A statistical physics approach to inference problems on random networks: Ising and kinetic Ising models

vorgelegt von
Ludovica Bachschmid Romano
geboren in Milano, Italien

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr.-rer.-nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Benjamin Blankertz
Gutachter:	Prof. Dr. Manfred Opper
Gutachter:	Prof. Dr. Johannes Berg
Gutachter:	Prof. Dr. David Saad

Tag der wissenschaftlichen Aussprache: 15. Dezember 2017

Berlin, 2018

Abstract

Recent advances in measurement technologies have resulted in the availability of large datasets from a variety of fields spanning the natural and social sciences. This posed the challenge to develop new statistical tools to extract relevant information from the data. A paradigmatic model that has been successfully applied to analyze large datasets is the Ising model of binary spins interacting through pairwise connections. In this thesis, we use methods of statistical physics to tackle several open problems related to modelling the stochastic dynamics of the Ising model and reconstructing the unknown network of interactions from data. First, we derive a novel mean-field solution to the discrete time parallel dynamics of the Ising model, based on a weak coupling expansion of the log-generating function with constrained first and second order moments over time, the result of which outperforms other mean field techniques in predicting single site magnetization. Next, for both the equilibrium and kinetic models, we analyze the inverse problem of learning the couplings between the variables based on a set of observations on spin configurations. Using the cavity and replica methods of statistical physics, we compare the performance of different inference algorithms, by analytical computation of the estimation error as a function of the size of the dataset, and study its deviation from asymptotic optimality. We also derive optimal algorithms for learning the couplings. Finally, we consider the case where a subset of the spin trajectories is observed while the rest are hidden. This enabled us to model systems where only a finite fraction of the system is experimentally accessible, but allowed the hidden variables to affect the dynamics of the observed variables. A central question is the prediction of the hidden spin state when the couplings are known. For the average case scenario, we investigate the theoretically optimal performance for predicting hidden spins by computing the error of the Bayes optimal predictor. We also derive a mean-field formalism to accurately estimate the single-site magnetisation of hidden spins for single instances of the network.

Zusammenfassung

Die verfügbare Datenmenge in Gebieten der Natur- und Sozialwissenschaften wächst stetig durch den technischen Fortschritt bei Methoden zur Datenerhebung. Dieser Zuwachs ist eine Herausforderung für Algorithmen, die Daten auf relevanten Informationen reduziert. Ein paradigmatisches Modell, das mit Erfolg auf groen Datenmengen angewendet wurde, ist das "Ising Modell" für binäre Spins, welche durch paarweise Wechselbeziehungen voneinander abhängig sind. In dieser Arbeit verwenden wir Methodik der statistischen Physik zur Lösung verschiedener offener Probleme, verbunden mit der Modellierung stochastischer Dynamik und der Rekonstruktion unbekannter Netzwerke durch Interaktionen, welche in Daten beobachtet werden. Als erstes leiten wir eine neue "Mean-field" Lösung her für die zeitdiskrete parallele Dynamik des Ising Modells. Hierzu entwickeln wir die logarithmische Moment generierende Funktion in den schwachen Kopplungen bedingt auf ersten und zweiten Momenten an jedem Zeitpunkt. Weiterhin untersuchen wir das inverse Ising Problem für sowohl das Gleichgewichts- als auch das kinetische Modell. Das heit, wir analysieren das Lernen der Kopplungen zwischen Variablen gegeben ein Set von Beobachtungen. Durch den Gebrauch von "Cavity-" und "Replikamethoden" aus der statistischen Physik vergleichen wir verschieden Inferenzalgorithmen durch analytische Berechnung des Schätzfehlers abhängig von der Menge der verfügbaren Daten und untersuchen die Abweichungen dieser von der optimalen Asymptotik. Auerdem leiten wir optimale Algorithmen zum Lernen der Kopplungen her. Am Ende betrachten wir den Fall, in dem ein Teil der Spintrajektorien bekannt ist, während ein Teil nicht beobachtet wird. Dies erlaubt uns Systeme zu modellieren, die uns experimentell nur unvollständig zugänglich sind, unter Berücksichtigung, dass die unbeobachteten Variablen die beobachteten Dynamiken beeinflussen können. Von zentraler Bedeutung ist die Vorhersage des Zustands der nicht beobachteten Spins, wenn die Kopplungen bekannt sind. Für den gemittelten Fall untersuchen wir das theoretisch optimale Ergebnis zur Vorhersage der nicht beobachteten Spins durch berechnen des Fehlers eines Bayes optimalen Prädiktors. Wir leiten einen "Mean-field" Formalismus für einzelne Instanzen von Netzwerken her, um die marginale Magnetisierung der nicht beobachteten Spins präzise zu schätzen.

Aknowledgements

I would like to express my gratitude to Prof. Manfred Opper for his supportive and patient supervision and for the countless things that I learned from him through our discussions. His insights and contagious enthusiasm for research have been a constant source of motivation over the years.

I gratefully acknowledge the Marie Curie Initial Training Network NETADIS for their funding and all the professors and students involved in the project: from lectures to informal chats at schools and conferences, the project provided extremely enriching experiences both from a scientific and a personal point of view. I wish to thank Prof. Peter Sollich for coordinating the project and for his attentive guidance during my secondment at King's College London; Prof. Yasser Roudi for hosting me at the Kavli Institute for Neuroscience at NTNU, for his incisive suggestions and encouragement; Prof. Andrea Pagnani for the inspiring discussions I had with him. Pascale Searle provided invaluable help in managing the project. Special thanks to Barbara and Claudia for their fruitful cooperation and for warmly welcoming me in London and Trondheim respectively; to Silvia, Barbara, and Carla for all the stimulating conversations, which began in discussions of physics and ended in a precious friendship.

I would also like to thank the members of my defense committee, Prof. Johannes Berg and Prof. David Saad, for carefully examining my thesis and providing constructive comments.

I am grateful to all the members of the KI group for creating an enjoyable work environment – especially to my office mates Christian and Noa for making room 4.014 a fun and relaxed space, to Philipp for sharing his passion for music, and to Cordula for helping me with all the administrative tasks.

Last but not least, I would like to extend my gratitude to my family for giving me unconditional love and support and to all my friends scattered over the globe for making my doctoral years so meaningful.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 The Ising model and the kinetic Ising model . . . 3
  - 1.2.1 The Ising model . . . 3
  - 1.2.2 Explaining correlations in neural spike trains . . . 5
  - 1.2.3 Glauber dynamics . . . 6
  - 1.2.4 Inferring effective connectivities in neuronal networks . 8

2 Thesis Outline **11**

3 Mean field approaches to dynamics on random networks: the kinetic Ising model **15**
- 3.1 Introduction . . . 15
- 3.2 Paper 1. . . . 18
- 3.3 Discussion . . . 53
- 3.4 Conclusions . . . 54

Appendices **59**
- Appendix 3.A TAP equations for the SK model . . . 59
  - 3.A.1 The cavity approach . . . 59
  - 3.A.2 Plefka's expansion . . . 61
- Appendix 3.B Mean field approaches to the kinetic Ising model: previous results . . . 63
- Appendix 3.C Algorithm with averaged moments . . . 67

4 Learning in kinetic Ising models **69**
- 4.1 Introduction . . . 69
- 4.2 Paper 2. . . . 71
- 4.3 Further results . . . 92
  - 4.3.1 The optimal linear estimator . . . 92
  - 4.3.2 Bayesian inference . . . 93
  - 4.3.3 Approximating the posterior by cavity arguments . . . 95
  - 4.3.4 A simple expectation propagation algorithm . . . 96
  - 4.3.5 Average case: a replica analysis . . . 99

4.3.6 Results . . . 101
4.4 Conclusions . . . 102

Appendices **105**
Appendix 4.A The replica method: from spin glasses to neural networks . . . 105
4.A.1 Spin glasses and the replica trick . . . 105
4.A.2 Statistical mechanics of learning: general setup . . . 108
Appendix 4.B Maximum likelihood estimator . . . 112
Appendix 4.C Mean field estimators for the stationary state . . . 113
Appendix 4.D Mean field estimators for the transient dynamics . . . 114
Appendix 4.E Expectation Propagation algorithm: generating function of the moments . . . 115
Appendix 4.F Fixed point of the Expectation Propagation algorithm 116
Appendix 4.G Complete Expectation Propagation . . . 117
Appendix 4.H Details of the replica calculation . . . 118

5 Learning Curves for the inverse Ising problem **121**
5.1 Introduction . . . 121
5.2 Paper 3. . . . 123
5.3 Further results . . . 152
5.4 Conclusions . . . 153

Appendices **157**
Appendix 5.A The likelihood and pseudo-likelihood functions . . . 157

6 Learning and inference in presence of hidden units **159**
6.1 Introduction . . . 159
6.2 Paper 4. . . . 161
6.3 Further results . . . 178
6.3.1 Inferring hidden states in a kinetic Ising model via the extended Plefka expansion . . . 178
6.3.2 Network reconstruction . . . 183
6.4 Conclusions . . . 187

Appendices **191**
Appendix 6.A Details of the extended Plefka expansion . . . 191
Appendix 6.B Sparse observations . . . 196

7 Summary and outlook **199**

1 Introduction

1.1 Motivation

Due to recent technological advances in data acquisition, the last two decades have witnessed a rapid increase in the amount and richness of data that can be collected in many fields of natural sciences, finance, social sciences, and communication networks. This has shifted the focus from the analysis of single system components to the attempt to understand the system as a whole. Remarkable examples can be found in biology, where the advent of multi-component recordings opened up entire new lines of research, such as genomics, transcriptomics, proteomics, metabolomics, and the analysis of simultaneous measurements of many neurons. As common feature, these large datasets encode the activity of large systems of many interacting units, where the direct connections between them are not directly measurable.

To understand how the system operates, it is crucial to reconstruct such underlying networks of interactions. The *inverse problem* of reconstructing the network starting from the empirical knowledge of certain observables, such as averages and correlations, has raised a lot of interest within the statistics, machine-learning, and statistical physics communities. The main challenge is to disentangle direct connections from mere correlations that could emerge from indirect influence of intermediate components. The task is even more difficult since the system is typically not entirely accessible, and the data are noisy.

