymbol{h}$ to obtain the perturbative series for the Gibbs free energy

$$G^{\lambda} = G^{(0)} + \lambda G^{(1)} + \lambda^2 G^{(2)} + \cdots, \tag{85}$$

where we suppressed the dependence of G on $\boldsymbol{J}$ and $\boldsymbol{m}$ to simplify the notation. Using the definition of the Gibbs free energy (38), we solve

$$G^{\lambda}(\boldsymbol{J}, \boldsymbol{m}) = \sum_i h_i m_i + F^{\lambda}(\boldsymbol{J}, \boldsymbol{h}) \tag{86}$$

with the local fields satisfying

$$m_i = -\frac{\partial F^{\lambda}(\boldsymbol{J}, \boldsymbol{h})}{\partial h_i}. \tag{87}$$

Plugging the perturbative series

$$\boldsymbol{h}^{\lambda} = \boldsymbol{h}^{(0)} + \lambda \boldsymbol{h}^{(1)} + \lambda^2 \boldsymbol{h}^{(2)} + \cdots \tag{88}$$

into Equation (87) and re-expanding the right-hand side in λ, one finds successive orders of $\boldsymbol{h}$. In particular, the lowest two orders are found easily as

$$\begin{aligned} h^{(0)} &= \operatorname{artanh} m_i, \\ h^{(1)} &= -\sum_j J_{ij} m_j. \end{aligned} \tag{89}$$

This gives the Gibbs free energy up to first order in λ,

$$\begin{aligned} G^{(0)}(\boldsymbol{m}) &= \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\}, \\ G^{(1)}(\boldsymbol{m}) &= -\sum_{i<j} J_{ij} m_i m_j. \end{aligned} \tag{90}$$

Continuing to the second order, one obtains the Onsager term (54)

$$G^{(2)}(\boldsymbol{m}) = -\frac{1}{2} \sum_{i<j} J_{ij}^2 (1 - m_i^2)(1 - m_j^2). \tag{91}$$

A systematic way to perform this expansion has been developed by Georges and Yedidia [85] leading to

$$\begin{aligned} G^{(3)}(\boldsymbol{m}) = {} & -\frac{2}{3} \sum_{(i,j)} J_{ij}^3 m_i (1 - m_i^2) m_j (1 - m_j^2) \\ & - \sum_{(i,j,k)} J_{ij} J_{jk} J_{ki} (1 - m_i^2)(1 - m_j^2)(1 - m_k^2), \end{aligned}$$

$$G^{(4)}(\boldsymbol{m}) = \frac{1}{12} \sum_{(i,j)} J_{ij}^4 (1 - m_i^2)(1 - m_j^2)$$

$$\times (1 + 3m_i^2 + 3m_j^2 - 15m_i^2 m_j^2)$$

$$- \sum_{(i,j,k,l)} J_{ij} J_{jk} J_{kl} J_{li} (1 - m_i^2)(1 - m_j^2)(1 - m_k^2)(1 - m_l^2)$$

$$- 2 \sum_{(i,j,k)} J_{ij}^2 J_{jk} J_{ki} m_i (1 - m_i^2) m_j (1 - m_j^2)(1 - m_k^2). \tag{92}$$

The mean-field and TAP reconstructions of couplings (53) and (57) then follow from Equation (40) and the first derivatives of G with respect to the magnetizations gives the corresponding reconstructions of the magnetic fields (50) and (58). Higher order terms of the Plefka expansion are discussed by Georges and Yedidia [85,244], but have not yet been applied to the inverse Ising problem. For a fully connected ferromagnetic model with all couplings set to J/N (to ensure an extensive Hamiltonian), one finds that already the second-order term $G^{(2)}$ vanishes relative to the first in the thermodynamic limit. For the SK model of a spin glass [202], couplings are of the order of $N^{-1/2}$, again to make the energy extensive. In that case, $G^{(0)}$, $G^{(1)}$, and $G^{(2)}$ turn out to be extensive, but the third order vanishes in the thermodynamic limit. In specific instances (not independently and identically distributed couplings), higher order terms of the Plefka expansion may turn out to be important.

The terms in J_{ij}, J_{ij}^2, J_{ij}^3, etc. appearing in Equations (90)–(92) resum to the results of the Bethe–Peierls ansatz [85]. If the couplings form a tree, these terms are the only ones contributing to the Gibbs free energy; terms like $J_{ij} J_{jk} J_{ki}$ in Equation (92) quantify the coupling strengths around a closed loop of spins and are zero if couplings form a tree. This finally shows that the Bethe–Peierls ansatz is exact on a tree. However, when the couplings contain loops, these terms contribute to the Gibbs free energy. Beyond evaluating the Plefka expansion at different orders, one can also resum the contributions from specific types of loops to infinite order. We will discuss such an approach in the next section.

2.2.14. *The Sessak–Monasson small-correlation expansion*

The Sessak–Monasson expansion is a perturbative expansion of the entropy $S(\boldsymbol{\chi}, \boldsymbol{m})$ in terms of the *connected correlation* $C_{ij} \equiv \chi_{ij} - m_i m_j$ [200]. For zero correlations C_{ij}, the couplings J_{ij} should also be zero. This motivates an expansion of the free energy $F(\boldsymbol{J}, \boldsymbol{h})$ in terms of the connected correlations around a non-interacting system; the Legendre transformation (34) then yields the perturbative series of the entropy function $S(\boldsymbol{\chi}, \boldsymbol{m})$. Equivalently, this is an expansion of Equations (23) for the fields and the couplings in the connected correlation $\boldsymbol{C}$. To make the perturbation explicit, we replace $\boldsymbol{C}$ by $\lambda \boldsymbol{C}$, where the perturbation parameter λ is to be set to one at the end. The couplings and fields are the solutions of

$$m_i = -\frac{\partial F(\boldsymbol{J}, \boldsymbol{h})}{\partial h_i},$$

$$\lambda C_{ij} + m_i m_j = -\frac{\partial F(\boldsymbol{J}, \boldsymbol{h})}{\partial J_{ij}}, \tag{93}$$

where the latter can be replaced by

$$\lambda C_{ij} = -\frac{\partial^2 F(\boldsymbol{J}, \boldsymbol{h})}{\partial h_i \partial h_j}. \tag{94}$$

We then expand the solution $\boldsymbol{h}$ and $\boldsymbol{J}$ in a power series around the uncorrelated case $\lambda = 0$

$$\boldsymbol{h}^\lambda = \boldsymbol{h}^{(0)} + \lambda \boldsymbol{h}^{(1)} + \frac{\lambda^2}{2}\boldsymbol{h}^{(2)} + \cdots ,$$
$$\boldsymbol{J}^\lambda = \boldsymbol{J}^{(0)} + \lambda \boldsymbol{J}^{(1)} + \frac{\lambda^2}{2}\boldsymbol{J}^{(2)} + \cdots . \tag{95}$$

At the zeroth order in the connected correlations, spins are uncorrelated, so $\boldsymbol{J}^{(0)} = 0$ and $h^{(0)} = \mathrm{artanh}(m_i)$. The expansion of model parameters induces an expansion of the Hamiltonian

$$H^\lambda = H^{(0)} + \lambda H^{(1)} + \frac{\lambda^2}{2}H^{(2)} + \cdots , \tag{96}$$

where

$$H^{(k)}(\boldsymbol{s}) = -\sum_{i<j} J_{ij}^{(k)} s_i s_j - \sum_i h_i^{(k)} s_i. \tag{97}$$

Likewise, the free energy $F(\boldsymbol{J},\boldsymbol{h}) = -\ln Z(\boldsymbol{J},\boldsymbol{h})$ (or the cumulant generator) can be expanded in λ,

$$F^\lambda = F^{(0)} + \lambda F^{(1)} + \lambda^2 F^{(2)} + \cdots$$
$$= F^{(0)} - \lambda \langle H^{(1)}\rangle_0 + \frac{\lambda^2}{2}[-\langle H^{(2)}\rangle_0 + \langle (H^{(1)})^2\rangle_0 - \langle H^{(1)}\rangle_0^2] + \cdots , \tag{98}$$

where the subscript $\langle\rangle_0$ refers to the thermal average under the zeroth-order Hamiltonian $H^{(0)}(\boldsymbol{s})$. For example, the first-order term is

$$F^{(1)} = -\langle H^{(1)}\rangle_0$$
$$= \sum_{i<j} J_{ij}^{(1)} \,\mathrm{th}(h_i^0)\,\mathrm{th}(h_j^0) + \sum_i h_i^{(1)}\,\mathrm{th}(h_i^0). \tag{99}$$

Equations (93) and (94) then read

$$m_i = -\frac{\partial F^\lambda}{\partial h_i^{(0)}} = -\frac{\partial F^{(0)}}{\partial h_i^{(0)}} - \lambda\frac{\partial F^{(1)}}{\partial h_i^{(0)}} - \lambda^2\frac{\partial F^{(2)}}{\partial h_i^{(0)}} - \cdots ,$$
$$\lambda C_{ij} = -\frac{\partial^2 F^\lambda}{\partial h_i^{(0)}\partial h_j^{(0)}} = -\frac{\partial^2 F^{(0)}}{\partial h_i^{(0)}\partial h_j^{(0)}} - \lambda\frac{\partial^2 F^{(1)}}{\partial h_i^{(0)}\partial h_j^{(0)}} - \cdots . \tag{100}$$

Evaluating these equations successively gives solutions for the different orders of $h_i^{(k)}$ and $J_{ij}^{(k)}$, which in turn yield expressions for reconstructed fields and couplings when the magnetizations $\boldsymbol{m}$ and the connected correlations $\boldsymbol{C}$ are identified with their empirical values. Recalling the zeroth-order solution $h_i^{(0)} = \mathrm{artanh}(m_i)$, the first-order contribution of the free energy (99) leads to

$$J_{ij}^{(1)} = \frac{C_{ij}}{[1 - (m_i)^2][1 - (m_j)^2]},$$
$$h_i^{(1)} = -\sum_{j\neq i} J_{ij}^{(1)} m_j. \tag{101}$$

Higher order terms are given in [200]. Also in [200], Sessak and Monasson give a diagrammatic framework suitable for the derivation of higher order terms in the couplings and sum specific

terms to infinite order. Roudi et al. [192] simplified the results yielding the reconstruction

$$J_{ij} = -(\boldsymbol{C}^{-1})_{ij} + J_{ij}^{\text{IP}} - \frac{C_{ij}}{(1 - m_i^2)(1 - m_j^2) - (C_{ij})^2} \tag{102}$$

for couplings with $i \neq j$, where J_{ij}^{IP} is the independent pair approximation (80). This result can be interpreted as follows [192]: The first term is the mean-field reconstruction, which is of a sub-series of the Sessak–Monasson expansion. One then adds the sub-series that constitutes the independent pair approximation (80). The last term is the overlap of the two series: the mean-field reconstruction of two spins considered independently, which needs to be subtracted. The resummation of other sub-series for special cases is also possible [200]; in [105] loop diagrams are resummed to obtain a reconstruction that is particularly robust against sampling noise.

2.3. *Logistic regression and pseudolikelihood*

Pseudolikelihood is an alternative to the likelihood function (16) and leads to the exact inference of model parameters in the limit of an infinite number of samples [4,102,156]. The computational complexity of this approach scales polynomially with the number of spin variables N, but also with the number of samples M. In practice, this is usually much faster than exact maximization of the likelihood function, whose computational complexity is exponential in the system size. Pseudolikelihood inference was developed by Julian Besag in 1974 in the context of statistical inference of data with spatial dependencies [25]. It is closely related to logistic regression. While pseudolikelihood and regression have been used widely in statistical inference [26,110,182], this approach was not well known in the physics community until quite recently [7,62,71,156].

Our derivation focuses on the physical character of regression analysis, which we then link to Besag's pseudolikelihood. Key observation is that although the likelihood function (16) depends on the model parameters in a complicated way, one can simplify this dependence by separating different groups of parameters. Let us consider how the statistics of a particular spin variable σ_i depends on the configuration of all other spins. We split the Hamiltonian into two parts

$$\begin{aligned} H(\boldsymbol{s}) &= H_i(\boldsymbol{s}) + H_{\backslash i}(\boldsymbol{s} \setminus s_i) \\ &= -h_i s_i - s_i \sum_{j \neq i} J_{ij} s_j + H_{\backslash i}(\boldsymbol{s} \setminus s_i), \end{aligned} \tag{103}$$

such that the first part $H_i(\boldsymbol{s})$ depends on the magnetic field h_i and the couplings of spin i to other spins J_{ij}, while the part $H_{\backslash i}(\boldsymbol{s} \setminus s_i)$ does not. $\boldsymbol{s} \setminus s_i$ denotes all spin variables except spin s_i. This splitting of variables is possible since the statistics of σ_i conditioned on the remaining spins $\{s_j\}_{j \neq i}$ is fully captured by h_i and $J_{ij}, (j \in 1, \ldots, N)$.

The expectation values of σ_i can be computed based on the partition function

$$Z(\boldsymbol{J}, \boldsymbol{h}) = \sum_{\boldsymbol{s} \setminus s_i} 2\cosh\left(h_i + \sum_{j \neq i} J_{ij} s_j\right) e^{-H_{\backslash i}(\boldsymbol{s} \setminus s_i)}, \tag{104}$$

where only spin i has been summed over. Differentiating the partition function in this form with respect to h_i and J_{ij} yields the expectation values

$$\langle \sigma_i \rangle = \left\langle \text{th}\left(h_i + \sum_{k \neq i} J_{ik} \sigma_k\right)\right\rangle \tag{105}$$

$$\langle \sigma_i \sigma_j \rangle = \left\langle \sigma_j \, \text{th} \left(h_i + \sum_{k \neq i} J_{ik} \sigma_k \right) \right\rangle, \tag{106}$$

where on both sides the thermal average is over the Boltzmann measure $e^{-H(s)}/Z$, but on the right-hand side only the spins $s \setminus s_i$ are averaged over.

In statistical physics, these equations are known as Callen's identities [44] and have been used to compute coupling parameters from observables in Monte Carlo simulations for the numerical calculation of renormalization-group trajectories [213]. While these equations are exact, the expectation values on the right-hand sides still contain an average over the $N-1$ spins other than i.