Inverse problems generally have no unique solution but can be tackled in a probabilistic formulation, where the system is modelled by a parametrised distribution; parameters are then inferred, either maximising the likelihood of the data or using Bayesian estimators (where further a priori information on the parameters is incorporated through a prior). However, exact inference requires the computation of high dimensional integrals and is intractable for most models of interest; hence, a lot of effort has been made to derive efficient approximate techniques for this task.

The field of statistical mechanics, whose focus is to study large systems of interacting particles and to unveil the relations between microscopic interactions and macroscopic observables, provides a whole series of techniques that can be

used both to construct inference algorithms based on suitable approximation schemes and to theoretically assess the algorithms' performance.

A model of statistical physics widely used for network reconstruction is the Ising model, describing an equilibrium system of binary variables (spins) interacting via pairwise connections. Despite being an obvious over-simplification of real-world systems, it can be used very effectively to disentangle direct interactions among the variables from spurious coupling effects. However, the assumption that the system is at equilibrium -which requires that the connections are symmetric- is unrealistic for many applications.

More recently, a simple extension to dynamics of the Ising model (hereafter denoted as the kinetic Ising model) has been employed for reconstructing biological, financial, and gene regulatory networks, where the equilibrium assumption is relaxed. Moreover, in the case of time-series data, the temporal structure encoded in the dataset can be exploited.

While the equilibrium Ising model has been extensively studied in the last decades and a wide body of literature has been devoted to develop approximate algorithms for the inverse Ising problem, the attention to its out-of-equilibrium dynamics is more recent, and even the problem of relating the system parameters to the time evolution of the observables is not yet fully solved.

In this context, the contribution of this thesis is towards both a theoretical analysis and the construction of new inference algorithms. First, for the kinetic Ising model, we will study the forward problem of predicting the time evolution of the single spin magnetization for fixed connections between the spins. We will focus on densely and weakly connected networks, for which an exact mean-field theory can be formulated in the thermodynamic limit of a large system. In contrast to the equilibrium case, the variables of interest are not single spins but entire spin trajectories, and the goal is to compute the marginal distribution of single-spin trajectories. The exact mean-field solution has so far been found only in the case of fully asymmetric interactions [MS11], in which two-times correlations are negligible. In the case of generic degree of symmetry of the couplings, a recent approach [MS14] which incorporates the effect of time correlations has improved on the prediction of single site magnetizations; still, it was not clear how the exact mean-field solution would look like in the thermodynamic limit. We will derive a novel approximate technique of the mean-field type to tackle the problem in the limit of an infinitely large system.

The second focus of the dissertation concerns the theoretical performance of inference algorithms, which estimate the couplings between the spins based on a set of observations of spin configurations. Various algorithms based on approximate schemes have been proposed for the inverse Ising problem, and more recently for the inverse kinetic Ising problem, with the goal of providing

computationally efficient inference tools for large networks [NZB17a]. While their performance has been only tested on specific instances of the problem, it is important to assess their statistical efficiency in a unified theoretical setting. Both for the kinetic and for the equilibrium case, we analytically compute the typical performance of various estimators of the couplings and provide new algorithmic implementations of the most efficient ones.

As a last point, we observe that in most biological networks, often only a small fraction of the system is experimentally accessible. Variables whose activity is recorded will also interact with variables not directly observable or detectable, usually referred to as hidden variables. Hence, recent works [TH13, DR13, Hua15, BHTR15, RT15, DB16] have introduced a model where a fraction of the spin trajectories are observed and a fraction are hidden. Network reconstruction is much harder in this scenario, and no satisfactory solution has been found in dense networks if the hidden variables are connected among themselves. Exact learning rules imply the summation over all possible configuration of hidden spins, which is intractable for large systems; one can resort to learning rules that are based on an approximate estimation of hidden nodes at fixed couplings, but the accuracy of inferring the state of the unobserved variables for given system parameters was not clear. In the last part of the dissertation, we address this problem by investigating the theoretically optimal performance for predicting the hidden spins. We will also introduce a novel technique to predict the single site magnetization of hidden spins for single instances of the network.

1.2 The Ising model and the kinetic Ising model

After introducing the Ising model and its simplest generalization to dynamics, we will show how they can be used to model the dependencies of spikes recorded from ensembles of neurons. Other application domains include determining the 3D structure of proteins [WWS$^+$09], analyzing gene expression data from gene regulatory networks [LBC$^+$06], inferring the fitness landscape in a evolutionary biology web of ecological interaction between species [BCG$^+$12]. A recent review of those and other methods can be found in [NZB17a].

1.2.1 The Ising model

The Ising model was introduced to study the macroscopic properties of magnetic materials [Hua87,LL69], many-body systems composed of molecules with a *magnetic moment* - a vector which tends to align with the magnetic field acting on the molecule. Magnetic moments of individual molecules are described

by binary variables, $\sigma_i = \pm 1$ (or Ising spins), localized at the vertices of a lattice. To each pair of variables at sites i, j we assign an interaction energy with value $-J_{ij}$ if the spins σ_i and σ_j are pointing in the same direction ($\sigma_i = \sigma_j$) and with value J_{ij} otherwise ($\sigma_i = -\sigma_j$). In some cases, each site i also has its own energy $-\sigma_i h_i$, due to the presence of a (local) external field h_i. The energy of the many-particle system, or Hamiltonian, is

$$H = -\sum_{i,j \in B} J_{ij}\sigma_i\sigma_j - \sum_{i=i}^{N} h_i\sigma_i, \tag{1.1}$$

where N is the total number of spins. The choice of the set of bonds B depends on the problem one is interested in. We will consider fully connected networks, where each spin is directly interacting with all the other spins in the network: $i = 1, \ldots, N$ and $j = 1, s \ldots, i-1$. In the canonical ensemble, the probability distribution of the variables $\boldsymbol{\sigma} = \{\sigma_1 \ldots \sigma_N\}$ is the Boltzmann-Gibbs distribution

$$P(\boldsymbol{\sigma}) = \frac{e^{-\beta H}}{Z}, \tag{1.2}$$

where $\beta = 1/T$ is the inverse temperature and the normalization factor Z,

$$Z = \sum_{\sigma_1=\pm 1} \sum_{\sigma_2=\pm 1} ... \sum_{\sigma_N=\pm 1} e^{-\beta H}, \tag{1.3}$$

is referred to as the partition function. In the following, we will use either the notation $\sum_{\boldsymbol{\sigma}}$ or $Tr_{\boldsymbol{\sigma}}$ to denote the sum over all possible spin configurations. The distribution (1.2) can be also interpreted from an information theoretic point of view (see, for instance, [Jay57]). Imagine we want to describe the probability distribution $P(\boldsymbol{\sigma})$ of a set of binary variables $\boldsymbol{\sigma}$. To model the system by making the minimum possible assumptions beyond what we can directly measure from the system itself, we can use the maximum entropy principle. Indeed, the Shannon entropy, defined as

$$S[P] = -\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log P(\boldsymbol{\sigma}), \tag{1.4}$$

quantifies the uncertainty of the set of random variables $\boldsymbol{\sigma}$: the larger the entropy, the less a priori information one has on the value of the variables. Hence, making the minimum assumptions on the form of $P(\boldsymbol{\sigma})$ corresponds to finding the distribution that maximizes the Shannon entropy. If something is known about the statistics of the variables, the maximization has to be performed under the corresponding constraints. Let us assume that we can compute sample averages of σ_i and $\sigma_i\sigma_j$ in the data; the maximum entropy

distribution subject to these constraints is (1.2), where $\{h_i, J_{ij}\}$ are the Lagrange multipliers that have to be chosen so that the averages $\{\langle\sigma_i\rangle, \langle\sigma_i\sigma_j\rangle\}$ with respect to (1.2) agree with experiments. In other words, the distribution of the Ising model can be seen as the less biased distribution that reproduces the observed averages and pairwise correlations between the variables.

While the forward Ising problem consists in predicting system observables – such as spin magnetizations and correlations – given a complete description of the system, in the inverse Ising problem we can measure magnetizations and correlations from a system whose parameters are unknown; the goal is to infer the parameters (i.e., couplings and local fields) from the data. We will discuss methods to perform this inference task in Chapters 4 and 5. In the following section, we provide an overview of how the inverse Ising problem has been applied to the field of computational neuroscience.

1.2.2 Explaining correlations in neural spike trains

A major challenge in neuroscience is to understand how neurons process information through collective interactions. Such understanding has been limited by the possibility to experimentally access only a tiny fraction of the exponential number of possible activity patterns. Recently, multi-electrode arrays techniques have allowed to simultaneously record the activity of hundreds of neurons, and the spatio-temporal resolution with which these recordings can be done is rapidly increasing. Although networks remain dramatically undersampled, there is now the possibility to address questions that were previously out of reach.

One central open question is the origin of the multi-neuron firing patterns observed in experiments [SBIB06], which seemed to be in contrast with the weak measured correlations between pairs of neurons. Recent works [SBIB06, CLM09, SFG+06] have used maximum entropy principles to explain how weak correlations among elements can have a strong effect on the state of the population as a whole.

For example, the authors of [SBIB06] analyze simultaneous recordings from 40 neurons in the salamander retina. Time is divided in $\Delta\tau = 20ms$ time steps, and the activity of each cell in a given time step is represented by a spin, with value $\sigma_i = 1$ if the neuron is spiking, $\sigma_i = 0$ if it is silent. The considered data is the set of observed simultaneous (i.e. within a single time bin) spike patterns, without regard to their temporal order. The authors use the Ising model as the minimal model that incorporates pairwise correlations and show that it can accurately predict the combinatorial patterns of spiking and silence in retinal ganglion cells; in contrast, models of independent neurons drastically fail. The comparison of theory and experiment is done for groups

of $N = 10$ neurons, which are small enough that the full distribution $P(\boldsymbol{\sigma})$ of a spin configuration can be sampled experimentally. The external fields and pairwise symmetric couplings are inferred by maximizing the likelihood of the data. Here, the couplings describe the interactions between neurons within the terms of the fitted model and are referred to as effective or functional connections. Hence, inferring the values of the couplings starting from the data allows to reconstruct the network of causal interactions between the neurons.

This minimal model has been generalized in different ways, for example by adding a stimulus-dependent field acting on the spins [GATSS13], or including higher order interactions [GSS11].

However, an evident limitation is that it does not address the temporal evolution of correlated states. Interestingly, the authors of [TJH$^+$08] point out that, if correlated states occurred in a temporally independent manner, concatenating the states sampled from the Ising model should give a reasonable estimate of the lengths of observed multi-neuron firing patterns. However, they observed that sequences of correlated states were significantly longer than predicted by concatenating states from the model. This suggested that temporal dependencies are a common feature of cortical network activity, and should be considered in the models.

Moreover, it was shown [TRMH13] that correlations in the statistics of neural spike trains could arise both as the effect of interaction between neurons or by sharing a common non-stationary input, where no interaction among neurons is present (see also [TMM$^+$14]).

Hence, a deeper insight could be achieved through dynamical models. Time-varying inputs and two-times correlations can be taken into account by a simple generalization to dynamics of the Ising model, i.e. the kinetic Ising model with Glauber update rule.

1.2.3 Glauber dynamics

Let us consider a set of Ising spins interacting trough couplings J_{ij} and with a (time-dependent) external field. They also interact with an external agency (e.g., a heat reservoir) which causes them to change their states randomly with time. The noise introduced by such external agency is parametrized by the inverse temperature β. Each spin $\sigma_i(t)$ is a stochastic function of time and makes random transitions between the values ± 1, according to the value of the neighboring spins. In particular, the local transition probability $w_i[\sigma_i(t + \Delta t)|\{\sigma_{j\in\partial i}(t)\}]$ that a site i at time $t + \Delta t$ has spin $\sigma_i(t + \Delta t)$, given

the value of its neighbors spins $\sigma_{j\in\partial i}(t)$ at time t, is [Gla63]:

$$\begin{aligned} w_i[\sigma_i(t+\Delta t)|\{\sigma_{j\in\partial i}(t)\}] &= \frac{e^{\beta\sigma_i(t+\Delta t)\theta_i(t)}}{2\cosh\beta\theta_i(t)}, \\ \theta_i(t) &= \sum_{j\in\partial i} J_{ij}\sigma(t)_j + h_i(t). \end{aligned} \tag{1.5}$$

Note that - contrary to the equilibrium case - we are not introducing any energy function; now the network of couplings is directed, and in general $J_{ij} \neq J_{ji}$. Self-interactions J_{ii} might be present or not: in the following sections of the thesis, we will focus on cases where such interactions are absent. The parameter β quantifies the randomness of the dynamics: for $\beta \to 0$ the dynamics is completely random, for $\beta \to \infty$ it is deterministic. Various update rules can be defined for this dynamics. One choice is to updated simultaneously all the spins at discrete time steps. Such *parallel* (or sinchronous) dynamics is defined by the Markov chain

$$P(\boldsymbol{\sigma}(t+1)) = \sum_{\boldsymbol{\sigma}^t} W[\boldsymbol{\sigma}(t+1);\boldsymbol{\sigma}(t)]P(\boldsymbol{\sigma}(t)), \tag{1.6}$$

with transition probability

$$W[\boldsymbol{\sigma}(t+1);\boldsymbol{\sigma}(t)] = \prod_i w_i[\sigma_i(t+1)|\{\sigma_{j\in\partial i}(t)\}]. \tag{1.7}$$

In case of symmetric interactions $J_{ij} = J_{ji}$ and stationary external fields $h_i(t) = h_i$, the dynamics obeys detailed balance and the equilibrium distribution can be written in the Boltzmann form $P_{\rm eq}(\boldsymbol{\sigma}) \sim e^{\beta H(\boldsymbol{\sigma})}$, where the H is the Peretto pseudo-Hamiltonian [Per84] (i.e., an Hamiltonian dependent on the inverse-temperature):

$$H(\boldsymbol{\sigma}) = -\sum_i h_i\sigma_i - \frac{1}{\beta}\sum_i \log 2\cosh[\beta\theta_i(\boldsymbol{\sigma})]. \tag{1.8}$$

A different dynamics is defined by the *sequential* (asynchronous) update rule, where at each time step only one randomly chosen spin is updated; the duration of each update is $1/N$, so that - on average - all spins have been updated once on a time scale $\mathcal{O}(N^0)$. A sequential dynamics with discrete time can be defined by the Markov chain (1.6) with transition probability [Coo01]

$$W[\boldsymbol{\sigma}(t+1);\boldsymbol{\sigma}(t)] = \frac{1}{N}\sum_i \left[\prod_{j\neq i}\delta_{\sigma_j(t+1),\,\sigma_j(t)}\right] w_i[\sigma_i(t+1)|\{\sigma_{j\in\partial i}(t)\}]. \tag{1.9}$$

In the continuous limit $N \to \infty$, the process obeys the master equation

$$\frac{d}{dt}P(\boldsymbol{\sigma}(t)) = \sum_i \left[\, P(F_i\boldsymbol{\sigma}(t))\, w_i(F_i\boldsymbol{\sigma}(t)) - P(\boldsymbol{\sigma}(t))\, w_i(\boldsymbol{\sigma}(t))\right], \tag{1.10}$$

where now

$$w_i(\boldsymbol{\sigma}(t)) = \frac{1}{2}\left[1 - \sigma_i \tanh \beta\theta_i(t)\right], \tag{1.11}$$

and where F_i is the flip operator: $F_i\boldsymbol{\sigma} = \{\sigma_1, \ldots, -\sigma_i, \ldots, \sigma_N\}$. In case of symmetric interactions and stationary external fields, the dynamics converges to the equilibrium distribution $P_{\text{eq}}(\boldsymbol{\sigma}) \sim e^{\beta H(\boldsymbol{\sigma})}$, where H is the Hamiltonian of the Ising model (1.1).

1.2.4 Inferring effective connectivities in neuronal networks

The Glauber parallel dynamics for an Ising system has been used to model networks of neurons that spike at a time-varying rate which depends on earlier spikes and on external covariates (such as a stimuli).

Spike trains recorded from N neurons are divided into small time bins, and a binary variable $\sigma_i(t)$ is assigned to each neuron i at each time bin t, with value 1 if the neuron has emitted one or more spikes in the time bin, -1 otherwise.

Two recent studies [RH11b, CFG$^+$15] simulated biologically realistic cortical models and showed that functional connections inferred using the kinetic Ising model can be successfully used for network reconstruction, i.e. to distinguish connected vs disconnected pairs of neurons. In addition, the authors of [CFG$^+$15] propose a method to overcome the limitations of a probabilistic dynamics with a single arbitrary time-step and to correct for the bias introduced by the arbitrary choice of the time-bin used to binarize the spike trains. A dynamic model - contrary to an equilibrium one - allows us to infer non-symmetric interactions, and real synaptic interaction are in general not symmetric. Yet, the exact relation between the inferred functional connections and the real synaptic connections is non-trivial and could be understood only analysing recordings from circuits where the actual physiological synapses between neurons are known [1].

However, some features of the inferred connections can be robust to changes in the model and give precious insights on the properties of the real system.

[1] So far, only few works [GKG$^+$13, VIS15, LCRP14] have validated connectivity estimates with some form of 'ground truth'; they showed that approaches based on Generalized Linear Models (GLMs) were successful in inferring the true connectivity of the circuit, while linear models and model-free approaches failed; this encourages for the choice of GLM-based approaches to estimate synaptic connectivity (note that the kinetic Ising model can be seen as a simplified GLM with a limited temporal memory).

A relevant example is the work by Dunn and collaborators [DMR15]. They analyse simultaneous recordings from tens of grid cells in two rats freely moving in a 2D environment. Grid cells are neurons that present a particular spatial selectivity, such that the positions in real space where one particular cell is firing form an hexagonal grid; the relative position of the grids of two distinct cells is called their relative phase. Fitting a kinetic Ising model with parallel update rule to the data, the authors find a systematic dependence of the couplings between two cells and their relative phase: cells with nearby phases have positive functional connection strengths, while those further apart have negative ones. The authors explain away various sources of correlations that could lead to spurious connections, such as the overlap of the firing fields, and head directional input. The result is relevant as, since attractor models of grid cells rely heavily on this type of effective connectivity, this work provides support for the idea of attractor dynamics in the grid cell assembly.

2 Thesis Outline

This dissertation is written in the form of a thesis by publication, where I collect the papers that I have co-authored in separate chapters. In each chapter, the paper is followed by unpublished results and by an appendix that briefly reviews the fundamental methods used in the paper.

Chapter 2 considers the forward problem of predicting the time evolution of system observables in the kinetic Ising model, assuming that the parameters are known. With the aim of analysing the system in a mean-field framework, we will introduce a novel technique referred to as the extended Plefka expansion. It is an extension to dynamics of the Plefka expansion for the Sherrington–Kirkpatrick model, where the novelty lies in constraining not only the first moments in the expansion, but also all marginal second moments. We conjecture that it provides the exact mean field solution to the forward problem in the thermodynamic limit of infinitely many particles, when the couplings are weak and long-ranged.

Publication included at page 17 (Publisher version): Bachschmid-Romano, Ludovica, et al. *Variational perturbation and extended Plefka approaches to dynamics on random networks: the case of the kinetic Ising model.* Journal of Physics A: Mathematical and Theoretical 49.43 (2016): 434003.

doi:10.1088/1751-8113/49/43/434003

In **Chapter 3**, we analytically study the performance of two algorithms for learning the couplings in the kinetic Ising model, focussing on the case of asymmetric couplings. The first one is based on the exact mean-field solution for the asymmetric model derived in [MS11]; the second is a Bayesian estimator of the couplings, where we approximate the posterior means - whose exact computation is intractable- using the cavity method of statistical physics. We compute the estimation error of these methods, as a function of the length of observed trajectories. The theoretical setting for our analysis is offered by the statistical mechanics of inverse problems, where the phase space consists of the couplings to be inferred, while the spin values are treated as fixed observations; the replica method of spin glasses is used to compute the average error of a given estimator of the couplings in the limit of large systems, where the ratio $\alpha = M/N$ remains finite - M being the size of the dataset and N the size of the system. The main challenge will be to treat analytically the distribution of the spin observations, but the analysis will be simplified by the fact that two-times

correlations decay after one time step for asymmetric networks. However, the equal-time correlation matrix will play a major role in determining the speed of learning and we will compute its statistics in order to get an explicit result for the estimation error. We also design an efficient algorithm to numerically implement the optimal Bayes estimator.