The crucial step and the only approximation involved is to replace the remaining averages in Equation (106) with the sample means

$$\langle \sigma_i \rangle^D = \left\langle \text{th} \left(h_i^{\text{PL}} + \sum_{k \neq i} J_{ik}^{\text{PL}} \sigma_k \right) \right\rangle^D,$$

$$\langle \sigma_i \sigma_j \rangle^D = \left\langle \sigma_j \, \text{th} \left(h_i^{\text{PL}} + \sum_{k \neq i} J_{ik}^{\text{PL}} \sigma_k \right) \right\rangle^D. \tag{107}$$

We now have a system of non-linear equations to be solved for h_i and J_{ij} for fixed i and various $j, j \neq i$. Standard methods to solve such equations are Newton–Raphson or conjugated gradient descent [93,180].

With these steps, we have broken down the problem of estimating the magnetic fields and the full coupling matrix to N separate problems of estimating one magnetic field and a single row of the coupling matrix for a specific spin i. Crucially, the Boltzmann average over 2^{N-1} states has been replaced with an average over all configurations of the samples. As a result, the computation of Equation (107) uses not only the sample magnetizations and correlations (sufficient statistics), but all spin configurations that have been sampled. The time required to evaluate Equation (107) is thus linear in the number of samples M. In general, the coupling matrix inferred in this way will be asymmetric, $J_{ij} \neq J_{ji}$ due to sampling noise. A practical solution is to use the average $\frac{1}{2}(J_{ij} + J_{ji})$ as estimate of the (symmetric) coupling matrix.

The set of Equations (107) can be viewed as solving a gradient-descent problem of a logistic regression [182]. The statistics of σ_i conditioned on the values of the remaining spins $\{s_j\}_{j \neq i}$ can be written as

$$p(s_i | \{s_j\}_{j \neq i}) = \frac{1}{2} \left[1 + s_i \, \text{th} \left(h_i + \sum_{j \neq i} J_{ij} s_j \right) \right]$$

$$= \frac{1}{1 + e^{-2s_i(h_i + \sum_{j \neq i} J_{ij} s_j)}}. \tag{108}$$

From this expression for the conditional probability, the ith row of couplings J_{i*} and the field h_i are simply the coefficients and the intercept in the multi-variate logistic regression of the response variable σ_i on the variables $\{s_j\}_{j \neq i}$. The log-likelihood for the ith row of couplings J_{i*} and the

magnetic field h_i of this regression problem is

$$L_{\mathrm{D}}^i(J_{i*}, h_i) = \frac{1}{M} \sum_{\mu} \ln p(s_i^{\mu} | \{s_j^{\mu}\}_{j \neq i})$$

$$= \frac{1}{M} \sum_{\mu} \ln \frac{1}{2} \left[1 + s_i^{\mu} \, \mathrm{th} \left(h_i + \sum_{j \neq i} J_{ij} s_j^{\mu} \right) \right]. \tag{109}$$

Setting the derivatives of this likelihood function with respect to the magnetic field h_i and the entries of J_{i*} to zero recovers (107).

Consider all couplings and fields together, one can sum (109) over all rows of couplings to obtain the so-called (log)-*pseudolikelihood* [26]

$$L^{\mathrm{PL}}(\boldsymbol{J}, \boldsymbol{h}) = \sum_i L_{\mathrm{D}}^i(J_{i*}, h_i), \tag{110}$$

which can be maximized with respect to all rows of the coupling matrix, yielding an asymmetric matrix as discussed above. The pseudolikelihood can also be maximized within the space of symmetric matrices, although the maximization problem is harder [7]; rather than solving N independent gradient-descent problems in N variables we have a single problem in $N(N+1)/2$ variables. In practice, maximising the pseudolikelihood without the symmetry constraint on the coupling matrix is preferred because of its simplicity and efficiency.

The reconstruction based on maximizing the pseudolikelihood is of a different nature than the previous methods, which approximated the Boltzmann measure. The Boltzmann measure specifies the probability of observing a particular spin configuration $\boldsymbol{s}$ in equilibrium. On the other hand, the conditional probability $\ln p(s_i^{\mu} | \{s_j^{\mu}\}_{j \neq i})$ only gives the probability of observing a given spin s_i conditioned on the remaining spin variables and cannot be used to generate the configurations of all spins. Yet, as a function of the model parameters, the log-pseudolikelihood (109) has the same maximum as the likelihood (16) in the limit of $M \to \infty$, when the sample average in Equation (107) coincides with the corresponding expectation values under the Boltzmann distribution.

Curiously, the pseudolikelihood can also be used to obtain a variant of the mean-field and TAP reconstructions. We follow Roudi and Hertz [189] and replace expressions such as $\langle \mathrm{th}(h_i + \sum_{j \neq i} J_{ij} \sigma_j) \rangle$ by $\mathrm{th}(h_i + \sum_{j \neq i} J_{ij} \langle \sigma_j \rangle)$, thus replacing the effective local field by its mean. The resulting approximation of Callen's identity (105) is

$$m_i^{\mathrm{MF}} = \mathrm{th} \left(h_i + \sum_{j \neq i} J_{ij} m_j^{\mathrm{MF}} \right), \tag{111}$$

which is the mean-field equation (51) for the magnetizations. Then, replacing σ_i by $\langle \sigma_i \rangle + (\sigma_i - \langle \sigma_i \rangle)$ in Equation (106) and expanding the equation to the lowest orders in $(\sigma_i - \langle \sigma_i \rangle)$ gives

$$C_{ij}^{\mathrm{PL-MF}} = \frac{1}{1 - (m_i^{\mathrm{MF}})^2} \sum_{k \neq i} J_{ik} C_{kj}^{\mathrm{PL-MF}}. \tag{112}$$

Identifying the pseudolikelihood mean-field magnetizations m_i^{MF} and connected correlations $C_{ij}^{\mathrm{PL-MF}}$ with the sample magnetizations m_i and sampled correlations C_{ij}, this linear equation

can be solved for $J_{ik}^{\text{PL}-\text{MF}}$ for a fixed i, giving

$$J_{ik}^{\text{PL}-\text{MF}} = [1 - m_i^2] \sum_{j \neq i} C_{ij} \times [(\boldsymbol{C}_{\backslash i})^{-1}]_{jk}, \tag{113}$$

where $\boldsymbol{C}_{\backslash i}$ is the submatrix of the correlation matrix with row and column i removed. This reconstruction is closely related, but not identical to the mean-field reconstruction (53). In particular, the diagonal terms are naturally excluded. Numerical experiments show that this reconstruction gives a good correlation between the reconstructed and true couplings, but the magnitude of the couplings is systematically underestimated. Continuing the expansion to second order in $(\sigma_i - \langle \sigma_i \rangle)$ leads to a variant of the TAP reconstruction (57).

2.4. *Comparison of the different approaches*

Table 1 gives a classification of the different reconstruction methods based on the thermodynamic potentials they employ and the approximations used to evaluate them. Some of these approximations are exact in particular limits, and fail in others. This has consequences for how well a reconstruction method works in a given regime. For instance, the TAP equations become exact for fully connected systems (with couplings between all spin pairs drawn from a distribution with variance $1/N$) in the thermodynamic limit. Hence, we expect the TAP reconstruction to perform well when couplings are uniformly distributed across spin pairs, and couplings are sufficiently weak, so there is no ergodicity breaking (see Section 2.5). Similarly, the Bethe–Peierls approximation is exact when the non-zero couplings between spin pairs form a tree. Hence, we expect the Bethe–Peierls reconstruction to work perfectly in this case, and to work well when the graph of couplings is locally treelike (so there are no short loops). The ACE, on the other hand, is expected to work well even when there are short loops.

In this section, we compare the reconstruction methods discussed so far. We consider the reconstruction problem of an Ising model (2) with couplings J_{ij}^0 between pairs of spins defined on a certain graph. We explore two aspects of the model parameters: the type of the interaction graph and the coupling strength (temperature). We consider three different graphs: the fully connected graph, a random graph with fixed degree as a representative of graphs with long loops, and the 2D square lattice as a representative of graphs with short loops. In this section, we denote the couplings of the model underlying the data by the superscript '0' to distinguish them from the inferred couplings. For each graph, every edge is assigned a coupling J_{ij}^0 drawn from a certain distribution. We use the Gaussian distribution with zero mean and standard deviation $\beta/\sqrt{N}$ for couplings on the fully connected graph, leading to the SK model. For the tree and square

Table 1. Classification of reconstruction methods based on the thermodynamic potentials used and the approximations employed to evaluate them.

Associated potential	Exact	Variational approximation	Perturbative expansion
$F(\boldsymbol{J}, \boldsymbol{h})$	Convex non-linear optimization		
$G(\boldsymbol{J}, \boldsymbol{m})$		Mean-field, Bethe–Peierls	Plefka, mean-field, TAP, Bethe–Peierls
$S(\boldsymbol{\chi}, \boldsymbol{m})$		Independent pair approximation, Cocco–Monasson	Sessak–Monasson

Note: Pseudolikelihood maximization falls outside this classification scheme, as it is not based on an approximation of the Boltzmann measure.

lattice, the uniform distribution on the interval $[-\beta, \beta]$ is used for the couplings. By tuning β, we effectively change the coupling strength. For the SK model, external magnetic fields field h_i^0 are drawn uniformly from the interval $[-0.3\beta, 0.3\beta]$.

At each value of β, we generate M samples (spin configurations) drawn from the Boltzmann distribution. Each sample is obtained by simulating $10^4 N$ Monte Carlo steps using the Metropolis transition rule starting from a random configuration. Although this is not always sufficient to guarantee that the system has reached equilibrium, the results are not sensitive to increasing this breaking-in time.

Figure 4(top left) compares the reconstructed couplings with the couplings of the original model at $\beta = 1.3$. We chose this value as it results in couplings which are sufficiently large so that the different methods perform differently (we will see that at low β, all methods perform similarly). On the fully connected graph, the ACE performs poorly as it sets too many couplings to zero. Mean-field reconstruction somewhat overestimates large couplings [16]. This error also affects the estimate of the magnetic fields. The TAP and Bethe–Peierls reconstructions correct this overestimate quite effectively. The reconstruction by pseudolikelihood stands out by providing an accurate reconstruction of the model parameters. One consequence is that in Figure 4, the symbols indicating the results from pseudolikelihood are largely obscured by those from the exact maximization of the likelihood (16). We also compare the correlations and magnetizations based on the reconstructed parameters with those of the original model. The correlations and
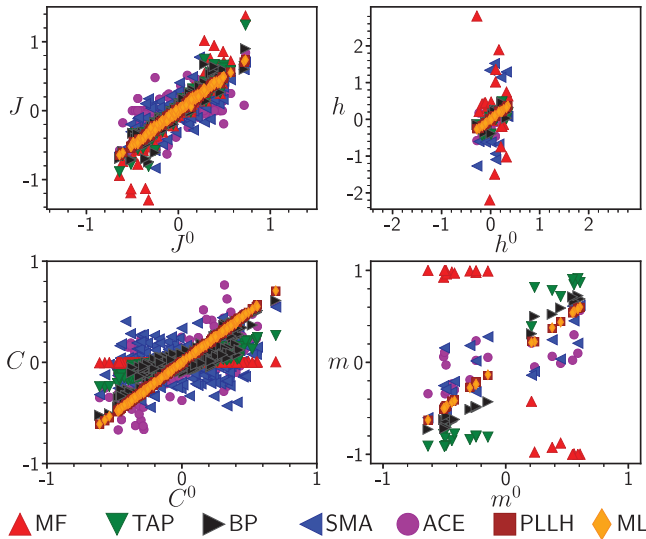

Figure 4. Reconstructing a fully connected Ising model with different methods. The scatter plots are generated from a particular realization of the SK model with $N = 20$, $\beta = 1.3$ and $M = 15,000$ samples (configurations drawn from the Boltzmann measure, see text). The colour legend indicates the different reconstruction methods: mean-field (MF), TAP, Bethe–Peierls (BP), Sessak–Monasson (SM), the ACE, maximum pseudolikelihood (MPL), and ML. The top plots show the reconstructed couplings/fields against the couplings/fields of the original model. The bottom plots show the connected correlations and magnetizations calculated from M samples generated using the underlying model parameters (on the x-axis) and the same quantities generated using the inferred parameters (y-axis). The Sessak–Monasson expansion was computed as described in [200] with the series in the magnetic fields truncated at the third order excluding the loop terms. The ACE was carried out using the ACE-package [17,53] with a maximum cluster size $k = 6$ and default parameters. The numerical maximization of the likelihood and pseudolikelihood was done using the Eigen 3 wrappers [89] for Minpack [155].

magnetizations are sampled as described above. We find significant bias in the results from all methods except for likelihood maximization and pseudolikelihood maximization (bottom panels of Figure 4).

Next, we explore how the reconstruction quality depends on the coupling strength β and the number of samples M. To this end, we define the relative reconstruction error

$$\gamma_J = \sqrt{\frac{\sum_{i<j}(J_{ij} - J_{ij}^0)^2}{\sum_{i<j}(J_{ij}^0)^2}}, \tag{114}$$

which compares the reconstructed couplings $\boldsymbol{J}$ to the couplings of the original model underlying the data $\boldsymbol{J}^0$. Figure 5, large panel, shows the relative reconstruction error γ_J as a function of coupling strength β. At low β, the connected correlations are small, so all reconstruction methods are equally limited by sampling noise: The relative reconstruction error increases with decreasing β as the couplings become small relative to the sampling error. This is also compatible with the result that at weak couplings the relative errors of all methods decrease with an increasing number of samples, as seen in the six small panels of Figure 5.

We find that the pseudolikelihood reconstruction outperforms the other (non-exact) methods over the entire range of β, and correctly reconstructs the model parameters even in the glassy phase at strong couplings [7]. This is because the conditional statistics of a single spin (conditioned on the other spins) is correctly described by Equation (108) at any coupling strength. Although at strong couplings, the error of the pseudolikelihood reconstruction grows with β, this can be compensated by increasing the number of samples (Figure 5, bottom right). This is not so for all other approximate methods (Figure 5, small panels): At high β, the approximation each method is based on breaks down, which cannot be compensated by a larger number of samples.

Figure 6 shows the reconstruction errors of different methods for a random graph of fixed degree (column A) and the square lattice (column B). Again, pseudolikelihood performs well in these cases, as does the Bethe–Peierls reconstruction. The ACE shows a remarkable behaviour: For both graphs, it has a very small reconstruction error at weak coupling strength, but breaks down at strong couplings. At weak couplings, where all other methods have similar reconstruction errors arising from sampling noise, however the ACE appears to avoid this source of error. The ACE explicitly assumes some couplings are exactly zero, so the reconstruction is biased towards graphs which are not fully connected. It thus uses *extra information* beyond the data.

Our comparison of the different methods is based on the knowledge of the underlying couplings, so the reconstructed couplings can be compared to the true underlying ones. In practice, the underlying couplings are not available. However, the likelihood (16) can be evaluated for different reconstructions, with the better reconstruction resulting in a higher value of the likelihood. Alternatively, the correlations and magnetizations of the reconstructed model can be compared with those observed in the data as in Figure 4. Whether regularizing terms improve the reconstruction can be decided based on statistical tests such as the Bayesian information criterion [198] or the Akaike information criterion [2].

An aspect not discussed so far is the inference of the interaction graph, this is, the distinction between non-zero and zero couplings. In practice, there is always some ambiguity between small non-zero couplings and couplings which are exactly zero, particularly at high sampling noise. A standard practice is to set a threshold and cut off small couplings below the threshold. Two exceptions are the ACE and the pseudolikelihood reconstruction. The ACE has a built-in procedure to set some couplings to zero. For the pseudolikelihood reconstruction, the problem is related to feature selection. Many methods for feature selection are available from statistics [93]; however, there is no consensus on the best method for the inverse Ising problem. One
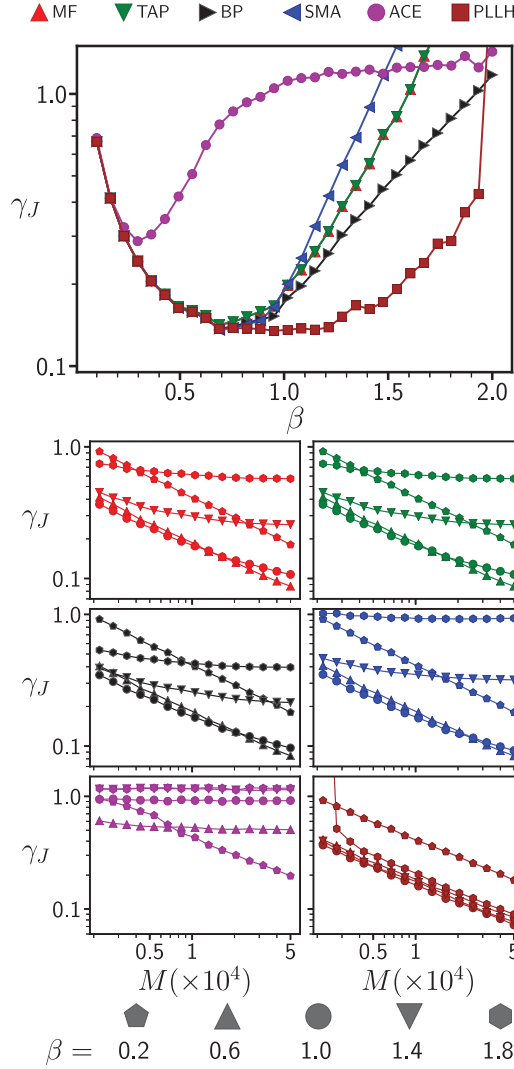
Figure 5. Reconstruction of a fully connected Ising model as a function of the coupling strength β and the number of samples M. The top panel shows the reconstruction error γ_J defined by Equation (114) as a function of β at a constant $M = 15,000$ samples. The smaller panels show γ_J as a function of the number of samples M at different coupling strengths β and for different methods (mean-field, TAP, Bethe–Peierls, Sessak–Monasson, ACE, and maximum pseudolikelihood). For these plots, a larger system size $N = 64$ was used, so evaluating the likelihood is not feasible; at lower system sizes we found pseudolikelihood performs as well as the exact maximization of the likelihood in the parameter range considered.

possibility is adding an ℓ_1-regularization term (known as Lasso, see Section 2.2) to the pseudo-likelihood [182]. However, there is a critical coupling strength, below which the ℓ_1 regularization of the pseudolikelihood fails to recover the interaction graph [149].

2.5. *Reconstruction and non-ergodicity*

At strong couplings, the dynamics of a system can become non-ergodic, which affects the sampling of configurations, and hence the reconstruction of parameters. Reconstruction on the
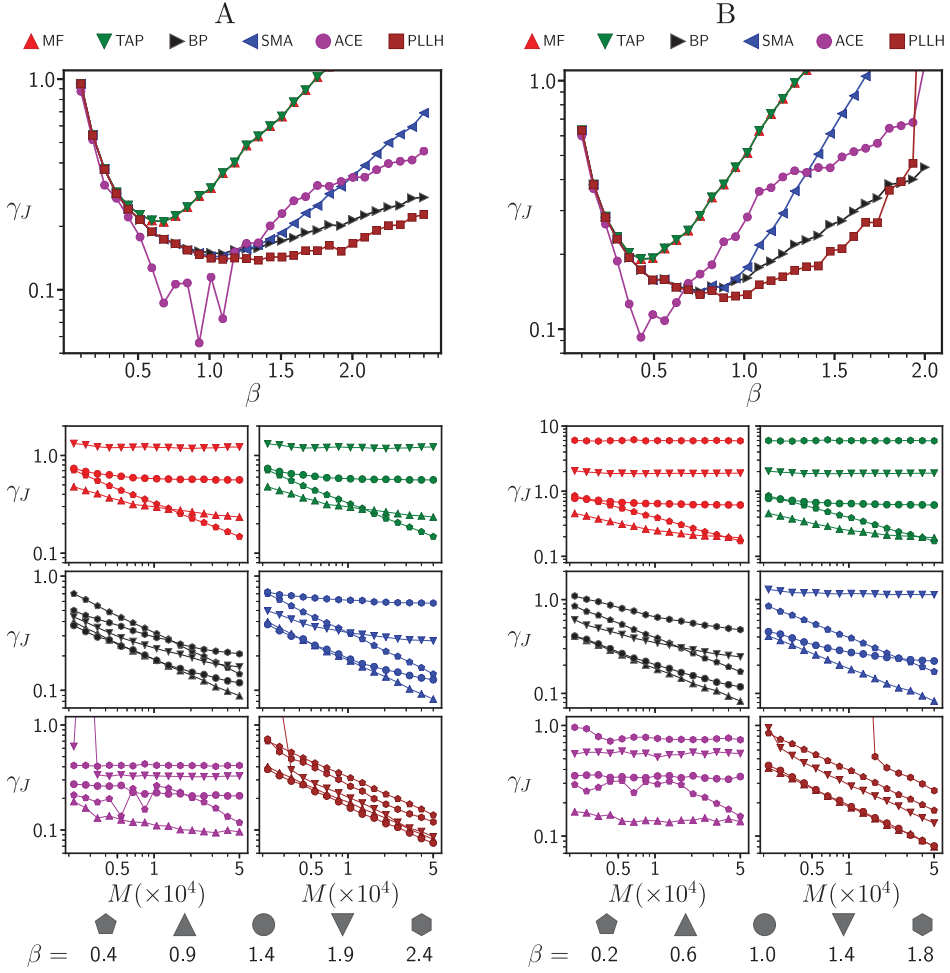
Figure 6. Reconstruction of the Ising model on a random graph of fixed degree $z = 3$ (column A) and on the square lattice (column B) as a function of coupling strength β and number of samples M. The underlying couplings J_{ij}^0 are drawn from the uniform distribution on the interval $[-\beta, +\beta]$, and the external magnetic fields are set to zero for simplicity. The system size is $N = 64$. The top panel shows the reconstruction error γ_J defined by Equation (114) as a function of β at constant $M = 15,000$ samples. The smaller panels show γ_J as a function of the number of samples M at different coupling strength β and for different methods (mean-field, TAP, Bethe–Peierls, Sessak–Monasson, ACE, and maximum pseudolikelihood).

basis of non-ergodic sampling is a fundamentally difficult subject where little is known to date.

At low temperatures (or strong couplings), a disordered spin system can undergo a phase transition to a state where spins are 'frozen' in different directions. This is the famous transition to the spin glass phase [146]. At the spin glass transition, the Gibbs free energy develops multiple valleys with extensive barriers between them (thermodynamic states). These free energy barriers constrain the dynamics of the system, and the system can remain confined to one particular valley for long times. A signature of this ergodicity breaking is the emergence of non-trivial spin magnetizations: At the phase transition, the self-consistent equations (51) and (55) develop

multiple solutions for the magnetizations m_i, corresponding to the different orientations the spins freeze into. The appearance of multiple free energy minima and the resulting ergodicity breaking has long been recognized as an obstacle to mean-field reconstruction [137]. In fact, all methods based on self-consistent equations (mean-field, TAP, Bethe–Peierls reconstruction) fail in the glassy phase [145].

The mean-field equations (51) describe individual thermodynamic states, not the mixture of many states that characterizes the Boltzmann measure. At large couplings (or low temperatures), mean-field and related approaches turn out not to be limited by the validity of self-consistent equations like (51) or (55), but by the identification of observed magnetizations and correlations with the corresponding quantities calculated under one of the self-consistent equations. The former may involve averages over multiple thermodynamic states, whereas each solution of the self-consistent equations describes a separate thermodynamic state. A solution to this problem is thus to consider correlations and magnetizations *within* a single thermodynamic state, where the mean-field result (52) is valid within the limitations of mean-field theory. The different thermodynamic states (free energy minima) can be identified from the data by searching for clusters in the sampled spin configurations. Collecting self-consistent equations from different minima and jointly solving these equations using the Moore–Penrose pseudo-inverse allows the reconstruction of couplings and fields [161].

The emergence of multiple states can lead to another problem: the samples need not come from the Gibbs measure in the first place, but might be taken only from a single thermodynamic state. In [35], the reconstruction from samples from a single thermodynamic state is studied for the concrete case of the Hopfield model. In the so-called memory regime, the Hopfield model has a number of thermodynamic states (attractor states) which scales linearly with the system size. Samples generated from a single run of the model's dynamics will generally come from a single state only. (The particular state the system is 'attracted' to depends on the initial conditions.) In this regime, the cavity-recursion equations (76) typically have multiple solutions, just like the TAP Equations (55) in the low-temperature phase. These solutions can be identified as fixed points of belief propagation, see Section 2.2.10. For a system where couplings do not form a tree, each individual fixed point only approximates the marginal probability distributions inside a single non-ergodic component of the full Gibbs measure. However, this approximation can become exact in the thermodynamic limit. (A necessary condition is that loops are sufficiently long, so connected correlations measured in a particular state decay sufficiently quickly with distance, where distance is measured along the graph of non-zero couplings.)

Assuming that one is able to find a fixed point of the cavity-recursion equations, it is a straightforward computation to express the Boltzmann weight (1) in terms of the full set of the belief propagation marginals and of the ratio between the cavity-recursion partition function and the true partition function of the model. For any fixed point, labelled by the index α, the Boltzmann weights are

$$p(s) = \frac{Z_{BP}^\alpha}{Z(J,h)} \prod_i p_i^\alpha(s_i) \prod_{i<j} \frac{p_{ij}^\alpha(s_i,s_j)}{p_i^\alpha(s_i)p_j^\alpha(s_j)}. \tag{115}$$

This result also applies to more general forms of the energy function [35]. It was first derived for the zero-temperature case in the context of the so-called tree-reweighted approximation to the partition function [117].

The probability distribution (115) can be used as the starting point for reconstruction in at least two ways. The simpler one consists in replacing $p_{ij}^\alpha(s_i,s_j)$ and $p_i^\alpha(s_i)$ by their sample estimates inside state α and then solving the identity between (115) and (1) with respect to J and h: $Z(J,h)$ cancels out and one is left with the independent pair approximation of Section 2.2.12. Solving

this identity can be done using belief propagation, see Section 2.2.11 and [35]. This approach suffers from the fact that in general there is no belief propagation fixed point for real data sets.

A more promising approach is to guide the belief propagation equations to converge to a fixed point corresponding to an appropriate ergodic component *close* to the empirical data. In fact, ignoring the information that the data come from a single ergodic component results in a large reconstruction error as one is effectively maximizing the wrong likelihood: the (reduced) free energy $-\ln Z$ appearing in the likelihood (16) needs to be replaced by a free energy restricted to a single thermodynamic state. An algorithmic implementation of this idea consists in restricting the spin configurations to a subset of configuration $\Omega_d(\xi)$, e.g. formed by a hypersphere of diameter d centered around the centroid ξ of the data samples. The system is forced to follow the measure

$$p_d(\mathbf{s}) \propto I[\mathbf{s} \in \Omega_d(\xi)] \, e^{\beta(\sum_{i<j} J_{ij}s_is_j + \sum_i h_is_i)}, \tag{116}$$

where the indicator function $I[\mathbf{s} \in \Omega_d(\xi)]$ for the set of configurations $\Omega_d(\xi)$ is one for $\mathbf{s} \in \Omega_d(\xi)$ and zero otherwise. The BP approach can be used to enforce that both magnetizations $\{m_i^{(d)}\}$ and correlations $\{c_{ij}^{(d)}\}$ are computed within the subspace $\Omega_d(\xi)$, see [35] for algorithmic details. In this way, the reconstruction of the couplings and local fields can be done by maximising the correct likelihood, i.e. replacing the partition function in the log-likelihood (16) with the partition function computed over the subset of configurations $\Omega_d(\xi)$. In this way, the parameters describing all thermodynamic states can be inferred from configurations of the system sampled from a single state.

Inverse problems in non-ergodic systems remain a challenging topic with many open questions, for instance if the approach applied above to the Hopfield model can be extended to infer generic couplings from samples from a single thermodynamic state.