Publication included at page 68 (Publisher version): Bachschmid-Romano, Ludovica, and Manfred Opper. *Learning of couplings for random asymmetric kinetic Ising models revisited: random correlation matrices and learning curves.* Journal of Statistical Mechanics: Theory and Experiment 2015.9 (2015): P09016.

doi:10.1088/1742-5468/2015/09/P09016

The formalism of Chapter 3 is then extended in **Chapter 4** to analyse the error of learning the couplings in an equilibrium model. The distribution of the data is even more difficult to treat, due to the presence of the normalising partition function, and we will use a combination of the cavity and replica methods of spin glasses to carry out the calculation and include the effect of correlations. We study the performance of algorithms based on the minimisation of a local cost function, focussing on the pseudo-likelihood and the mean-field estimators. Surprisingly, we will find that a simple quadratic cost function is the one that achieves minimal error, and the explicit estimator associated with it can be entirely constructed from data.

Publication included at page 120 (Publisher version): Bachschmid-Romano, Ludovica, and Manfred Opper. *A statistical physics approach to learning curves for the inverse Ising problem.* Journal of Statistical Mechanics: Theory and Experiment 2017.6 (2017): 063406.

doi:10.1088/1742-5468/aa727d

Chapter 5 treats an extension of the kinetic Ising model, where a fraction of the trajectories is observed and a fraction is hidden. Using the replica method, we analytically compute the average error of the Bayes optimal estimator of hidden spins, which is obtained from the posterior distribution of unobserved spins given the observed ones. We then turn to the study of single instances of the network. Using the extended Plefka expansion, we derive a set of mean-field equations characterising the dynamics of the hidden-spin variables, and we can accurately estimate the single-site magnetisation of hidden spins. In the end, we discuss the applicability of our result as building block for an algorithm aimed at reconstructing the network connections.

Publication included at page 158 (Publisher version): Bachschmid-Romano, Ludovica, and Manfred Opper.*Inferring hidden states in a random kinetic Ising model: replica analysis.* Journal of Statistical Mechanics: Theory and Experiment 2014.6 (2014): P06013.

doi:10.1088/1742-5468/2014/06/P06013

In the sixth and last chapter, we summarise the conclusions of the single chapters and indicate future research directions.

3 Mean field approaches to dynamics on random networks: the kinetic Ising model

3.1 Introduction

Mean-field methods of statistical physics allow for a tractable description of complex systems of many interacting variables. Based on the assumption that the fluctuations around the average value of the order parameters are small, they provide a solution in which each variable is subject to an effective local field, considering the interactions with other degrees of freedom, on average. Many highly non-trivial mean-field approximations were used to derive the main equilibrium properties of the Ising spin glass model [MPV87][1], and made it possible to introduce efficient algorithms for statistical inference and optimisation [MM09]. In recent years, a growing interest has been dedicated to extending such mean-field techniques to the dynamic counterpart of the Ising model.

Various kinds of dynamics can be defined for the Ising model. We are interested in studying the Glauber dynamics with parallel update rule, for systems where the matrix of couplings is fixed and can have any degree of symmetry, bearing in mind the applicability of this framework to model the dynamics in biological networks, based on time-series data.

However, initial mean field approaches to the dynamics of spin systems were developed for the disorder averaged case scenario. Soft spin models were the first to be considered. If the connections between the spins are symmetric, the dynamics of the network can be described as a relaxation of a global energy function towards local minima. A relaxation dynamic of the Langevin type has been extensively analysed (for a review, see [BCKM98, Cug03]), as it provides a framework for studying off-equilibrium behaviours (such as the phenomenon of ageing in glassy systems) and unveiled strong analogies between disordered systems and other types of glasses where disorder is absent, such as structural fragile glasses [Par06]. If the connections are not symmetric, an

[1] For a brief introduction to spin glasses and disorder averages, see section 4.A.1.

energy function cannot be defined; however, a Langevin equation formalism was developed in [CS87], where a set of local self-consistent equations for the spin correlations and response functions is derived.

For discrete spins, the discontinuous nature of the variables rules out an approach based on Langevin equations. Instead, the stochasticity of the dynamics can be formulated in a probabilistic setting, typically in terms of a Glauber rule [Gla63] that - given the value of the spin variable at the current time - specifies the probability of observing a given spin configuration at a following time. Time can be considered either as a discrete or a continuous variable, and the spin values can be updated all at the same time or sequentially one by one, giving rise to diverse types of dynamics. For spin glass models with hard spins, a path integral formalism to describe such Glauber dynamics with continuous time was introduced by Sommers [Som87]. Crisanti and Sompolinsky [CS88] observed that the mean-field equations for a network with partially asymmetric couplings are quite difficult to solve, but they simplify remarkably in the particular case of a network with fully asymmetric couplings: two-times correlation decay to zero, and - in the N large limits - the local fields can be replaced by a time-dependent Gaussian random field.

An alternative approach was proposed by Coolen and collaborators [CS93, CS94, CLS96]. Their dynamical replica analysis is based on a generating functional formalism and derives deterministic flow equations for macroscopic state variables. Such works were followed by [NY96], where the Glauber dynamics of the SK model is studied at high temperatures. The authors explicitly compute the microscopic probability distribution of the spin configuration as a function of time, via a high-temperature expansion.

The role of the degree of symmetry on the transient dynamics of a system at zero temperature was analysed in [EO94]. A combination of dynamical functional methods and Monte Carlo simulations allows us to identify – by varying the average symmetry of the couplings – a transition from ergodic-dynamics to a phase where a finite fraction of spins freezes[2].

Models with fixed quenched disorder were studied more recently. Dynamical TAP equations have been first derived for the spherical p-spin model in [Bir99]. This work also analyses the conditions under which the dynamics is a relaxation in the TAP free-energy landscape. Analogous TAP equations were also recently found in [BSO16], where the model is extended to comprise generic continuous variables and nonlinear interaction terms; the solution is found by a generating functional approach closely related to the one that we will discuss in Paper 1.

[2] A review of those and other methods used for both soft and hard spin models can be found in [HKP91, Coo01].

For Ising spins, by information geometric arguments, Kappen and Spanjers [KS00] argued that asymmetric networks at the stationary state obey the same TAP equations (3.6) valid for the equilibrium model. This result is a good approximation for weak couplings, but does not provide the exact mean field description, as was later proved by Mezard and Sakellariou [MS11]. The authors, following the approach of [CS88], observe that asymmetric networks exhibit small correlations among spins at various times. A central limit theorem argument shows that effective fields are Gaussian distributed, and the resulting mean-field description of the dynamics is exact in the thermodynamic limit for weak and long-range couplings[3]. The whole transient dynamics for networks with an arbitrary degree of symmetry was first studied in [RH11a], where, using a generating functional approach, the authors derive TAP-like equations via a small couplings expansion. Another derivation of these equations using information geometry was reported in [AM12]. However, in the limit of an asymmetric network, their result does not agree with the exact one of [MS11]. Saad and Mahmoudi extended the work of [MS11] to the case of couplings with arbitrary symmetry [MS14]. The authors still consider the effective fields as Gaussian distributed but introduce non-zero covariance between spins at different times and provide recursive equations to compute correlations at all times. The result improves on the other methods and recover the exact theory for asymmetric networks; however, for a arbitrary degree of symmetry, the exactness of this method in the limit $N \to \infty$ remains an open question.

Paper 1 introduces novel approaches to the problem. First, two methods for deriving a naive mean-field equation in the static case are extended to the kinetic case: a variational approach based on minimising the Kullback-Leibler divergence between the true distribution of spins and the distribution of independent trajectories, and a saddle-point approximation to the generating functional. Then, two novel approximations are presented. In a variational perturbative approximation, the action in the path integral representation of the generating functional is expanded around a quadratic function in the fields and conjugate fields; the latter function depends on variational parameters that are optimised to obtain minimum sensitivity of the approximating functional to the variational parameters. In other words, the generating functional of the dynamics is approximated by a Gaussian distribution, and its parameters are later optimised. The result will strongly depend on the constraints imposed on the parameters of such Gaussian distribution, in particular on its covariance matrix. Finally, we present an extension of the Plefka expansion for dynamics introduced in [RH11a].

[3]The result is not consistent with the one of [KS00]).

Plefka's expansion [Ple82] was originally performed to derive mean-field (TAP) equations for the equilibrium Sherrington-Kirkpatrick model, by expanding the free energy at fixed magnetisation (Gibbs potential) in powers of the interaction strength. In the kinetic case, the variables of interest are no longer single spins but entire spin trajectories, and a path integral formalism is introduced to compute averages over trajectories; a Gibbs free energy cannot be defined in the out-of-equilibrium scenario, but the logarithm of the partition function at fixed moments over time will provide the analogous function to be expanded. While in the first generalisation to dynamics [RH11a] of Plefka's expansion only marginal first moments over time are fixed, in Paper 1 we will show that also the second order moments must be considered for a correct analysis. The result will outperform other current methods in predicting the time evolution of the single-spin magnetisations, and we conjecture that it provides the correct mean field solution to the forward problem in the thermodynamic limit of infinitely many particles when the couplings are weak and long-ranged. In the discussion section, we will further compare among different methods.

An introduction to Plefka's expansion (and to the cavity method) is given in section 3.A, where we derive the mean field (TAP) equations for the equilibrium Sherrington-Kirkpatrick model. Section 3.B reviews previous approaches to the transient dynamics of an Ising model with parallel Glauber update rule.

3.2 Paper 1.

Author's contribution: I performed the analytical and numerical calculations relative to: Section 5 (Extended Plefka expansion); Appendix D (Details on the extended Plefka expansion); Appendix E (The YuleWalker equations). I wrote the relative sections in the paper. I contributed to writing section 6 (Numerical results), Section 7 (Summary and Conclusions) and to the preparation of the figures.