2.6. *Parameter reconstruction and criticality*

Empirical evidence for critical behaviour has been reported in systems as diverse as neural networks [21,22] and financial markets [41,80], leading to the intriguing hypothesis that some information-processing systems may be operating at a critical point [151]. A key signature of criticality is broad tails in some quantities, for instance the distribution of returns in a financial market, or avalanches of neural activity, whose sizes are distributed according to a power law. A second sign of criticality involves an inverse statistical problem: A statistical model like the Ising system with parameters chosen to match empirical data (such as neural firing patterns or financial data) shows signs of a phase transition [141,151,152,211,221,223]. Specifically, the heat capacity of an Ising model with parameters J_{ij} and h_i reconstructed from data shows a peak in the heat capacity as a function of the temperature. Temperature is introduced by changing the couplings to βJ_{ij} and βh_i. Varying the inverse temperatures β, the heat capacity $C_h \equiv -\beta^2 \partial \langle H \rangle / \partial \beta$ shows a pronounced maximum near or at $\beta = 1$, that is, at the model parameter values inferred from the data. The implication is that the parameters of the reconstructed model occupy a very particular area in the space of all model parameters, namely one resulting in critical behaviour.

An example is shown in Figure 7. It is based on recordings of 160 neurons in the retina of a salamander taken by the Berry lab [221,223]. As described in Section 1.1, time is discretized into intervals of 20 ms duration, and the spin variable s_i takes on the values $+1$ if neuron i fires during a particular interval, and -1 if it does not fire. Correlations and magnetizations are then computed by averaging over 297 retinal stimulation time courses, where during each time course, the retina is subjected to the same visual stimulation of 953×20ms duration. During most time intervals, a neuron is typically silent. As a result, spin variables (each representing a different neuron) have negative magnetizations, with mean and standard deviation -0.93 ± 0.06 over all spins. We note
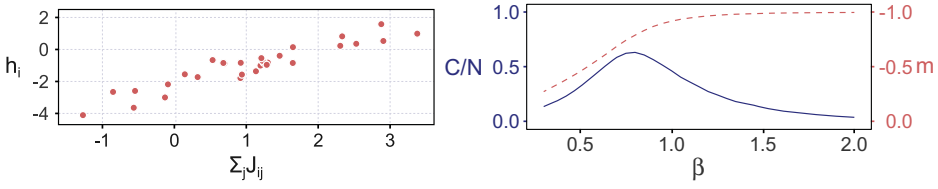
Figure 7. The Ising model with parameters matching the statistics of neural firing patterns. Couplings and external fields are generated from neural data [221,223] as described in the text. The system size here is $N = 30$. Left: Fields and couplings turn out not to be independent, but obey a linear–affine relationship, which is due to neural firing rates and hence magnetizations being approximately constant across neurons. Right: We simulate the model with reconstructed couplings at different temperatures using Monte Carlo simulations. The heat capacity (solid line) shows a peak near $\beta = 1$, and the magnetization $m = (1/N) \sum_i m_i$ (dashed, axis on the right) goes from zero to minus one as the inverse temperature is increased.

that the magnetizations are fairly homogeneous across spins. Connected correlations between spins are small, with a slight bias towards positive correlations; mean and standard deviation over all spin pairs are 0.006 ± 0.017. Both points turn out to be important.

We use the pseudolikelihood method of Section 2.3 to reconstruct the model parameters from the firing patterns the first $N = 30$ neurons. Similar results are found for different system sizes. Figure 7(left) shows that the reconstructed model parameters indeed occupy a particular region of the space of model parameters: External fields h_i are linked to the sum over couplings $\sum_j J_{ij}$ by a linear–affine relationship. If the magnetizations $\{m_i\}$ are all equal to some m, such a relationship follows directly from the mean-field equation $h_i = \text{artanh } m_i - \sum_j J_{ij} m_j$, along with slope $-m$ and zero-offset artanh m.

Now, we simulate the Ising model with the reconstructed parameters, but rescale both couplings and fields by β. At inverse temperature $\beta = 1$, magnetizations and correlations are close to the magnetizations and correlations found in the original data. However, away from $\beta = 1$, different fields $h_i = (1/\beta) \text{artanh} m_i - (1/\beta) \sum_j J_{ij} m_j$ (in the mean-field approximation) would be needed to retain the same magnetization. However, as the fields remain fixed, the magnetizations change instead, reaching $-1(0)$ in the limit $\beta \to \infty(0)$. Between these limits, the magnetizations, and hence the energy changes rapidly, leading to a peak in the heat capacity. For small, largely positive connected correlations C_{ij} we expect couplings to scale as $1/N$, resulting in a phase transition described by the Curie–Weiss model. Indeed, a finite-size analysis shows the peak getting sharper and moving closer to $\beta = 1$ as the system size is increased [141,151,152,211,223]. From this, we posit that the peak in the heat capacity indeed signals a ferromagnetic phase transition, however, not necessarily one underlying the original system. Instead, we generically expect such a ferromagnetic transition whenever a system shows sufficient correlations and magnetizations that are neither plus nor minus one: The corresponding couplings and fields then lie near (not at) a critical point, which is characterized by the emergence of non-zero magnetizations. Changing the temperature will drive the system to either higher or lower magnetizations, and away from the critical point.

This effect may not be limited to a ferromagnetic transition induced by the empirical magnetization. In [141], Mastromatteo and Marsili point out a link between the criticality of inferred models and information geometry [13,157]: susceptibilities like the magnetic susceptibility can be interpreted as the entries in the so-called Fisher information matrix used to calculated the covariances of ML estimates [56]. A high susceptibility implies that different parameter values can be distinguished on the basis of limited data, whereas low susceptibilities mean that the likelihood does not differ sufficiently between different parameter values to significantly favour

one parameter value over another. Thus, a critical point corresponds to a high density of models whose parameters are distinguishable on the basis of the data [141].

3. Non-equilibrium reconstruction

In a model of interacting magnetic spins such as Equation (2), each pair of spins $i < j$ contributes a term $J_{ij}s_is_j$ to the Hamiltonian. The resulting effective fields on a spin i, $\sum_j J_{ij}s_j + h_i$, involve symmetric couplings $J_{ij} = J_{ji}$; spin i influences spin j as much as j influences i. A stochastic dynamics based on such local fields, such as Monte Carlo dynamics, obeys detailed balance, which allows one to determine the steady-state distribution [83].

In applications such as neural networks or gene regulatory networks discussed in Sections 1.1 and 1.2, we have no reason to expect symmetric connections between neurons or between genes; a synaptic connection from neuron i to neuron j does not imply a link in the reverse direction. Stochastic systems generally relax to a steady state at long times, where observables no longer change with time. However, in systems with asymmetric couplings, the resulting NESS violates detailed balance and differs from the Boltzmann distribution. As a result, none of the results of Section 2 on equilibrium reconstruction apply to systems with asymmetric couplings.

In this section, we consider the inverse Ising problem in a non-equilibrium setting. We first review different types of spin dynamics, and then turn to the problem of reconstructing the parameters of a spin system from either time-series data or from samples of the NESS.

3.1. *Dynamics of the Ising model*

The Ising model lacks a prescribed dynamics, and there are many different dynamical rules that lead to a particular steady-state distribution. One particularly simple dynamics is the so-called Glauber dynamics [86], which allows the derivation of a number of analytical results. Other dynamical rules can be used for parameter reconstruction in the same way at least in principle. What dynamical rule is suitable for particular systems is however an open question. An approach which sidesteps this issue is to apply the maximum entropy principle to stochastic trajectories, known as the principle of maximum caliber [181]. In [139,158], such an approach is applied to analyse neural dynamics.

3.1.1. *Sequential Glauber dynamics*

Glauber dynamics can be based on discrete time steps, and at each step either one or all spins have new values assigned to them according to a stochastic rule. We first consider a sequential dynamics: In the transition from time t to $t + 1$, the label of a spin variable is picked randomly, say i. The value of the spin variable σ_i is then updated, with $s_i = \pm 1$ sampled from the probability distribution

$$p(s_i(t + 1) \,|\, \boldsymbol{s}(t)) = \frac{\exp\{s_i(t + 1)\theta_i(t)\}}{2\cosh(\theta_i(t))}, \tag{117}$$

where the effective local field is denoted

$$\theta_i(t) = \sum_j J_{ij}s_j(t) + h_i. \tag{118}$$

One way to implement this dynamics is to set

$$s_i(t + 1) = \text{sign}(\theta_i(t) + \xi(t)), \tag{119}$$

where $\xi_i(t)$ is drawn independently at each step from the distribution $p(\xi) = 1 - \text{th}^2(\xi)$.

If the couplings J_{ij} between spins are symmetric and there are no self-couplings J_{ii}, the sequential Glauber dynamics (117) obeys detailed balance, and the distribution of spin configurations at long times relaxes to a steady state described by the Boltzmann distribution (11). However, the NESS arising at long times from Glauber dynamics with non-symmetric couplings is generally not known. Nevertheless, there are several exact relations that follow from Equation (117), and those can be exploited for inference.

Suppose that spin i is picked for updating in the step from time t to $t+1$. Averaging over the two possible spin configurations at time $t+1$, we obtain the magnetization

$$m_i(t+1)_{s(t)} \equiv \langle \sigma_i(t+1) \rangle_{s(t)} = \text{th}(\theta_i(t)), \tag{120}$$

where the average is conditioned on the configuration of all other spins at time t through the effective local field θ_i. If spin i is not updated in this interval, the conditioned magnetization is trivially $m_i(t+1)_{s(t)} = s_i(t)$. In the steady state, effective fields $\Theta_i = \sum_j J_{ij}\sigma_j(t) + h_i$ and spins configurations $\boldsymbol{\sigma}$ are random variables whose distribution no longer depends on time; their averages give

$$m_i \equiv \langle \sigma_i \rangle = \langle \text{th}(\Theta_i) \rangle = \left\langle \text{th}\left(\sum_j J_{ij}\sigma_j + h_i\right) \right\rangle. \tag{121}$$

Similarly, the pair correlation between spins at sites $i \neq j$ at equal times obeys in the NESS

$$\chi_{ij} \equiv \langle \sigma_i\sigma_j \rangle = \tfrac{1}{2}\langle \sigma_i \text{th}(\Theta_j) \rangle + \tfrac{1}{2}\langle \sigma_j \text{th}(\Theta_i) \rangle, \tag{122}$$

with the two terms arising from instances where the last update of i was before the last update of j, and vice versa. Likewise, the pair correlation of spin configurations at consecutive time intervals is in the steady state

$$\phi_{ij} \equiv \langle \sigma_i(t+1)\sigma_j(t) \rangle = \frac{1}{N}\langle \text{th}(\Theta_i)\sigma_j \rangle + \frac{N-1}{N}\chi_{ij}. \tag{123}$$

These relationships are exact, but are hard to evaluate since the averages on the right-hand sides are over the statistics of spins in the NESS, which is generally unknown. Below, these equations will be used in different ways for the reconstruction of the model parameters.

The dynamics (117) defines a Markov chain of transitions between spin configurations differing by at most one spin flip. Equivalently, one can also define an asynchronous dynamics described by a Master equation, where time is continuous and the time between successive spin flips is a continuous random variable.

3.1.2. *Parallel Glauber dynamics*

Glauber dynamics can also be defined with parallel updates, where *all* spin variables can change their configurations in the time interval between t and $t+1$ according to the stochastic update rule

$$p(s(t+1) \,|\, s(t)) = \frac{\exp\{\sum_i s_i(t+1)\theta_i(t)\}}{\prod_i 2\cosh(\theta_i(t))}. \tag{124}$$

This update rule defines a Markov chain consisting of stochastic transitions between spin configurations. The resulting dynamics is not realistic for biological networks, as the synchronous update requires a central clock. It is, however, implemented easily in technical networks, and is widely used for its simplicity. For a symmetric coupling matrix, the steady state can still be specified in closed form using Peretto's pseudo-Hamiltonian [175].

Magnetizations and correlations in the NESS obey simpler relationships for parallel updates than for sequential updates; with the same arguments as above, one obtains

$$
\begin{aligned}
m_i &= \langle \text{th}(\Theta_i) \rangle, \\
\chi_{ij} &= \langle \text{th}(\Theta_i)\,\text{th}(\Theta_j) \rangle, \\
\phi_{ij} &= \langle \text{th}(\Theta_i)\sigma_j \rangle.
\end{aligned}
\tag{125}
$$

3.2. *Reconstruction from time series data*

The reconstruction of system parameters is surprisingly easy on the basis of time series data, which specifies the state of each spin variable at M successive time points. An application where such data is widely available is the reconstruction of neural networks from temporal recordings of neural activity. Given a stochastic update rule like Equations (117) or (124), the (log-) likelihood given a time series $D = \{s(t)\}$ is

$$
L_{\mathrm{D}}(\boldsymbol{J}, \boldsymbol{h}) = \frac{1}{M} \sum_{t=1}^{M-1} \ln p(s(t+1) \,|\, s(t)).
\tag{126}
$$

The likelihood can be evaluated over any time interval, both in the NESS or even before the steady state has been reached. This approach is not limited to non-equilibrium systems; whenever time series data are available, it can be used equally well to reconstruct symmetric couplings.

3.2.1. *Maximization of the likelihood*

For Glauber dynamics with parallel updates, the likelihood is

$$
L_{\mathrm{D}}(\boldsymbol{J}, \boldsymbol{h}) = \frac{1}{M} \sum_{t=1}^{M-1} \sum_{i} [s_i(t+1)\theta_i(t) - \ln 2 \cosh(\theta_i(t))].
\tag{127}
$$

Unlike the likelihood (16) arising in equilibrium statistics, the non-equilibrium likelihood can be evaluated easily, as the normalization is already contained in the term $\cosh(\theta)$. Derivatives of the likelihood (127) are

$$
\begin{aligned}
\frac{\partial L_{\mathrm{D}}}{\partial h_i}(\boldsymbol{J}, \boldsymbol{h}) &= \frac{1}{M} \sum_{t=1}^{M-1} t[s_i(t+1) - \text{th}\,\theta_i(t)], \\
\frac{\partial L_{\mathrm{D}}}{\partial J_{ij}}(\boldsymbol{J}, \boldsymbol{h}) &= \frac{1}{M} \sum_{t=1}^{M-1} [s_i(t+1)s_j(t) - \text{th}(\theta_i(t))s_j(t)].
\end{aligned}
\tag{128}
$$

These derivatives can be evaluated in MN^2 computational steps and can be used to maximize the likelihood by gradient ascent [189].

The derivatives of the likelihood can also be written as time averages over the data, which makes a conceptual connection with logistic regression apparent [189]. The derivative with respect to fields gives $\partial L_{\mathrm{D}}/\partial h_i = \langle \sigma_i(t+1) \rangle_T^{\mathrm{D}} - \langle \text{th}(\sum_j J_{ij}\sigma_j(t) + h_i) \rangle_T^{\mathrm{D}}$ and similarly for the derivative with respect to the couplings (the subscript refers to the temporal average, the super-script to configurations taken from the data). Parallel Glauber dynamics (124) defines a logistic regression giving the statistics of $\sigma_i(t+1)$ as a function of $\{s_j(t)\}$. In a hypothetical data set of P trajectories of the system $D = \{s^p(t)\}_{t=1}^{M}$, each realization can be considered as $M-1$ realizations

$(s_i^p(t+1), \{s_j^p(t)\})$ of such regression pairs. The discussion in Section 2.3 then applies directly, giving rise to regression equations for fields and couplings.

For Glauber dynamics with sequential updates, the situation is not quite so simple. In each time interval one spin is picked for updating, but if this spin is assigned the configuration it had before it is impossible to tell which spin was actually chosen [249]. The solution is to sum the likelihood (126) over all spins (each of which might have been the one picked for updating with probability $1/N$). It turns out that both the intervals when a spin was actually flipped, and the intervals when no spin was flipped are required for the inference of couplings and fields.

3.2.2. *Mean-field theory of the NESS*

As in the case of equilibrium reconstruction, mean-field theory offers an approximation to the ML reconstruction that can be evaluated quickly. The speed-up is not quite as significant as it is in equilibrium, because the likelihood (127) can already be computed in polynomial time in N and M.

In Section 2.2.13, the mean-field equation (51) and the TAP equation (55) were derived in an expansion around a factorizing ansatz for the equilibrium distribution. Kappen and Spanjers showed that, remarkably, exactly the same equations emerge as first- and second-order expansions around the same ansatz for the NESS as well [113]. Suppose that the (unknown) steady-state distribution of configurations $p(s)$ in the NESS is 'close' to another distribution with the *same magnetizations*, $q(s) = \prod_i((1 + m_i s_i)/2)$. According to Equation (121), this distribution describes the NESS of a different model with fields $h_i^{(q)} = \mathrm{artanh} m_i$ and couplings $J_{ij}^{(q)} = 0$. Next, we consider a small change in these fields and couplings and ask how the magnetizations change. To first order this change is given by the derivatives of Equation (121) evaluated at $h_i = h_i^{(q)}$ and $J_{ij} = 0$

$$\Delta m_i = \sum_j \frac{\partial \langle s_i \rangle}{\partial h_j}\bigg|_q \Delta h_j + \sum_{k,j} \frac{\partial \langle s_i \rangle}{\partial J_{kj}}\bigg|_q \Delta J_{kj} + \cdots$$

$$= (1 - m_i^2)\Delta h_i + (1 - m_i^2)\sum_j \Delta J_{ij} m_j + \cdots . \tag{129}$$

Setting $\Delta h_i = h_i - h_i^{(q)}$ and $\Delta J_{ij} = J_{ij} - J_{ij}^{(q)} = J_{ij}$ and demanding that magnetizations remain unchanged under this change of fields and couplings gives to first order $h_i^{(q)} = h_i + \sum_j J_{ij} m_j$ and thus

$$m_i = \mathrm{th}\left(h_i + \sum_j J_{ij} m_j\right). \tag{130}$$

Carrying the expansion (129) to second order in Δh_j and ΔJ_{jk} yields the TAP equations (55) [113]. As the equations for the magnetizations (121) and correlations (125) are identical for sequential and parallel dynamics, the mean-field and TAP equations apply equally to both types of dynamics. Roudi and Hertz [188] extended these results to time scales before the NESS is reached using a generating functional approach.

The next step is to apply the mean-field approximation to the correlations (125) for parallel Glauber updates [188,250]. For sequential updates, analogous result can be derived from Equations (122) and (123). We expand the effective local field $\Theta_i = h_i + \sum_j J_{ij} m_j + \sum_j J_{ij}\delta\sigma_j$ around the mean field $\theta_i^{\mathrm{MF}} = h_i + \sum_j J_{ij} m_j$ writing $\delta\sigma_i \equiv \sigma_i - m_i$. Expanding the th Θ_i -term of

the two-time pair correlation (125) in a formal expansion in powers of $\delta\sigma$ gives

$$
\begin{aligned}
\phi_{ij} &= \langle \mathrm{th}(\Theta_i)\sigma_j \rangle \\
&= \langle \mathrm{th}\,\theta_i^{\mathrm{MF}}\sigma_j \rangle + (1 - \mathrm{th}^2\theta_i^{\mathrm{MF}}) \sum_l J_{il}\langle \delta\sigma_l\sigma_j \rangle + \cdots \\
&= m_i m_j + (1 - m_i^2) \sum_l J_{il}(\chi_{lj} - m_l m_j) + \cdots .
\end{aligned}
\tag{131}
$$

To first order, this equation can be read as a matrix equation in the connected correlations functions, $D_{ij} \equiv \phi_{ij} - m_i m_j = \sum_{m,l} A_{im} J_{ml} C_{lj}$ with $A_{im} = \delta_{im}(1 - m_i^2)$ and $C_{ij} = \chi_{ij} - m_i m_j$. Inverting this relationship leads to the mean-field reconstruction

$$
\boldsymbol{J}^{\mathrm{MF}} = \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1},
\tag{132}
$$

based on sample averages of magnetizations and connected correlations in the NESS. The reconstruction based on the TAP equation can be derived analogously [188,250]; the result is of the same form as Equation (132) with $A_{im} = \delta_{im}(1 - m_i^2)(1 - F_i)$, where F_i is the smallest root of

$$
F_i(1 - F_i)^2 = (1 - m_i^2) \sum_j (J^{\mathrm{MF}})_{ij}^2 (1 - m_j^2).
\tag{133}
$$

3.2.3. *The Gaussian approximation*

For an asymmetric coupling matrix, with no correlation between J_{ij} and J_{ji}, the statistics of the effective local fields in the NESS is remarkably simple [147]: In the thermodynamic limit, Θ_i turns out to follow a Gaussian distribution at least in some regimes. This distribution is characterized by a mean θ_i^{MF}, standard deviation Δ_i, and covariance ϵ_{ij}. Using the definition of the effective local field (118), these parameters are linked to the spin observables by

$$
\begin{aligned}
g_i &= h_i + \sum_j J_{ij} m_j, \\
\Delta_i^2 &= \sum_{lk} J_{il} C_{lk} J_{ik}, \\
\epsilon_{ij} &= \sum_{lk} J_{il} C_{lk} J_{jk}.
\end{aligned}
\tag{134}
$$

The key idea is that one can transform back and forth between $\{\sigma_i(t)\}$ and $\{\Theta_i(t)\}$ via

$$
\begin{aligned}
\boldsymbol{\Theta}(t) &= \boldsymbol{J}\boldsymbol{\sigma}(t) + \boldsymbol{h}, \\
\boldsymbol{\sigma}(t) &= \boldsymbol{J}^{-1}(\boldsymbol{\Theta}(t) - \boldsymbol{h})
\end{aligned}
\tag{135}
$$

and evaluate the correlation functions within the Gaussian theory. With $\theta_i = \theta_i^{\mathrm{MF}} + \sqrt{\Delta_i}x$ and $\theta_j = \theta_j^{\mathrm{MF}} + \sqrt{\Delta_j}y$, where x and y are univariate Gaussian random variables, one obtains from

$D_{ik} = \langle \text{th}(\Theta_i)\sigma_k \rangle - \langle \text{th}(\Theta_i) \rangle \langle \sigma_k \rangle$

$$\sum_k J_{jk} D_{ik} = \langle \text{th}(\Theta_i)(\Theta_j - h_j) \rangle - \langle \text{th}(\Theta_i) \rangle \langle \Theta_j - h_j \rangle$$

$$= \langle \text{th}(\theta_i^{\text{MF}} + \sqrt{\Delta_i} x)\sqrt{\Delta_j} y \rangle$$

$$= \epsilon_{ij} \langle 1 - \text{th}^2(\theta_i^{\text{MF}} + \sqrt{\Delta_i} x) \rangle. \tag{136}$$

In the last step, we have used the fact that covariances between spin variables are small [147]. Inserting the result for the covariance (134) gives again an equation of the same form as Equation (132), however, with $A_{im} = \delta_{im} \int dx (1/\sqrt{2\pi}) e^{-x^2/2}[1 - \text{th}^2(\theta_i^{\text{MF}} + \sqrt{\Delta_i} x)]$. Mean-field theory neglects the fluctuations here and renders this term as $1 - m_i^2$, whereas the fluctuations can be captured more accurately under the Gaussian theory. However, A_{im} cannot be determined directly from the data alone, but also require the couplings; Mézard and Sakellariou [147] give an iterative scheme to infer the parameters of the effective local fields (134) as well as the coupling matrix and magnetic fields. The typical-case performance of the Gaussian theory in the thermodynamic limit has been analysed within the framework of statistical learning [10], finding the Gaussian theory breaks down at strong couplings and a small number of samples.

The Gaussian distribution of local fields is not limited to the asymmetric Ising model. In fact, the asymmetric Ising model is one particular example from a class of models called generalized linear models. In this model class, the Gaussian approximation has been used in the context of neural network reconstruction [227] already prior to its application to the asymmetric Ising model.

3.2.4. *Method comparison*

We compare the results of the mean-field approximation, the Gaussian approximation, and the maximization of the exact likelihood (127). As in Section 2.4, we draw couplings from a Gaussian distribution with mean zero and standard deviation $\beta/\sqrt{N}$, but now couplings J_{ij}^0 are statistically independent of J_{ji}^0, so the matrix of couplings is in general asymmetric. Fields are drawn from a uniform distribution on the interval $[-0.3\beta, 0.3\beta]$. We then sample a time series of $M = 15,000$ steps by parallel updates under Glauber dynamics (124). The scatter plots in the top row of Figure 8 compares couplings and fields reconstructed by different methods with the couplings and fields of the original model. It shows that couplings are significantly underestimated by the mean-field reconstruction, a bias which is avoided by the Gaussian theory. The right-hand plot shows the relative reconstruction error (114) against β. As in the equilibrium case shown in Figure 5, reconstruction by any method is limited at small β by sampling noise. At strong couplings β, mean-field theory breaks down. Also at strong couplings, the iterative algorithm for Gaussian approximation reconstruction converges very slowly and stops when the maximum number of iterations is reached (here set to 50,000 steps).

3.3. *Outlook: reconstruction from the steady state*

Time series data are not available in all applications. This prompts the question how to reconstruct the parameters of a non-equilibrium model from independent samples of the steady state. It is clear that, unlike for equilibrium systems, pairwise correlations are insufficient to infer couplings: The matrix of correlations is symmetric, whereas the matrix of couplings is asymmetric for non-equilibrium systems. Hence, there are twice as many free parameters as there are observables. Similarly, using an equilibrium model like Equation (1) (maximum entropy model) on data
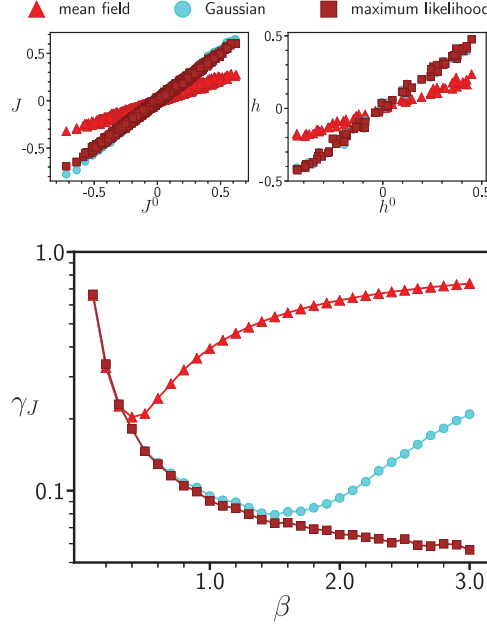
Figure 8. Reconstructing the asymmetric Ising model from time series. The scatter plots in the upper panels compare the reconstructed couplings and fields with the couplings and fields underlying the original model. The system size is $N = 64$, the standard deviation of the original couplings is $\beta/\sqrt{N}$ with $\beta = 1.5$, the original fields are uniformly distributed between $[-0.3\beta, 0.3\beta]$, and $M = 15{,}000$ time steps are sampled. The bottom plot gives the reconstruction error as a function of β, see text.

generated by a non-equilibrium model would give parameters matching the two-spin observables, but entirely different from the true parameters underlying the data.

One solution uses three-spin correlations to infer the couplings [63]. One problem of this approach is that connected three-spin correlations are small since the effective local fields are well described by a multi-variate Gaussian distribution (see Section 3.2.3).

A second approach uses perturbations of the NESS: We measure one set of pair correlations at certain (unknown) model parameters, and then a second set of pair correlations of a perturbed version of the system. Possible perturbations include changing one or several of the couplings by a known amount, changing the magnetic fields, or fixing particular spins to a constant value. This generates two sets of coupled equations specifying two symmetric correlations matrices, which can be solved for one asymmetric coupling matrix. Conceptually, such an approach is well known in biology, where altering parts of a system and checking the consequences is a standard mode of scientific inquest. Neural stimulation can lead to a rewiring of neural connections, and the effects of this neural plasticity can be tracked in neural recordings [183]. An exciting development is optogenetic tools, which allow to stimulate and monitor the activity of individual neurons *in vivo* [220]. In the context of inferring gene regulatory networks from gene expression data, perturbation-based approaches have been used both with linear [217] and non-linear models of gene regulation [148,159], see also Section 1.2. In this context, it is also fruitful to consider the genetic variation occurring in a population of cells as a source of perturbations. Such an approach has already been used in gene regulatory networks [186,253], and with the expansion of tools to analyse large numbers of single cells [43,248], it may soon spread to other types of networks.