Variational perturbation and extended Plefka approaches to dynamics on random networks: the case of the kinetic Ising model

2016 J. Phys. A: Math. Theor. 49 434003

(http://iopscience.iop.org/1751-8121/49/43/434003)

IOP Publishing Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **49** (2016) 434003 (33pp) doi:10.1088/1751-8113/49/43/434003

Variational perturbation and extended Plefka approaches to dynamics on random networks: the case of the kinetic Ising model

L Bachschmid-Romano[1,4], C Battistin[2,4], M Opper[1] and Y Roudi[2,3]

[1] Department of Artificial Intelligence, Technische Universität Berlin, Marchstraße 23, Berlin, D-10587, Germany
[2] Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim, Norway
[3] Institute for Advanced Study, Princeton, USA

E-mail: ludovica.bachschmidromano@tu-berlin.de and claudia.battistin@ntnu.no

Received 19 April 2016, revised 22 July 2016
Accepted for publication 3 August 2016
Published 3 October 2016

Abstract

We describe and analyze some novel approaches for studying the dynamics of Ising spin glass models. We first briefly consider the variational approach based on minimizing the Kullback–Leibler divergence between independent trajectories and the real ones and note that this approach only coincides with the mean field equations from the saddle point approximation to the generating functional when the dynamics is defined through a logistic link function, which is the case for the kinetic Ising model with parallel update. We then spend the rest of the paper developing two ways of going beyond the saddle point approximation to the generating functional. In the first one, we develop a variational perturbative approximation to the generating functional by expanding the action around a quadratic function of the local fields and conjugate local fields whose parameters are optimized. We derive analytical expressions for the optimal parameters and show that when the optimization is suitably restricted, we recover the mean field equations that are exact for the fully asymmetric random couplings (Mézard and Sakellariou 2011 *J. Stat. Mech.* 2011 L07001). However, without this restriction the results are different. We also describe an extended Plefka expansion in which in addition to the magnetization, we also fix the correlation and response functions. Finally, we numerically study the performance of these approximations for

[4] These authors contributed equally to this work.

Sherrington–Kirkpatrick type couplings for various coupling strengths and the degrees of coupling symmetry, for both temporally constant but random, as well as time varying external fields. We show that the dynamical equations derived from the extended Plefka expansion outperform the others in all regimes, although it is computationally more demanding. The unconstrained variational approach does not perform well in the small coupling regime, while it approaches dynamical TAP equations of (Roudi and Hertz 2011 *J. Stat. Mech.* 2011 P03031) for strong couplings.

Keywords: random graphs, non-equilibrium processes, spin glasses, variational methods, perturbational methods

(Some figures may appear in colour only in the online journal)

1. Introduction

The kinetic Ising spin glass model is a prototypical model for studying the dynamics of disordered systems. Previous work on this topic focused both on studying the average—over couplings—behavior of various order parameters, such as magnetizations, correlations and response functions, and in more recent years, developing approximate methods for relating the dynamics of a given realization of the model to its parameters. The latter line of work has received a lot of attention in recent years, in part, because of the applications it has on developing approximate inference methods for point processes which in turn are receiving particular attention due to the on going improvements in data acquisition techniques in various disciplines in life sciences.

Most of the early work on the topic dealt with systems with symmetric interactions, until Crisanti and Sompolinsky [3] studied the disorder averaged dynamics of Ising models with various degrees of symmetry and Kappen and Spanjers [4] derived naive mean field and TAP equations for the stationary state of the Ising model for arbitrary couplings, in both cases considering Glauber dynamics. Roudi and Hertz [2] derived dynamical TAP equations (hereafter denoted by RH-TAP) for both discrete time parallel and continuous time Glauber dynamics using Plefka's method [5], originally used for studying equilibrium spin glass models, extended to dynamics. This was followed by [6] who reported another derivation of these equations using information geometry following the approach of [4]. Mezard and Sakellariou [1] developed a mean field method (hereafter denoted by MS-MF) which is exact for large networks with independent random couplings; see also [7]. Two schemes for improving the existing mean-field description were proposed in [8] an elegant generalized mean field methods was followed in [9].

In the current paper we follow up on these efforts and report some new results on the dynamics of kinetic Ising model with parallel dynamics. We first look at the relationship between the saddle point approximation to the path integral representation of the dynamics and the simplest variational approach based on minimizing the Kullback–Leibler (KL) divergence between the true distribution of the spin trajectories and a factorized distribution. Although for the standard kinetic Ising model the two methods yield the same equations of motion, we see that this is not in general the case when the probability of spin configurations at a given time given those of the previous time is not a logistic function of the fields. After this, we consider two approaches for going beyond the saddle point solution of the path

integral representation of the dynamics of the standard kinetic Ising model with parallel dynamics (defined in more detail in the following sections).

In one of these approaches, which we refer to as gaussian average variational method, we perform a Taylor expansion of the action in the path integral representation of the generating functional around a quadratic function of the fields and conjugate fields. As described in described in detail in section 4, we then choose the parameters of this function such that the resulting functional minimally depends on these parameters. We derive analytical expressions for these optimal solutions and show that for a fully asymmetric network under a further assumption about the interaction between the fields and the conjugate fields, we can recover the equations of motion for the magnetization identical to MS-MF equations [1]. Without this assumption we observe that the resulting equations are different from MS-MF. In the second approach, we go beyond the saddle point by performing an extended Plefka expansion. The standard Plefka expansion for the equilibrium model involves performing a small coupling approximation of the free energy at fixed magnetization and is the approach that was originally taken in [2]. As we show here, however, similar to the soft spin models [10, 11], a better description of the dynamics can be achieved by not only fixing the magnetizations but also pairwise correlation and response functions while expanding around the uncoupled model.

2. The dynamical model

We consider the synchronous dynamics of N interacting binary spins in the time window $[0, T]$ defined by

$$P(\mathbf{s}^{0:T}) = \prod_{t=0}^{T-1} P(\mathbf{s}(t+1)|\mathbf{s}(t))P(\mathbf{s}(0)), \tag{1}$$

in which

$$P(\mathbf{s}(t+1)|\mathbf{s}(t)) = \prod_{i=1}^{N} f_{s_i(t+1)}(H_i(t)), \tag{2}$$

where

$$H_i(t) = h_i(t) + \sum_{j=1}^{N} J_{ij} s_j(t) \tag{3}$$

is the total field acting on spin i at time t composed of the external field h_i and the fields felt from other spins in the system. The function $f_{s_i(t+1)}(H_i(t))$ is a generic transfer function or conditional probability of the state of the spin i at time $(t+1)$ given the field at time t. Our goal will be to calculate the mean magnetizations of the spins.

The generating functional of the distribution $P(\mathbf{s}^{0:T})$, expressed as a path integral is

$$Z[\psi, \mathbf{h}, \mathbf{J}] = \langle e^{i\psi^\top \mathbf{s}} \rangle_P = \frac{1}{2^N (2\pi)^{NT}} \int D\mathcal{G} e^{-L[\mathbf{h}, \psi, \mathcal{G}]}, \tag{4}$$

where $\langle ... \rangle_P$ denotes averaging with respect to the history of trajectories defined by (1) and (2), and

$$L[\mathbf{h},\,\boldsymbol{\psi},\,\mathcal{G}] \equiv -\mathrm{i}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\hat{g}_i(t)(g_i(t)-h_i(t)) + \sum_{i=1}^{N}\sum_{t=0}^{T}\ln \mathrm{Tr}_{s_i(t)} f_{s_i(t)}(g_i(t-1))\mathrm{e}^{\mathrm{i}s_i(t)\left[\psi_i(t)-\sum_j J_{ji}\hat{g}_j(t)\right]} \tag{5}$$

having set $\hat{g}_i(T)=0$ and $f_{s_i(0)}(g_i(-1))=1$. Notice that we assumed the initial state $\mathbf{s}(0)$ to be uniformly distributed, manifested in the factor $1/2^N$ in (4), and that we refer to the two auxiliary variables with the compact notation $\mathcal{G}\equiv\{g_i(t),\,\hat{g}_i(t)\}_{i=1\ldots N,t=0\ldots T}$ and $D\mathcal{G}=\prod_{i,t}\mathrm{d}g_i(t)\mathrm{d}\hat{g}_i(t)$.

The magnetization of spin i at time t can then be obtained as the first derivative of the log-generating functional:

$$m_i(t) = -\mathrm{i}\lim_{\boldsymbol{\psi}\to\mathbf{0}}\frac{\partial \ln(Z[\boldsymbol{\psi},\,\mathbf{h},\,\mathbf{J}])}{\partial\psi_i(t)}. \tag{6}$$

Let us make a brief note on how the the integral representation of the generating functional in (4) and (5) has been derived. This is done by first replacing $H_i(t)$ in (2) by $g_i(t)$ and integrating over all $g_i(t)$ while enforcing that at each time step and for each spin $g_i(t)=H_i(t)$ by inserting δ-functions, $\delta[g_i(t)-H_i(t)]$, in the integral. One then writes this delta function in its integral representation

$$\delta[g_i(t)-H_i(t)] = \int\frac{\mathrm{d}\hat{g}_i(t)}{2\pi}\exp\{\mathrm{i}\hat{g}_i(t)[g_i(t)-H_i(t)]\} \tag{7}$$

which is how the $\hat{g}_i(t)$ appear in the equations. This rewriting of the generating functional constitutes the first steps in the Martin–Siggia–Rose–De Domenicis–Peliti formalism [12, 13] once it is adapted for hard spins. For more details about this approach and a pedagogical review on its application to soft and hard spin dynamics see [14, 15].

A logistic transfer function f in (2), such that $f_{s_i(t+1)}(H_i(t)) = \frac{1}{2}(1+s_i(t+1)\tanh H_i(t))$, yielding the following probability distribution over spin paths

$$\mathrm{P}(\mathbf{s}^{0:T}) = \frac{1}{2^N}\prod_{i=1}^{N}\prod_{t=0}^{T-1}\frac{\mathrm{e}^{s_i(t+1)H_i(t)}}{2\cosh(H_i(t))}, \tag{8}$$

corresponds to the *standard* kinetic Ising model with parallel update studied in previous work [1, 2, 6, 9].

This path integral representation in (4) allows us to explicitly perform the trace over the spins in the generating functional of (4) and (5) yielding

$$\begin{aligned} L[\mathbf{h},\,\boldsymbol{\psi},\,\mathcal{G}] \equiv & -\sum_{i=1}^{N}\sum_{t=0}^{T-1} lc\left[g_i(t-1)+\mathrm{i}\psi_i(t)-\mathrm{i}\sum_l J_{li}\hat{g}_l(t)\right] \\ & -\mathrm{i}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\hat{g}_i(t)(g_i(t)-h_i(t)) \\ & -\sum_{i=1}^{N} lc[g_i(T-1)+\mathrm{i}\psi_i(T)] + \sum_{i=1}^{N}\sum_{t=0}^{T-1} lc[g_i(t)], \end{aligned} \tag{9}$$

where we have set $g_i(-1)=0\ \forall i$ and

$$lc[x] \equiv \log\cosh(x). \tag{10}$$

3. Mean field

As a prologue to our more important results in the following sections, in this section we review the derivation of mean field equations for the dynamical model in (2) using two approaches. These are the saddle point approximation to the path integral representation of the generating functional in (4), and the minimization of the KL distance between the true distribution P in (1) and a factorized one. Despite being formally different methods, in the literature they are both often referred to as *mean field* and it is indeed well know that for the specific case of the equilibrium Ising model, they lead to the very same set of equations, known as *naïve mean field equations* [16]. Throughout this section the transfer function f in (2) is considered a generic function of the field $H_i(t)$. Only towards the end of this section we are going to consider f as a logistics function of the kinetic Ising model.