4.   Conclusions

At the end of this overview, we step back and summarize the aims and motivations behind the inverse Ising problem, discuss the efficacy of different approaches, and outline different areas of research that may involve the statistical mechanics of inverse problems in the future.

Motivation. The inverse Ising problem arises in the context of very different types of questions connected with the inference of model parameters. The first and most straightforward question appears when data actually is generated by a process obeying detailed balance and has pairwise couplings between binary spins. The problem of inferring the parameters of such a model can then be phrased in terms of the equilibrium statistics (11), the maximization of the likelihood (16), and the use of approximation schemes discussed in Section 2.

The second question arises when data are generated by a different, and possibly entirely unknown type of process, and we seek a statistical description of the data in terms of a simpler model matching only particular aspects of the observed data. An example is models with maximum entropy given pairwise correlations (Section 2.2.3) used to describe neural data in Section 1.1.

The assumptions behind the Ising model (pairwise couplings, binary spins, etc.) can be relaxed. For instance, multi-valued Potts spin variables have been used extensively in models of biological sequences [194,238]. The extension of the Ising model to models with three- and four-spin couplings has not been used yet in the context of inference. Nevertheless, for many of the methods of Section 2, the extension beyond pairwise couplings would be straightforward. A much larger rift lies between equilibrium and non-equilibrium models. Parameter inference of a non-equilibrium model is often based on data beyond independent samples of the steady state, specifically time-series data.

Methods. The practical question how to infer the couplings and fields that parametrize an Ising model has been answered using many different approximations with different regimes of validity. Section 2.4 gives an overview. For data sampled from the equilibrium distribution (1), the pseudolikelihood approach of Section 2.3 gives a reconstruction close to the optimal one (in fact, asymptotically close in the number of samples). However, this is paid for by a computational effort that scales with the number of samples. For the non-equilibrium regime and a time series of configurations sampled from the stochastic dynamics, the likelihood of a model can be evaluated exactly comparatively easily.

Both the pseudolikelihood and the likelihood of a time series can be understood in the language of regression. Thus a single framework links two of the most successful approaches, in both the equilibrium and non-equilibrium setting. Regression singles out one spin variable as a dependent variable and treats the remainder as independent variables. It then characterizes the statistics of the dependent variable given configurations of the independent variables.

The application of novel concepts to the inverse Ising problem and the development of new algorithms continues. An exciting new direction is interaction screening [130,232]. Vuffray, Misra, Lokhov, and Chertkov introduce the objective function

$$\frac{1}{M} \sum_{\mu=1}^{M} e^{H_i(s^\mu)} = \frac{1}{M} \sum_{\mu=1}^{M} \exp\left\{ -\sum_{j \neq i} J_{ij} s_i^\mu s_j^\mu - h_i s_i^\mu \right\} \tag{137}$$

to be minimized with respect to the ith column of the coupling matrix and the magnetic field on spin i by convex optimization. $H_i(s)$ denotes the part of the Hamiltonian containing all terms in s_i, see Section 2.3. Sparsity or other properties of the model parameters can be effected by appropriate regularization terms. This objective function aims to find those parameters which 'screen the interactions' (and the magnetic field) in the data, making the sum over samples in Equation (137)

as balanced as possible. To illustrate the appeal of this objective function, we look at the infinite sampling limit, where the summation over samples in the objective function (137) can be replaced by a summation over the configurations, reweighted by the Boltzmann distribution with the true couplings $\boldsymbol{J}^*$ and fields $\boldsymbol{h}^*$,

$$\frac{1}{M} \sum_{\mu=1}^{M} e^{H_i(\boldsymbol{s}^{\mu})} \approx \sum_{\boldsymbol{s} \backslash s_i} \frac{1}{Z(\boldsymbol{J}^*, \boldsymbol{h}^*)} e^{\sum_{k>j; k, j \neq i} J_{kj}^* s_k s_j + \sum_{k \neq i} h_k^* s_k}$$
$$\times \sum_{s_i} e^{-\sum_j (J_{ij} - J_{ij}^*) s_i s_j - (h_i - h_i^*) s_i}. \tag{138}$$

Since the last sum as a function of J_{ij} and h_i (for a fixed i) is convex and even when being reflected around J_{ij}^* and h_i^*, so is the whole objective function. It follows immediately that this objective function has a unique global minimum at $J_{ij} = J_{ij}^*$ and $h_i = h_i^*$.

Of course, the sample average over $e^{H_i(\boldsymbol{s})}$ is affected by sampling fluctuations when the number of samples is small. However, nevertheless the minimum of this objective function nearly saturates the information theoretic bounds on the reconstruction of sparse Ising models of Santhanam and Wainwright [193] (see Section 2.2.4).

The flat histogram method in Monte Carlo simulations uses a similar rebalancing with respect to the Boltzmann measure [234]. Also, there may be conceptual links between interaction screening and the fluctuations theorems like the Jarzynski equality [106], which also take sample averages over exponentials of different thermodynamic quantities.

Another recent development concerning the sparse inverse Ising problem is the use of Bayesian model selection techniques by Bulso et al. [38].

More to come. Over the last two decades, interest in inverse statistical problems has been driven by technological progress and this progress is likely to continue opening up new applications. Beyond the extrapolation of technological developments, there are several broad areas at the interface of statistical mechanics, statistics, and machine learning where inverse statistical problems like the inverse Ising problem might play a role in the future.

- *Stochastic control theory.* Stochastic control theory seeks to steer a stochastic system towards certain desired states [111]. An inverse problem arises when the parameters describing the stochastic system are only partially known. As a result, its response to changes in external control parameters ('steering') must be predicted on the basis of its past dynamics. A recent application is the control of cell populations, specifically populations of cancer cells [77]. Such cell populations evolve stochastically due to random fluctuations of cell duplications and cell deaths. Birth and death rates of at least part of the population can be controlled by therapeutic drugs.
- *Network inference.* Like regulatory connections between genes discussed in Section 1.2, metabolic and signalling interactions also form intricate networks. An example where such a network optimizes a specific global quantity is flux-balance analysis [165], which models the flow of metabolites through a metabolic network. Inverting the relationship between metabolic rates and metabolite concentrations allows in principle to infer the metabolic network on the basis of observed metabolite concentrations [61].
- *Causal analysis.* The *do*-calculus developed by Judea Pearl seeks to establish causal relationships behind statistical dependencies [172]. *do*-calculus is based on interventions such as fixing a variable to a particular value, and then observing the resulting statistics of other variables. Recently, Aurell and Del Ferrano have found a link between *do*-calculus and the dynamic cavity method from statistical physics [6].

- *Maximum entropy and dynamics.* The notion of a steady-state distribution with maximum entropy has been generalized to a maximum entropy distribution over *trajectories* of a dynamical system [181]. Like the maximum entropy models of Sections 1.1–1.3, this approach can be used to derive simple effective models of the *dynamics* of a system, whose parameters can be inferred from time series. In fact, Glauber dynamics with parallel updates gives the maximum entropy distribution of $\sigma(t+1)$ given $s(t)$. So far, applications have been in neural modelling [139,158], in the effective dynamics of quantitative traits in genetics [18,32], and in flocking dynamics [46].

- *Hidden variables.* Even in large-scale data sets, there will be variables that are unobserved. Yet, those hidden variables affect the statistics of other variables, and hence the inferences we make about interactions between the observed variables [19,69,191,222]. This can lead to a signature of critical behaviour, even when the original system is not critical [140,196].

- *High-dimensional statistics and inference.* In many applications, the number of systems parameters to be inferred is of the same order of magnitude or exceeds the sample size. The field of high-dimensional statistics deals with this regime, and a fruitful interaction with statistical physics has emerged over the last decade [247], spurred by applications like compressed sensing. Vuffray et al. [232] propose the objective function (137) for the inverse Ising problem in the high-dimensional regime. Other objective functions might perform even better, and the optimal objective function may depend on the number of samples and the statistics of the underlying couplings. Berg [23] and Bachschmid-Romano and Opper [11] use the statistical mechanics of disordered systems to find the objective function which minimizes the difference between the reconstructed and the underlying couplings for Gaussian distributed couplings.

- *Restricted Boltzmann machines and deep learning.* Recently, (deep) feed-forward neural networks have re-established themselves as powerful learning architectures, leading to spectacular applications in the area of computer vision, speech recognition, data visualization, and game playing [123]. This progress has also demonstrated that the most challenging step in data analysis is the extraction of features from unlabelled data. There is a wide range of methods for feature extraction, the most well-known ones being convolutional networks for images [123,124], or auto-encoders [97,231], and restricted Boltzmann machines (RBMs) [97] for less structured data. RBMs are possibly one of the most general methods, although the algorithms used for finding the optimal parameters are heuristic and approximate in many respects. For instance, in the case of multiple layers, RMBs correspond to generic Boltzmann machines (with feedback loops), for which learning is a hard computational problem. Usually, the sub-optimal solution which is adopted is used to train each layer independently. These observations are not surprising since RBMs are nothing but an inverse Ising problem with a layer of visible spins connected to a layer of hidden spins. Empirical data are available only for the visible spins. The role of the hidden variables is to compress information and identify structural features in the data. Learning consists of finding the visible-to-hidden couplings $J_{ij} = J_{ji}$ and the local fields such that summing over to the hidden variables gives back a (marginalized) probability distribution over the visible variables which is maximally consistent with the data, i.e. has minimal KL divergence (18) from the empirical distribution of the data.

- *Learning phases of matter.* One aspect of the emerging applications of machine learning in quantum physics is the identification of non-trivial quantum states from data. This is of particular interest in phases of matter where the order parameter is either unknown or hard to compute (like the so-called entanglement entropy [104,125]). Techniques from machine learning, specifically deep learning with neural networks, have recently been used to classify quantum states without knowledge of the underlying Hamiltonian [45].

Acknowledgments

Disclosure statement

Funding

References

[1] D.H. Ackley, G.E. Hinton and T.J. Sejnowski, Cogn. Sci. 9(1) (1985) p.147.
[2] H. Akaike, IEEE Trans. Automat. Control 19(6) (1974) p.716.
[3] S.-I. Amari, H. Nakahara, S. Wu and Y. Sakai, Neural Comput. 15(1) (2003) p.127.
[4] B.C. Arnold and D. Strauss, Sankhya A (1991) p.233.
[5] E. Aurell, PLoS Comput. Biol. 12(5) (2016) p.e1004777.
[6] E. Aurell and G. Del Ferraro, J. Phys. Conf. Ser. 699 (2016) p.012002.
[7] E. Aurell and M. Ekeberg, Phys. Rev. Lett. 108(9) (2012) p.090201.
[8] E. Aurell, C. Ollion and Y. Roudi, Eur. Phys. J. B 77(4) (2010) p.587.
[9] K. Baba, R. Shibata and M. Sibuya, Aust. N. Z. J. Stat. 46(4) (2004) p.657.
[10] L. Bachschmid-Romano and M. Opper, J. Stat. Mech. Theory Exp. 2015(9) (2015) p.P09016.
[11] L. Bachschmid-Romano and M. Opper, A statistical physics approach to learning curves for the inverse Ising problem, preprint (2017). Available at arXiv:1705.05403
[12] M. Bailly-Bechet, A. Braunstein, A. Pagnani, M. Weigt and R. Zecchina, BMC Bioinform. 11(1) (2010) p.355.
[13] V. Balasubramanian, Neural Comput. 9(2) (1997) p.349.
[14] J.R. Banavar, A. Maritan and I. Volkov, J. Phys. Condens. Matter 22(6) (2010) p.063101.
[15] R. Bapat, S.J. Kirkland and M. Neumann, Linear Algebra Appl. 401 (2005) p.193.
[16] J.P. Barton, S. Cocco, E. De Leonardis and R. Monasson, Phys. Rev. E 90(1) (2014) p.012132.
[17] J.P. Barton, E. De Leonardis, A. Coucke and S. Cocco, Bioinformatics 32(20) (2016) p.3089.
[18] N.H. Barton and H.P. de Vladar, Genetics 181(3) (2009) p.997.
[19] C. Battistin, J. Hertz, J. Tyrcha and Y. Roudi, J. Stat. Mech. Theory Exp. 2015(5) (2015) p.P05021.
[20] R.J. Baxter, *Exactly Solvable Models in Statistical Mechanics*, Academic Press, London, 1982.
[21] C. Bedard, H. Kroeger and A. Destexhe, Phys. Rev. Lett. 97(11) (2006) p.118102.
[22] J.M. Beggs and N. Timme, Front. Physiol. 3 (2012) p.163.
[23] J. Berg, Statistical mechanics of the inverse Ising problem and the optimal objective function, preprint (2016). Available at http://arxiv.org/abs/1611.04281
[24] J. Berg, S. Willmann and M. Lässig, BMC Evol. Biol. 4(1) (2004) p.42.
[25] J. Besag, J. R. Stat. Soc. B 36(2) (1974) p.192.
[26] J. Besag, J. R. Stat. Soc. B 48(3) (1986) p.259.
[27] H. Bethe, Proc. R. Soc. Lond. A, 150 (1935) p.552.

[28] W. Bialek, A. Cavagna, I. Giardina, T. Mora, O. Pohl, E. Silvestri, M. Viale and A.M. Walczak, Proc. Natl. Acad. Sci. USA 111(20) (2014) p.7212.

[29] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale and A.M. Walczak, Proc. Natl. Acad. Sci. USA 109(13) (2012) p.4786.

[30] W. Bialek and R. Ranganathan, Rediscovering the power of pairwise interactions, preprint (2007). Available at arXiv:0712.4397

[31] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Heidelberg, 2006.

[32] K. Boďová, G. Tkačik and N.H. Barton, Genetics 202(4) (2016) p.1523.

[33] N.M. Bogolyubov, V.F. Brattsev, A.N. Vasil'ev, A.L. Korzhenevskii and R.A. Radzhabov, Theor. Math. Phys. 26(3) (1976) p.230.

[34] S.S. Borysov, Y. Roudi and A.V. Balatsky, Eur. Phys. J. B 88(12) (2015) p.1.

[35] A. Braunstein, A. Ramezanpour, R. Zecchina and P. Zhang, Phys. Rev. E 83(5) (2011) p.056114.

[36] T. Broderick, M. Dudik, G. Tkacik, R.E. Schapire and W. Bialek, Faster solutions of the inverse pairwise Ising problem, preprint (2007). Available at arXiv:0712.2437

[37] N.E. Buchler, U. Gerland and T. Hwa, Proc. Natl. Acad. Sci. USA 100(9) (2003) p.5136.

[38] N. Bulso, M. Marsili and Y. Roudi, J. Stat. Mech. Theory Exp. 2016(9) (2016) p.093404.

[39] L. Burger and E. Van Nimwegen, PLoS Comput. Biol. 6(1) (2010) p.e1000633.

[40] A.N. Burkitt, Biol. Cybern. 95(1) (2006) p.1.

[41] T. Bury, Physica A 392(6) (2013) p.1375.

[42] T. Bury, J. Stat. Mech. Theory Exp. 2013(11) (2013) p.P11004.

[43] C. Cadwell, A. Palasantza, X. Jiang, P. Berens, Q. Deng, M. Yilmaz, J. Reimer, S. Shen, M. Bethge, K. Tolias, R. Rickard Sandberg and T. Andreas, Nat. Biotechnol. 34 (2015) p.199.

[44] H.B. Callen, Phys. Lett. 4(3) (1963) p.161.

[45] J. Carrasquilla and R.G. Melko, Nat. Phys. 13 (2017) p. 431.

[46] A. Cavagna, I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone and A.M. Walczak, Phys. Rev. E 89(4) (2014) p.042707.

[47] H.J. Changlani, H. Zheng and L.K. Wagner, J. Chem. Phys. 143(10) (2015) p.102814.

[48] J.T. Chayes, L. Chayes and E.H. Lieb, Commun. Math. Phys. 93(1) (1984) p.57.

[49] M. Chertkov and V.Y. Chernyak, J. Stat. Mech. (2006) p.P06009.

[50] C. Chow and C. Liu, IEEE Trans. Inf. Theory 14 (1968) p.462.

[51] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson and M. Weigt, Inverse statistical physics of protein sequences: A key issues review, preprint (2017). Available at http://arxiv.org/abs/1703.01222

[52] S. Cocco, S. Leibler and R. Monasson, Proc. Natl. Acad. Sci. USA 106(33) (2009) p.14058.

[53] S. Cocco and R. Monasson, Phys. Rev. Lett. 106(9) (2011) p.090601.

[54] S. Cocco and R. Monasson, J. Stat. Phys. 147(2) (2012) p.252.

[55] S. Cocco, R. Monasson and M. Weigt, J. Phys. Conf. Ser. 473 (2013) p.012010.

[56] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, Hoboken, NJ, 2006.

[57] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1961.

[58] A.E. Dago, A. Schug, A. Procaccini, J.A. Hoch, M. Weigt and H. Szurmant, Proc. Natl. Acad. Sci. USA 109(26) (2012) p.E1733.

[59] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T.M. Allen, M. Altfeld, M. Carrington, D.J. Irvine, B.D. Walker and A.K. Chakraborty, Proc. Natl. Acad. Sci. USA 108(28) (2011) p.11530.

[60] D. de Juan, F. Pazos and A. Valencia, Nat. Rev. Genet. 14(4) (2013) p.249.

[61] D. De Martino, F. Capuani and A. De Martino, Phys. Biol. 13(3) (2016) p.036005.

[62] A. Decelle and F. Ricci-Tersenghi, Phys. Rev. Lett. 112(7) (2014) p.070603.

[63] S. Dettmer, H.C. Nguyen and J. Berg, Phys. Rev. 5(E 94) (2016) p.052116.

[64] P. D'haeseleer, S. Liang and R. Somogyi, Bioinformatics 16(8) (2000) p.707.

[65] K.A. Dill and J.L. MacCallum, Science 338(6110) (2012) p.1042.

[66] R.A. DiStasio Jr, É. Marcotte, R. Car, F.H. Stillinger and S. Torquato, Phys. Rev. B 88(13) (2013) p.134104.

[67] Y. Dorsett and T. Tuschl, Nat. Rev. Drug Discov. 3(4) (2004) p.318.

[68] R.M. Dowben and J.E. Rose, Science 118(3053) (1953) p.22.

[69] B. Dunn and Y. Roudi, Phys. Rev. E 87(2) (2013) p.022127.

[70] T.P. Eggarter, Phys. Rev. B 9 (1974) p.2989.

[71] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt and E. Aurell, Phys. Rev. E 87(1) (2013) p.012707.

[72] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning*, Cambridge University Press, Cambridge, UK, 2001.

[73] K. Faust, J.F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower, PLoS Comput. Biol. 8(7) (2012) p.e1002606.

[74] A.L. Ferguson, J.K. Mann, S. Omarjee, T. Ndung'u, B.D. Walker and A.K. Chakraborty, Immunity 38(3) (2013) p.606.

[75] U. Ferrari, Phys. Rev. E 94(2) (2016) p.023301.

[76] A. Fischer and C. Igel, *An introduction to restricted Boltzmann machines*, in *Iberoamerican Congress on Pattern Recognition*, Springer, Heidelberg, 2012, p. 14.

[77] A. Fischer, I. Vázquez-García and V. Mustonen, Proc. Natl. Acad. Sci. USA 112(4) (2015) p.1007.

[78] C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler and M. Zehl, RNA 7(02) (2001) p.254.

[79] N. Friedman, Science 303(5659) (2004) p.799.

[80] X. Gabaix, P. Gopikrishnan, V. Plerou and H.E. Stanley, Nature 423(6937) (2003) p.267.

[81] R.G. Gallager, IRE Trans. Inf. Theory 8(1) (1962) p.21.

[82] E. Ganmor, R. Segev and E. Schneidman, Proc. Natl. Acad. Sci. USA 108(23) (2011) p.9679.

[83] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer, Heidelberg, 1985.

[84] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein and P.O. Brown, Mol. Biol. Cell 11(12) (2000) p.4241.

[85] A. Georges and J.S. Yedidia, J. Phys. A: Math. Gen. 24 (1991) p.2173.

[86] R.J. Glauber, J. Math. Phys. 4(2) (1963) p.294.

[87] U. Göbel, C. Sander, R. Schneider and A. Valencia, Proteins: Struct. Funct. Bioinform. 18(4) (1994) p.309.

[88] E. Granot-Atedgi, G. Tkačik, R. Segev and E. Schneidman, PLoS Comput. Biol 9(3) (2013) p.e1002922.

[89] G. Guennebaud, B. Jacob, et al., Eigen v3. preprint (2010). Available at http://eigen.tuxfamily.org

[90] S.F. Gull and J. Skilling, IEE Proc. F Commun. Radar Signal Process 131(6) (1984) p.646.

[91] M. Habeck, Phys. Rev. E 89 (2014) p.052113.

[92] N. Halabi, O. Rivoire, S. Leibler and R. Ranganathan, Cell 138(4) (2009) p.774.

[93] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, Heidelberg, 2009.

[94] D.R. Hekstra, S. Cocco, R. Monasson and S. Leibler, Phys. Rev. E 88(6) (2013) p.062714.

[95] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Vol. 1, Basic Books, New York, 1991.

[96] G.J. Hickman and T.C. Hodgman, J. Bioinform. Comput. Biol. 7(06) (2009) p.1013.

[97] G.E. Hinton and R.R. Salakhutdinov, Science 313(5786) (2006) p.504.

[98] J.D. Hoheisel, Nat. Rev. Genet. 7(3) (2006) p.200.

[99] T.A. Hopf, L.J. Colwell, R. Sheridan, B. Rost, C. Sander and D.S. Marks, Cell 149(7) (2012) p.1607.

[100] T.A. Hopf, J.B. Ingraham, F.J. Poelwijk, M. Springer, C. Sander and D.S. Marks, Quantification of the effect of mutations using a global probability model of natural sequence variation, preprint (2015). Available at arXiv:1510.04612.

[101] A. Hyvärinen, J. Mach. Learn. Res. 6(April) (2005) p.695.

[102] A. Hyvärinen, Neural Comput. 18(10) (2006) p.2283.

[103] T.E. Ideker, V. Thorsson and R.M. Karp, *Discovery of regulatory interactions through perturbation: inference and experimental design*, in *Pacific Symposium on Biocomputing*, Vol. 5, Hawaii, 2000, p. 302.

[104] R. Islam, R. Ma, P.M. Preiss, M.E. Tai, A. Lukin, M. Rispoli and M. Greiner, Nature 528(7580) (2015) p.77.

[105] H. Jacquin and A. Rançon, Phys. Rev. E 94 (2016) p.042118.

[106] C. Jarzynski, Phys. Rev. Lett. 78(14) (1997) p.2690.

[107] E.T. Jaynes, Phys. Rev. 106(4) (1957) p.620.

[108] E.T. Jaynes, *E.T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, Springer, Heidelberg, 1989.

[109] D.T. Jones, T. Singh, T. Kosciolek and S. Tetchner, Bioinformatics 31(7) (2015) p.999.

[110] J.D. Kalbfleisch, *Pseudo-Likelihood*, Wiley, Hoboken, NJ, 2005.

[111] H.J. Kappen, J. Marro, P.L. Garrido, and J. Torres, *AIP Conference Proceedings*, AIP Publishing, New York, 2007.

[112] H.J. Kappen and F. Rodríguez, Adv. Neural Inf. Process. Syst. (1998) p.280.

[113] H.J. Kappen and J.J. Spanjers, Phys. Rev. E 61(5) (2000) p.5658.

[114] R. Kikuchi, Phys. Rev. 81 (1951) p.988.

[115] M.A. Kohanski, D.J. Dwyer and J.J. Collins, Nat. Rev. Microbiol. 8(6) (2010) p.423.

[116] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, MA, 2009.

[117] V. Kolmogorov, IEEE Trans. Pattern Anal. Mach. Intell. 28(10) (2006) p.1568.

[118] P.L. Krapivsky, S. Redner and E. Ben-Naim, *A Kinetic View of Statistical Physics*, Cambridge University Press, Cambridge, 2010.

[119] J. Krumsiek, K. Suhre, T. Illig, J. Adamski and F.J. Theis, BMC Syst. Biol. 5(1) (2011) p.21.

[120] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard and D. Baker, Science 302(5649) (2003) p.1364.

[121] J.E. Kulkarni and L. Paninski, Network: Comput. Neural Syst. 18(4) (2007) p.375.

[122] W. Kunkin and H.L. Frisch, Phys. Rev. 177(1) (1969) p.282.

[123] Y. LeCun, Y. Bengio and G. Hinton, Nature 521(7553) (2015) p.436.

[124] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, Proc. IEEE 86(11) (1998) p.2278.

[125] M. Levin and X.-G. Wen, Phys. Rev. Lett. 96 (2006) p.110405.

[126] T.R. Lezon, J.R. Banavar, M. Cieplak, A. Maritan and N.V. Fedoroff, Proc. Natl. Acad. Sci. USA 103(50) (2006) p.19033.

[127] Y.-Y. Liu, Y. Wang, T.R. Walsh, L.-X. Yi, R. Zhang, J. Spencer, Y. Doi, G. Tian, B. Dong and X. Huang, et al., Lancet Infec. Dis. 16(2) (2016) p.161.

[128] J.W. Locasale and A. Wolf-Yadlin, PLoS ONE 4(8) (2009) p.e6522.

[129] S.W. Lockless and R. Ranganathan, Science 286(5438) (1999) p.295.

[130] A.Y. Lokhov, M. Vuffray, S. Misra and M. Chertkov, Optimal structure and parameter learning of Ising models, preprint (2016). Available at arXiv:1612.05024.

[131] D.J. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, 2003.

[132] J.H. Macke, P. Berens, A.S. Ecker, A.S. Tolias and M. Bethge, Neural Comput. 21(2) (2009) p.397.

[133] J.H. Macke, M. Opper and M. Bethge, Phys. Rev. Lett. 106(20) (2011) p.208102.

[134] E.Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A.R. Bialas, N. Kamitaki, E.M. Martersteck, J.J. Trombetta, D.A. Weitz, J.R. Sanes, A.K. Shalek, A. Regev and S.A. McCarroll, Cell 161(5) (2015) p.1202.

[135] J.K. Mann, J.P. Barton, A.L. Ferguson, S. Omarjee, B.D. Walker, A. Chakraborty and T. Ndung'u, PloS ONE 10(8) (2014) p.e1003776.

[136] É. Marcotte, R.A. DiStasio Jr, F.H. Stillinger and S. Torquato, Phys. Rev. B 88(18) (2013) p.184432.

[137] E. Marinari and V. Van Kerrebroeck, J. Stat. Mech. Theory Exp. 2010(02) (2010) p.P02008.

[138] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina and C. Sander, PloS ONE 6(12) (2011) p.e28766.

[139] O. Marre, S. El Boustani, Y. Frégnac and A. Destexhe, Phys. Rev. Lett. 102(13) (2009) p.138101.

[140] M. Marsili, I. Mastromatteo and Y. Roudi, J. Stat. Mech. Theory Exp. 2013(09) (2013) p.P09003.

[141] I. Mastromatteo and M. Marsili, J. Stat. Mech. Theory Exp. 2011(10) (2011) p.P10012.

[142] W. Maysenhölder, Phys. Status Solidi B 139 (1987) p.399.

[143] L. Merchan and I. Nemenman, J. Stat. Phys. 162(5) (2016) p.1294.

[144] M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, Oxford, 2009.

[145] M. Mézard and T. Mora, J. Physiol. Paris 103(1–2) (2009) p.107.

[146] M. Mézard, G. Parisi and M.A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, Singapore, 1987.

[147] M. Mézard and J. Sakellariou, J. Stat. Mech. (2011) p.L07001.

[148] E.J. Molinelli, A. Korkut, W. Wang, M.L. Miller, N.P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, D.B. Solit, C.A. Pratilas, M. Weigt, A. Braunstein, A. Pagnani, R. Zecchina and C. Sander, PLoS Comput. Biol. 9(12) (2013) p.e1003290.

[149] A. Montanari and J.A. Pereira, *Which graphical models are difficult to learn?*, in *Advances in Neural Information Processing Systems, Vancouver*, D. Koller, ed., Curran Associates, Inc., Red Hook, NY, 2009, p.1303.