3.1. Saddle point mean field

In the equilibrium case, one way to derive the naïve mean field equations is as the equations describing the saddle point approximation to a path integral representation of the free energy, while the TAP equations are those derived by calculating the gaussian integral around the saddle point [17]. (Another way is by means of Plefka expansion, which at this point we do not discuss but will get back to later on). Let us consider this saddle point approach for the kinetic model in (2) and the corresponding generating functional (4). Defining a complex measure q as

$$q(s_i(t)|g_i(t-1),\hat{\mathbf{g}}(t),\psi_i(t)) = \frac{f_{s_i(t)}(g_i(t-1))\mathrm{e}^{\mathrm{i}s_i(t)\left[\psi_i(t)-\sum_j J_{ji}\hat{g}_j(t)\right]}}{F_{it}(g_i(t-1),\hat{\mathbf{g}}(t),\psi_i(t))}, \tag{11}$$

where $F_{it}(g_i(t-1),\hat{\mathbf{g}}(t),\psi_i(t))$ is the normalization constant, the saddle point equations for the generating functional of (4), namely the stationary points of the function $L[\mathbf{h},\psi,\mathcal{G}]$, in (5), $\hat{g}_i^{\mathrm{SP}}(t)$ and $g_i^{\mathrm{SP}}(t)$, read

$$\hat{g}_i^{\mathrm{SP}}(t) = \mathrm{i}\left\langle \frac{f'_{s_i(t+1)}(g_i^{\mathrm{SP}}(t))}{f_{s_i(t+1)}(g_i^{\mathrm{SP}}(t))}\right\rangle_{q(s_i(t+1)|g_i^{\mathrm{SP}}(t),\hat{\mathbf{g}}^{\mathrm{SP}}(t+1),\psi_i(t+1))}, \tag{12a}$$

$$g_i^{\mathrm{SP}}(t) = h_i(t) + \sum_j J_{ij}\langle s_j(t)\rangle_{q(s_j(t)|g_j^{\mathrm{SP}}(t-1),\hat{\mathbf{g}}^{\mathrm{SP}}(t),\psi_j(t))}, \tag{12b}$$

where we have defined $f'_x(y) \equiv \frac{\partial f_x(y)}{\partial y}$. Notice that in the limit $\boldsymbol{\psi}\to\mathbf{0}$, $\hat{\mathbf{g}}^{\mathrm{SP}}=\mathbf{0}$ is a self-consistent solution of the previous saddle point equation (12*a*), while (12*b*) turns into

$$g_i^{\mathrm{SP}}(t) = h_i(t) + \sum_j J_{ij}\langle s_j(t)\rangle_{f_{s_j(t)}(g_j^{\mathrm{SP}}(t-1))}. \tag{13}$$

The approximate log generating functional $\ln Z[\boldsymbol{\psi},\mathbf{h},\mathbf{J}] \simeq -L[\mathbf{h},\boldsymbol{\psi},\boldsymbol{\mathcal{G}}^{\mathrm{SP}}] + \text{const.}$ allows us to estimate the magnetizations using (6) and (13) as

$$m_i(t) = \langle s_i(t)\rangle_{f_{s_i(t)}\left(h_i(t-1)+\sum_j J_{ij}m_j(t-1)\right)}. \tag{14}$$

These are the saddle point mean field equations for a general function f. Note that the marginal here yields the same expression as the conditional probability in (2), namely $f_{s_i(t+1)}(H_i(t))$ except that in (14), the fluctuating field $H_i(t)$ has been replaced by an effective

(mean) field $H_i^{\text{eff}}(t) = h_i(t) + \sum_j J_{ij} m_j(t)$, in analogy with the physical intuition behind the original formulation of the mean field theory by Weiss [18].

3.2. Mean field from KL distance

A second way of deriving mean field equations, usually employed in the machine learning community, is based on a variational approximation. Within this framework, one approximates the model distribution $P(\mathbf{s}^{0:T})$ with a Markovian process $Q(\mathbf{s}^{0:T})$ that factorizes over the spin trajectories [19]. In other words, assuming

$$Q(\mathbf{s}^{0:T}) = \prod_{t=0}^{T-1} Q(\mathbf{s}(t+1)|\mathbf{s}(t)) Q(\mathbf{s}(0)), \tag{15}$$

where

$$Q(\mathbf{s}(t+1)|\mathbf{s}(t)) = \prod_{j=1}^{N} Q(s_j(t+1)|s_j(t)), \tag{16}$$

one minimizes the KL divergence, $D_{\text{KL}}[Q(\mathbf{s}^{0:T})\|P(\mathbf{s}^{0:T})]$, between the approximate distribution $Q(\mathbf{s}^{0:T})$ and the model $P(\mathbf{s}^{0:T})$. In the case of the model defined in (2) and an approximate distribution satisfying (15) and (16), the KL-divergence can be rewritten as

$$D_{\text{KL}}[Q(\mathbf{s}^{0:T})\|P(\mathbf{s}^{0:T})] \equiv \text{Tr}_{\mathbf{s}^{0:T}} Q(\mathbf{s}^{0:T}) \ln \frac{Q(\mathbf{s}^{0:T})}{P(\mathbf{s}^{0:T})}, \tag{17a}$$

$$= \sum_t \text{Tr}_{\mathbf{s}(t)} Q(\mathbf{s}(t)) \text{Tr}_{\mathbf{s}(t+1)} Q(\mathbf{s}(t+1)|\mathbf{s}(t)) \ln \frac{Q(\mathbf{s}(t+1)|\mathbf{s}(t))}{P(\mathbf{s}(t+1)|\mathbf{s}(t))}$$

$$= \sum_t \text{Tr}_{s_j(t)} Q(s_j(t)) \text{Tr}_{s_j(t+1)} Q(s_j(t+1)|s_j(t)) \ln \frac{Q(s_j(t+1)|s_j(t))}{u_{jt}(s_j(t+1)|s_j(t))}$$

$$+ \sum_t \sum_{i \neq j} \text{Tr}_{s_i(t)} Q(s_i(t)) \text{Tr}_{s_i(t+1)} Q(s_i(t+1)|s_i(t)) \ln Q(s_i(t+1)|s_i(t)), \tag{17b}$$

where the first line is just the definition of the KL-divergence, in the second line we have exploited the Markovian property of P and Q and assumed $P(\mathbf{s}(0)) = Q(\mathbf{s}(0))$, while in the last line we have use the factorizability of Q over spin trajectories. Notice that the last equality is valid for any choice of j and that we have defined u_{jt} as

$$u_{jt}(s_j(t+1)|s_j(t)) \equiv \exp\{ \text{Tr}_{\{\mathbf{s}_{\setminus j}(t+1), \mathbf{s}_{\setminus j}(t)\}} Q(\mathbf{s}_{\setminus j}(t+1), \mathbf{s}_{\setminus j}(t)) \ln P(\mathbf{s}(t+1)|\mathbf{s}(t)) \} \tag{18}$$

and $\mathbf{s}_{\setminus j}(t)$ denotes all components of $\mathbf{s}(t)$ apart from j. Observe that thanks to the Markovian property of the two distributions P and Q we were able to reduce the average over a NT dimensional space to a sum of T averages over $2N$ dimensional spaces.

In order to determine the variational mean field equations, one has to minimize the KL-divergence in the space of marginals $Q(s_j(t))$ and transition probabilities $Q(s_j(t+1)|s_j(t))$. Given that these are not independent, we enforce the constraints:

$$Q(s_j(t+1)) = \text{Tr}_{s_j(t)} Q(s_j(t+1)|s_j(t)) Q(s_j(t)) \tag{19}$$

using Lagrange multipliers $\lambda(s_j(t))$, ultimately optimizing the following cost function:

$$\mathcal{L} \equiv D_{\text{KL}}[Q(\mathbf{s}^{0:T})\|P(\mathbf{s}^{0:T})] - \sum_{j,t} \text{Tr}_{s_j(t)} \lambda(s_j(t)) \{ Q(s_j(t)) - \text{Tr}_{s_j(t-1)} Q(s_j(t)|s_j(t-1)) Q(s_j(t-1)) \}. \tag{20}$$

The stationary points of $\mathcal{L}$ in (20) are the zeros of the functional derivatives

$$\frac{\delta\mathcal{L}}{\delta Q(s_j(t))} = \mathrm{Tr}_{s_j(t+1)} Q(s_j(t+1)|s_j(t)) \ln \frac{Q(s_j(t+1)|s_j(t))}{u_{jt}(s_j(t+1)|s_j(t))} - \lambda(s_j(t)) + \mathrm{Tr}_{s_j(t+1)} \lambda(s_j(t+1)) Q(s_j(t+1)|s_j(t)), \tag{21a}$$

$$\frac{\delta\mathcal{L}}{\delta Q(s_j(t+1)|s_j(t))} = Q(s_j(t)) \left\{ \ln \frac{Q(s_j(t+1)|s_j(t))}{u_{jt}(s_j(t+1)|s_j(t))} + 1 + \lambda(s_j(t+1)) \right\} \tag{21b}$$

that can be reduced to the relation:

$$Q(s_j(t+1)|s_j(t)) = \frac{u_{jt}(s_j(t+1)|s_j(t))}{\mathrm{Tr}_{s_j(t+1)} u_{jt}(s_j(t+1)|s_j(t))}. \tag{22}$$

It is worth emphasizing that this solution is valid for any Markov chain P and any approximate Markov distribution Q that factorizes over the spin trajectories. From now on we will require the spins at time t to be conditionally independent under the model distribution, as in (2). This assumption and a little algebra allow us to simplify (22) as follows:

$$Q[s_j(t+1)|s_j(t)] = \frac{\exp\left\{ \mathrm{Tr}_{\mathbf{s}_{\setminus j}(t)} Q[\mathbf{s}_{\setminus j}(t)] \ln f_{s_j(t+1)} \left(h_j(t) + \sum_l J_{jl} s_l(t) \right) \right\}}{\mathrm{Tr}_{s_j(t+1)} \exp\left\{ \mathrm{Tr}_{\mathbf{s}_{\setminus j}(t)} Q[\mathbf{s}_{\setminus j}(t)] \ln f_{s_j(t+1)} \left(h_j(t) + \sum_l J_{jl} s_l(t) \right) \right\}}, \tag{23}$$

where we imposed the normalizability to Q.

If there are no self-couplings in the model distribution P, the right-hand side of (23) will not depend on $s_j(t)$ and consequently the solution for the joint distribution $Q(\mathbf{s}^{0:T})$ will factorize in time. The spin independent 1st order Markov chain Q that best approximates the model P defined in (2) with $J_{jj} = 0$, is actually a 0th order Markov chain. Additionally the absence of self-interactions in P makes (23) an explicit relation between the marginal of spin j at time $t+1$ and the marginals of all spins but j at the previous time step t. Since we are dealing with a system of binary units, marginals are fully determined by their first moments, thus the marginal of spin j at time $t+1$, in (23), becomes a function of the magnetizations at time t. Taking one step further one can easily verify that the first moments of (23) equal the naïve mean field magnetizations of (14) if the transition probability $f_{s_i(t+1)}(H_i(t))$ belongs to the exponential family with the field $H_i(t)$ as natural parameter

$$f_{s_i(t+1)}(H_i(t)) = \frac{\exp[a(s_i(t+1)) H_i(t)]}{\mathrm{Tr}_{s_i(t)} \exp[a(s_i(t+1)) H_i(t)]}, \tag{24}$$

where $a(\cdot)$ is a generic function of the state $s_i(t+1)$. For the kinetic Ising model $a(\cdot)$ is the identity function and the equations for the magnetizations read:

$$m_i(t) = \tanh\left[h_i(t-1) + \sum_j J_{ij} m_j(t-1) \right], \tag{25}$$

equivalent to (14) and know as the dynamical Naïve mean field equations [2].

4. Gaussian average method

What we have shown so far is that the saddle point approximation to the generating functional for the kinetic Ising model and the one based on the KL divergence match each other,

although this is not the case for non-logistics transfer functions. In this section, we study an improvement over the saddle point approximation. Our approach is to find the optimal gaussian distribution for approximating the generating functional perturbatively, and then using the resulting approximation to calculate the magnetizations. This can be thought as an extension to complex measures of a standard variational method: it was taken by Müschlegel and Zittartz [20] for the equilibrium Ising model, while a general framework is set in [21]. We describe this approach in detail in this section.

4.1. Optimization

We consider the first order Taylor expansion of the log-generating functional defined in (4) and (5) around a gaussian integral:

$$-\ln(Z[\boldsymbol{\psi}, \mathbf{h}, \mathbf{J}]) \simeq -\ln \int \tilde{D}\boldsymbol{\mathcal{G}} + \frac{\int \tilde{D}\boldsymbol{\mathcal{G}}(L - L_s)}{\int \tilde{D}\boldsymbol{\mathcal{G}}} + NT \ln 2\pi + N \ln 2, \tag{26}$$

where we have defined the complex gaussian measure

$$\tilde{D}\boldsymbol{\mathcal{G}} = D\boldsymbol{\mathcal{G}} \mathrm{e}^{-L_s}, \tag{27a}$$

$$L_s = \frac{1}{2}(\boldsymbol{\mathcal{G}} - \bar{\boldsymbol{\eta}})^{\top} \mathbf{S} (\boldsymbol{\mathcal{G}} - \bar{\boldsymbol{\eta}}) \tag{27b}$$

parametrized by the interaction matrix $\mathbf{S}$ and the mean $\bar{\boldsymbol{\eta}}$. Here we split the vectors $\bar{\boldsymbol{\eta}}$ into $\{\boldsymbol{\eta}(t), \hat{\boldsymbol{\eta}}(t)\}_{t=0}^{T-1}$ and $\boldsymbol{\eta}(t)$ into $\{\eta_i(t)\}_{i=1}^{N}$ similar to $\boldsymbol{\mathcal{G}}$. From now on we will use the form of the action L in (9) since we are going to focus on the standard parallel update kinetic Ising model.

The choice of a quadratic form for L_s allows us to easily calculate many of the terms in (26), simplifying the expression for the log-generating functional as

$$\begin{aligned}
-\ln Z = {} & \frac{1}{2}\ln\det \mathbf{S} - NT - \mathrm{i}\sum_{i,t} \mathbf{S}^{-1}_{2Nt+i;2Nt+N+i} - \mathrm{i}\sum_{i,t} \eta_i(t)\hat{\eta}_i(t) \\
& + \mathrm{i}\sum_{i,t} \hat{\eta}_i(t) h_i(t) - \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \sum_i \sum_{t=1}^{T-1} \int \tilde{D}\boldsymbol{\mathcal{G}}' lc\,[g_i'(t-1) + \eta_i(t-1) + \mathrm{i}\psi_i(t) \\
& - \mathrm{i}\sum_l J_{li}\hat{g}_l'(t) - \mathrm{i}\sum_l J_{li}\hat{\eta}_l(t)\Big] \\
& - \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \sum_i \int \tilde{D}\boldsymbol{\mathcal{G}}' lc \Big[\mathrm{i}\psi_i(0) - \mathrm{i}\sum_l J_{li}\hat{g}_l'(0) - \mathrm{i}\sum_l J_{li}\hat{\eta}_l(0)\Big] \\
& - \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \sum_i \int \tilde{D}\boldsymbol{\mathcal{G}}' lc\,[g_i'(T-1) + \eta_i(T-1) + \mathrm{i}\psi_i(T)] \\
& + \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \sum_{i,t} \int \tilde{D}\boldsymbol{\mathcal{G}}' lc\,[g_i'(t) + \eta_i(t)] + N\ln 2,
\end{aligned} \tag{28}$$

where we have replaced (9) in (26) and we have performed the change of variables $\boldsymbol{\mathcal{G}}' = \boldsymbol{\mathcal{G}} - \bar{\boldsymbol{\eta}}$. Notice that when not stated otherwise the sum over t runs from $t = 0$ to $t = T - 1$. From now on we will just drop off the superscript $'$ from variables $\boldsymbol{\mathcal{G}}$.

If all measures were real probability measures, the first order approximation on the right-hand side of (26) would be an upper bound to the free energy $-\ln Z$. In this case a minimization of the bound with respect to the variational parameters would be the obvious choice

for optimizing the approximation. Since integrations in our case are over complex measures this argument cannot be applied. Instead, we base our optimization on the idea of the variational perturbation method [22]: if the Taylor series expansion of the log generating functional (4)–(9) would be continued to infinite order it would represent the functional and the resulting series would be entirely independent of the parameters of the gaussian measure (27*a*). On the other hand, the truncated series (26) inherits a dependence on the variational parameters $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$, $\mathbf{S}$. Hence, one would expect that the truncation represents the most sensible approximation if it depends the least on these parameters. One should therefore choose their optimal values such that the approximation to $\ln Z$ is the most insensitive to variations of these parameters. This simply corresponds to computing the stationary values of the log generating functional in the $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$, $\mathbf{S}$ space. This requirement of minimum sensitivity to the variational parameters was introduced in [23] as an approximation protocol.

Using the logic in the previous paragraph and setting the first derivative of the expression for $-\ln Z$ in (28) with respect to $\hat{\eta}_j(t)$ to zero, one gets the equation for stationary $\eta_j(t)$, the first moment of the gaussian form for $g_j(t)$:

$$\eta_i(t) = h_i(t) + \sum_j J_{ij}\mu_i(t), \tag{29}$$

where we have defined for $t = 1, \ldots, T-1$:

$$\mu_i(t) = \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \int \tilde{D}\boldsymbol{\mathcal{G}} \tanh\left[g_i(t-1) + \eta_i(t-1) + \mathrm{i}\psi_i(t) - \mathrm{i}\sum_l J_{li}(\hat{g}_l(t) + \hat{\eta}_l(t)) \right], \tag{30}$$

while for $t = 0$ and $t = T$ we have respectively:

$$\mu_i(0) = \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \int \tilde{D}\boldsymbol{\mathcal{G}} \tanh\left[\mathrm{i}\psi_i(0) - \mathrm{i}\sum_l J_{li}(\hat{g}_l(0) + \hat{\eta}_l(0)) \right], \tag{31a}$$

$$\mu_i(T) = \frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \int \tilde{D}\boldsymbol{\mathcal{G}} \tanh\left[g_i(T-1) + \eta_i(T-1) + \mathrm{i}\psi_i(T) \right]. \tag{31b}$$

Solving $-\partial \ln Z / \partial \eta_i(t) = 0$ gives:

$$\hat{\eta}_i(t) = \mathrm{i}\mu_i(t+1) - \mathrm{i}\frac{\sqrt{\det \mathbf{S}}}{(2\pi)^{NT}} \int \tilde{D}\boldsymbol{\mathcal{G}} \tanh\left[g_i(t) + \eta_i(t) \right]. \tag{32}$$

Looking for the stationary points of (26) with respect to $\mathbf{S}^{-1}$ corresponds to solving the following set of equations:

$$\frac{\partial \ln Z}{\partial S_{ij}^{-1}(t,\,t')} = -\frac{1}{2}S_{ji}(t',\,t) - \mathrm{i}\delta_{ji+N}\,\delta_{tt'} + \frac{\det \mathbf{S}^{1/2}}{(2\pi)^{NT}}\sum_{m,s}\int \tilde{D}\boldsymbol{\mathcal{G}}\left\{\frac{1}{2}\partial_{it,jt'}\ln 2\cosh\left[g_m(s)+\eta_m(s)\right]\right\} - \frac{\det \mathbf{S}^{1/2}}{(2\pi)^{NT}}\sum_{m,s}\int \tilde{D}\boldsymbol{\mathcal{G}}\left\{\frac{1}{2}\partial_{it,jt'}\ln 2\cosh\left[g_m(s)+\eta_m(s) + \mathrm{i}\psi_m(s+1) - \mathrm{i}\sum_l J_{lm}\left(\hat{g}_l(s+1)+\hat{\eta}_l(s+1)\right)\right]\right\} = 0, \tag{33}$$

where we have defined $\partial_{it,jt'} \equiv \frac{\partial^2}{\partial \boldsymbol{\mathcal{G}}_i(t)\partial \boldsymbol{\mathcal{G}}_j(t')}$.