[150] C. Moore and S. Mertens, *The Nature of Computation*, Oxford University Press, Oxford, 2011.

[151] T. Mora and W. Bialek, J. Stat. Phys. 144(2) (2011) p.268.

[152] T. Mora, S. Deny and O. Marre, Phys. Rev. Lett. 114(7) (2015) p.078105.

[153] T. Mora, A.M. Walczak, W. Bialek and C.G. Callan, Proc. Natl. Acad. Sci. USA 107(12) (2010) p.5405.

[154] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J. Onuchic, T. Hwa and M. Weigt, Proc. Natl. Acad. Sci. USA 108(49) (2011) p.E1293.

[155] J.J. Moré, B.S. Garbow and K.E. Hillstrom, User Guide for MINPACK-1, ANL-80-74, Argonne National Laboratory, 1980.

[156] A. Mozeika, O. Dikmen and J. Piili, Phys. Rev. E 90(1) (2014) p.010101.

[157] I.J. Myung, V. Balasubramanian and M.A. Pitt, Proc. Natl. Acad. Sci. USA 97(21) (2000) p.11170.

[158] H. Nasser and B. Cessac, Entropy 16(4) (2014) p.2244.

[159] S. Nelander, W. Wang, B. Nilsson, Q.-B. She, C. Pratilas, N. Rosen, P. Gennemark and C. Sander, Mol. Syst. Biol. 4(1) (2008) p. 216.

[160] H.C. Nguyen and J. Berg, J. Stat. Mech. (2012) p.P03004.

[161] H.C. Nguyen and J. Berg, Phys. Rev. Lett. 109(5) (2012) p.050602.

[162] M.E.J. Obien, K. Deligkaris, T. Bullmann, D.J. Bakkum and U. Frey, Front. Neurosci. 8 (2014).

[163] I.E. Ohiorhenuan, F. Mechler, K.P. Purpura, A.M. Schmid, Q. Hu and J.D. Victor, Nature 466(7306) (2010) p.617.

[164] M. Opper and D. Saad (ed.), *Advanced Mean-Field Methods: Theory and Practice*, The MIT Press, Cambridge, MA, 2001.

[165] J.D. Orth, I. Thiele and B.Ø. Palsson, Nat. Biotechnol. 28(3) (2010) p.245.

[166] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G.A. Pavlopoulos, D.E. Kim, H. Kamisetty, N.C. Kyrpides and D. Baker, Science 355(6322) (2017) p.294.

[167] L. Paninski, J. Comput. Neurosci. 21(1) (2006) p.71.

[168] L. Paninski, J.W. Pillow and E.P. Simoncelli, Neural Comput. 16(12) (2004) p.2533.

[169] G. Parisi and F. Slanina, J. Stat. Mech. Theory Exp. (2006) p.L02003.

[170] J.M. Parrondo, J.M. Horowitz and T. Sagawa, Nat. Phys. 11(2) (2015) p.131.

[171] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, Burlington, 1988.

[172] J. Pearl, Biometrika 82(4) (1995) p.669.

[173] D. Pe'er, A. Regev, G. Elidan and N. Friedman, Bioinformatics 17(suppl 1) (2001) p.S215.

[174] R. Peierls, *On Ising's model of ferromagnetism*, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 32, Cambridge University Press, 1936, p. 477.

[175] P. Peretto, Biol. Cybern. 50(1) (1984) p.51.

[176] C. Peterson and J. Anderson, Complex Syst. 1 (1987) p.995.

[177] P. Picotti, M. Clément-Ziza, H. Lam, D.S. Campbell, A. Schmidt, E.W. Deutsch, H. Röst, Z. Sun, O. Rinner, L. Reiter, M.J. Shen, Q.A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, M. RL, A. Beyer and R. Aebersold, Nature 494(7436) (2013) p.266.

[178] J.W. Pillow, J. Shlens, L. Paninski, A. Sher, A.M. Litke, E. Chichilnisky and E.P. Simoncelli, Nature 454(7207) (2008) p.995.

[179] T. Plefka, J. Phys. A: Math. Gen. 15 (1982) p.1971.

[180] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipies in $C++$*, Cambridge University Press, Cambridge, 2002.

[181] S. Pressé, K. Ghosh, J. Lee and K.A. Dill, Rev. Mod. Phys. 85(3) (2013) p.1115.

[182] P. Ravikumar, M.J. Wainwright and J.D. Lafferty, Ann. Stat. 38(3) (2010) p.1287.

[183] J.M. Rebesco, I.H. Stevenson, K. Koerding, S.A. Solla and L.E. Miller, Front. Syst. Neurosci. 4 (2010) p.39.

[184] M. Rechtsman, F. Stillinger and S. Torquato, Phys. Rev. E 73(1) (2006) p.011406.

[185] F. Ricci-Tersenghi, J. Stat. Mech. (2012) p.P08015.

[186] M.V. Rockman, Nature 456(7223) (2008) p.738.

[187] Y. Roudi, E. Aurell and J.A. Hertz, Front. Comput. Neurosci. 3 (2009) p.22.

[188] Y. Roudi and J. Hertz, J. Stat. Mech. Theory Exp. 2011(03) (2011) p.P03031.

[189] Y. Roudi and J. Hertz, Phys. Rev. Lett. 106(4) (2011) p.048702.

[190] Y. Roudi, S. Nirenberg and P.E. Latham, PLoS Comput. Biol. 5(5) (2009) p.e1000380.

[191] Y. Roudi and G. Taylor, Curr. Opin. Neurobiol. 35 (2015) p.110.

[192] Y. Roudi, J. Tyrcha and J. Hertz, Phys. Rev. E 79(5) (2009) p.051915.

[193] N.P. Santhanam and M.J. Wainwright, IEEE Trans. Inf. Theory 58(7) (2012) p.4117.

[194] M. Santolini, T. Mora and V. Hakim, PLoS ONE 9(6) (2014) p.e99015.

[195] E. Schneidman, M.J. Berry, R. Segev and W. Bialek, Nature 440(7087) (2006) p.1007.

[196] D.J. Schwab, I. Nemenman and P. Mehta, Phys. Rev. Lett. 113(6) (2014) p.068102.

[197] D.A. Schwarz, M.A. Lebedev, T.L. Hanson, D.F. Dimitrov, G. Lehew, J. Meloy, S. Rajangam, V. Subramanian, P.J. Ifft, Z. Li, A. Ramakrishnan, A. Tate, K.Z. Zhuang and M.A.L. Nicolelis, Nat. Methods 11(6) (2014) p.670.

[198] G. Schwarz, et al., Ann. Stat. 6(2) (1978) p.461.

[199] G. Sella and A.E. Hirsh, Proc. Natl. Acad. Sci. USA 102(27) (2005) p.9541.

[200] V. Sessak and R. Monasson, J. Phys. A: Math. Theor. 42(5) (2009) p.055001.

[201] K. Shekhar, C.F. Ruberman, A.L. Ferguson, J.P. Barton, M. Kardar and A.K. Chakraborty, Phys. Rev. E 88(6) (2013) p.062705.

[202] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. 35(26) (1975) p.1792.

[203] J. Shlens, G.D. Field, J.L. Gauthier, M.I. Grivich, D. Petrusca, A. Sher, A.M. Litke and E.J. Chichilnisky, J. Neurosci. 26(32) (2006) p.8254.

[204] C. Sima, J. Hua and S. Jung, Curr. Genomics 10(6) (2009) p.416.

[205] N. Slonim, G.S. Atwal, G. Tkačik and W. Bialek, Proc. Natl. Acad. Sci. USA 102(51) (2005) p.18297.

[206] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner and R. Ranganathan, Nature 437(7058) (2005) p.512.

[207] J. Sohl-Dickstein, P.B. Battaglino and M.R. DeWeese, Phys. Rev. Lett. 107(22) (2011) p.220601.

[208] M.E. Spira and A. Hai, Nat. Nanotechnol. 8(2) (2013) p.83.

[209] H.E. Stanley, *Introduction to Phase Transitions and Critical Phenomena*, Oxford University Press, Oxford, 1987.

[210] R.R. Stein, D.S. Marks and C. Sander, PLoS Comput. Biol. 11(7) (2015) p.e1004182.

[211] G.J. Stephens, T. Mora, G. Tkačik and W. Bialek, Phys. Rev. Lett. 110(1) (2013) p.018701.

[212] J.I. Sułkowska, F. Morcos, M. Weigt, T. Hwa and J.N. Onuchic, Proc. Natl. Acad. Sci. USA 109(26) (2012) p.10340.

[213] R.H. Swendsen, Phys. Rev. Lett. 52 (1984) p.1165.

[214] T. Tanaka, Phys. Rev. E 58(2) (1998) p.2302.

[215] T. Tanaka, Neural Comput. 12(8) (2000) p.1951.

[216] A. Tang, D. Jackson, J. Hobbs, W. Chen, J.L. Smith, H. Patel, A. Prieto, D. Petrusca, M.I. Grivich, A. Sher, P. Hottowy, W. Dabrowski, A.M. Litke and J.M. Beggs, J. Neurosci. 28(2) (2008) p.505.

[217] J. Tegner, M.K.S. Yeung, J. Hasty and J.J. Collins, Proc. Natl. Acad. Sci. USA 100(10) (2003) p.5944.

[218] D.J. Thouless, P.W. Anderson and R.G. Palmer, Philos. Mag. 35 (1977) p.593.

[219] Y. Tikochinsky, N.Z. Tishby and R.D. Levine, Phys. Rev. A 30(5) (1984) p.2638.

[220] D. Tischer and O.D. Weiner, Nat. Rev. Mol. Cell Biol. 15(8) (2014) p.551.

[221] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek and M.J. Berry II, PLoS Comput. Biol. 10(1) (2014) p.e1003408.

[222] G. Tkačik, O. Marre, T. Mora, D. Amodei, M.J. Berry II and W. Bialek, J. Stat. Mech. Theory Exp. 2013(03) (2013) p.P03011.

[223] G. Tkačik, T. Mora, O. Marre, D. Amodei, S.E. Palmer, M.J. Berry and W. Bialek, Proc. Natl. Acad. Sci. USA (2015) p.201514188.

[224] G. Tkačik, J.S. Prentice, V. Balasubramanian and E. Schneidman, Proc. Natl. Acad. Sci. USA 107(32) (2010) p.14419.

[225] G. Tkačik, E. Schneidman, I. Berry, J. Michael and W. Bialek, Spin glass models for a network of real neurons, preprint (2009). Available at arXiv:0912.5409.

[226] S. Torquato, Soft Matter 5(6) (2009) p.1157.

[227] T. Toyoizumi, K.R. Rad and L. Paninski, Neural Comput. 21(5) (2009) p.1203.

[228] W. Truccolo, U.T. Eden, M.R. Fellows, J.P. Donoghue and E.N. Brown, J. Neurophysiol. 93(2) (2005) p.1074.

[229] J. Tyrcha, Y. Roudi, M. Marsili and J. Hertz, J. Stat. Mech. Theory Exp. 2013(03) (2013) p.P03005.

[230] E. van Nimwegen, PLoS Comput. Biol. 12(5) (2016) p.e1004726.

[231] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, J. Mach. Learn. Res. 11(December) (2010) p.3371.

[232] M. Vuffray, S. Misra, A.Y. Lokhov and M. Chertkov, Adv. Neural Inf. Process. Syst. (2016) p.2595.

[233] C. Walsh, Nature 406(6797) (2000) p.775.

[234] J.-S. Wang, Physica A 281(1) (2000) p.147.

[235] S. Wang, S. Sun, Z. Li, R. Zhang and J. Xu, PLOS Comput. Biol. 13(1) (2017) p.e1005324.

[236] Z. Wang, M. Gerstein and M. Snyder, Nat. Rev. Genet. 10(1) (2009) p.57.

[237] T.L.H. Watkin, A. Rau and M. Biehl, Rev. Mod. Phys. 65(2) (1993) p.499.

[238] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch and T. Hwa, Proc. Natl. Acad. Sci. USA 106(1) (2009) p.67.

[239] M. Welling and Y.W. Teh, Artif. Intell. 143(1) (2003) p.19.

[240] M. Welling and Y.W. Teh, Neural Comput. 16 (2004) p.197.

[241] Q.F. Wills, K.J. Livak, A.J. Tipping, T. Enver, A.J. Goldson, D.W. Sexton and C. Holmes, Nat. Biotechnol. 31(8) (2013) p.748.

[242] R. Wise, T. Hart, O. Cars, M. Streulens, R. Helmuth, P. Huovinen and M. Sprenger, Br. Med. J. 317(7159) (1998) p.609.

[243] K. Wood, S. Nishida, E.D. Sontag and P. Cluzel, Proc. Natl. Acad. Sci. USA 109(30) (2012) p.12254.

[244] J.S. Yedidia, *An idiosyncratic journey beyond mean field theory*, in *Advanced Mean-Field Methods: Theory and Practice*, M. Opper and D. Saad, eds., The MIT Press, Cambridge, MA, 2001, p. 21.

[245] J.S. Yedidia, W. Freeman and Y. Weiss, IEEE Trans. Inf. Theory 51(7) (2005) p.2282.

[246] J.S. Yedidia, W.T. Freeman and Y. Weiss, Adv. Neural Inf. Process. Syst. 13 (2001).

[247] L. Zdeborová and F. Krzakala, Adv. Phys. 65(5) (2016) p.453.

[248] A. Zeisel, A.B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler and S. Linnarsson, Science 347(6226) (2015) p.1138.

[249] H.-L. Zeng, M. Alava, E. Aurell, J. Hertz and Y. Roudi, Phys. Rev. Lett. 110(21) (2013) p.210601.

[250] H.-L. Zeng, E. Aurell, M. Alava and H. Mahmoudi, Phys. Rev. E 83(4) (2011) p.041135.

[251] B. Zhang and P.G. Wolynes, Proc. Natl. Acad. Sci. USA 112(19) (2015) p.6062.

[252] G. Zhang, F. Stillinger and S. Torquato, Phys. Rev. E 88(4) (2013) p.042309.

[253] J. Zhu, B. Zhang, E.N. Smith, B. Drees, R.B. Brem, L. Kruglyak, R.E. Bumgarner and E.E. Schadt, Nat. Genet. 40(7) (2008) p.854.