4.2. Equations for the magnetizations

In the previous subsection we derived expressions for the parameters of the gaussian used for perturbative approximation of the log-generating functional at fixed ψ. Now we want to derive an expression for the magnetizations using (6). We will first perform the derivative of (28) with respect to ψ; notice that even $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$ and $\mathbf{S}$ are ψ dependent, such that (6) reads:

$$m_i(t) = -\,\mathrm{i}\lim_{\psi\to 0}\frac{\partial \ln Z}{\partial \psi_i(t)} + \sum_{j,t'}\frac{\partial \eta_j(t')}{\partial \psi_i(t)}\frac{\partial \ln Z}{\partial \eta_j(t')} + \sum_{j,t'}\frac{\partial \hat{\eta}_j(t')}{\partial \psi_i(t)}\frac{\partial \ln Z}{\partial \hat{\eta}_j(t')} + \sum_{l,j,t',t''}\frac{\partial S_{l,j}(t',\,t'')}{\partial \psi_i(t)}\frac{\partial \ln Z}{\partial S_{l,j}(t',\,t'')}. \tag{34}$$

However, since in our optimization scheme we looked for the stationary values of $\ln Z$ with respect to the variational parameters, $\partial \ln Z/\partial \psi$ will only consist of its explicit derivative with respect to ψ, leading to:

$$m_i(t) = \lim_{\psi\to 0}\mu_i(t) \tag{35}$$

for all $t = 0,\ldots,T$ and $\mu_i(t)$ has been defined in (30)–(31*b*).

4.3. The optimized values of the parameters

In principle, one needs to solve the full set of equations (29)–(33) and take the limit of $\boldsymbol{\psi} \to 0$ to calculate the magnetization in (35). This is obviously a very difficult task to do analytically given the high dimensional integrals that appear in (30)–(33) and that the equations have to be solved simultaneously. The solutions, however, can be very much simplified if we assume

$$\lim_{\psi\to 0}\hat{\eta}_i(t) = 0 \tag{36}$$

$\forall\, i,\ \forall\, t$. With (36), which we will justify in section 4.4 below, the optimal interaction matrix $\mathbf{S}$ in (33) in the limit $\boldsymbol{\psi} \to 0$ assumes the following block tridiagonal structure:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}(0,0) & \mathbf{S}(0,1) & 0 & 0 & 0 & \cdots \\ \mathbf{S}(1,0) & \mathbf{S}(1,1) & \mathbf{S}(1,2) & 0 & 0 & \cdots \\ 0 & \mathbf{S}(2,1) & \mathbf{S}(2,2) & \mathbf{S}(2,3) & 0 & \cdots \\ 0 & 0 & \mathbf{S}(3,2) & \mathbf{S}(3,3) & \mathbf{S}(3,4) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \tag{37}$$

where

$$\mathbf{S}(t,t) = \left[\begin{array}{c|c} 0 & -\mathrm{i}\mathbb{I} \\ \hline -\mathrm{i}\mathbb{I} & \boldsymbol{\gamma}(t) \end{array}\right], \quad \mathbf{S}(t,t+1) = \left[\begin{array}{c|c} 0 & \boldsymbol{\lambda}(t,t+1) \\ \hline 0 & 0 \end{array}\right], \tag{38}$$

the blocks $\mathbf{S}(t,t')$ are of size $2N \times 2N$, $\mathbf{S}(t,t+1) = \mathbf{S}(t+1,t)^{\top}$ and

$$\gamma_{ij}(t) = \sum_k J_{ik} J_{jk} (1 - m_k(t)^2) \qquad t = 0, \ldots, T-1, \tag{39a}$$

$$\lambda_{ij}(t,t+1) = \mathrm{i} J_{ji} (1 - m_i(t+1)^2) \qquad t = 0, \ldots, T-2. \tag{39b}$$

Observe that the matrix $\mathbf{S}$ in (37) is a symmetric complex matrix (not Hermitian), whose Hermitian part is positive symmetric. (Recall that the Hermitian part of a matrix $\mathbf{S}$ is defined as $(\mathbf{S} + \mathbf{S}^{\dagger})/2$.) This is consistent with its derivation given that—as pointed out in [24]—the gaussian integral $\int \tilde{D}\boldsymbol{\mathcal{G}}$ converges only if the Hermitian part of $\mathbf{S}$ is a positive symmetric matrix.

In (39*a*) and (39*b*) we implicitly state that $\det \mathbf{S} = 1$: as a matter of fact it can be proven to be a mere consequence of the block structure of the matrix $\mathbf{S}$, as shown in appendix A. Since $\int \tilde{D}\boldsymbol{\mathcal{G}} = (2\pi)^{NT}\sqrt{\det \mathbf{S}}$ this means that the gaussian integral and the model log generating functional match in the limit $\boldsymbol{\psi} \to \mathbf{0}$.

Finally we can substitute the optimal values of the variational parameters in (35) and exploit (32) to get:

$$m_i(t) = \frac{1}{(2\pi)^{NT}} \int \tilde{D}\boldsymbol{\mathcal{G}} \tanh[g_i(t-1) + h_i(t-1) + \sum_j J_{ij} m_j(t-1)] \tag{40}$$

for $t = 1, \ldots, T-1$.

We are now left to evaluate a multidimensional integral in (40). In fact the integration in (40) can be reduced to a one-dimensional integral marginalizing the multivariate gaussian distribution, yielding

$$m_i(t) = \int \frac{\mathrm{d}x \mathrm{e}^{-x^2/2}}{\sqrt{2\pi}} \tanh\left[x\sigma(t) + h_i(t-1) + \sum_j J_{ij} m_j(t-1)\right], \tag{41}$$

$$\sigma(t) = \sqrt{(\mathbf{S}^{-1})_{2N(t-1)+i,2N(t-1)+i}}, \tag{42}$$

where the integral is now over $x = g_i(t) \Big/ \sqrt{(\mathbf{S}^{-1})_{2N(t-1)+i,2N(t-1)+i}}$, a normally distributed, zero mean unit variance, random variable.

For performing the one-dimensional integral in (41), we need to compute the entries of the inverse of matrix $\mathbf{S}$. In appendix B we demonstrate that, given $\mathbf{S}$ as defined in (37)–(39*b*), the entries of $\mathbf{S}^{-1}$ in which we are interested in can be calculated recursively as

$$\mathbf{S}^{-1}_{2Nt+i,2Nt+i} = \tilde{\gamma}_{ii}(t) \quad \text{where} \quad \tilde{\boldsymbol{\gamma}}(t) = \boldsymbol{\gamma}(t) - \boldsymbol{\lambda}(t-1,t)^{\top} \tilde{\boldsymbol{\gamma}}(t-1) \boldsymbol{\lambda}(t-1,t). \tag{43}$$

As we show in appendix B, $\tilde{\gamma}_{ii}(t)$ can only take positive values and therefore the integral in (41) is physically well-defined.

Recalling the definitions of the matrices $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, one can verify that the magnetizations in (41) only depend on the past magnetizations $m_j(t')$ with $t' < t$, $j = 1,\ldots, N$. Since this dependence goes back to $t' = 0$ it is natural to wonder if the error in estimating the past magnetizations would accumulate impairing the inference process. We notice (not included in section 6) that for the gaussian average method knowledge of the history of the experimental magnetization—knowing $\tilde{\gamma}_{ii}(t-1)$ when computing $m_i(t)$ with (41)—does not affects the reconstruction significantly. Whether we are using experimental magnetizations or approximate ones in (43), we observe that $\tilde{\gamma}_{ii}(t-1)$ grows exponentially with time for strong couplings while it converges to a finite value for weak couplings. This behavior can be understood by studying the stability of the map (43) of $\tilde{\boldsymbol{\gamma}}(t-1)$ into $\tilde{\boldsymbol{\gamma}}(t)$ that defines a dynamical system, as we do in appendix C. Averaging over the disorder one realizes that this dynamical system is chaotic for couplings strength above a certain critical value. Its critical value depends on the degree of symmetry of the connectivity and on the presence of an external field.

4.4. *The solution* $\lim_{\psi\to 0}\hat{\eta}_i(t) = 0$

In principle, the value of limit of $\psi \to 0$ of $\hat{\eta}_i(t)$ that satisfy the optimality equations, may be non-zero. In this section, we justify the choice of $\lim_{\psi\to 0}\hat{\eta}_i(t) = 0$ that we made in the previous section. We first note that zero is a good candidate for the optimal value of $\hat{\eta}_i(t) = \langle \hat{g}_i(t)\rangle_{L_s}$—here $\langle\ldots\rangle_{L_s}$ indicates the average under the complex measure e^{-L_s} in (27*a*) and (27*b*)—since

$$\lim_{\psi\to 0}\langle \hat{g}_i(t)\rangle_L = 0, \tag{44}$$

where L for the kinetic Ising model has been defined in (9) and $\langle\ldots\rangle_L$ indicates the average under the complex measure e^{-L}. This choice for the mean in the $\hat{g}$s can be justified by analogy with the mean in the gs: the stationary value for latter is also the saddle point value of the kinetic Ising generating functional, while the saddle point in the $\hat{g}$s is conventionally set to zero.

Furthermore, we can show that $\lim_{\psi\to 0}\hat{\eta}_i(t) = 0$ yields a consistent solution. To do this we first note that by inverting the matrix $\mathbf{S}$, as shown in appendix B, two point correlation functions $\langle g_i(t-1)\hat{g}_j(t)\rangle_{L_s'}$ and $\langle \hat{g}_i(t)\hat{g}_j(t)\rangle_{L_s'}$ are both zero, where notation $\langle\ldots\rangle_{L_s'}$ indicates averages under the gaussian measure $\mathrm{e}^{-L_s'}$, with $L_s' = \frac{1}{2}\boldsymbol{\mathcal{G}}\mathbf{S}\boldsymbol{\mathcal{G}}$. Consequently, we have

$$\begin{aligned}
\lim_{\psi\to 0}\mu_i(t) &= \left\langle \tanh\left[g_i(t-1) + \eta_i(t-1) - \mathrm{i}\sum_l J_{li}\hat{g}_l(t)\right]\right\rangle_{L_s'} \\
&= \sum_{n=1}^{\infty} a_n \left\langle \left[g_i(t-1) + \eta_i(t-1) - \mathrm{i}\sum_l J_{li}\hat{g}_l(t)\right]^{2n-1}\right\rangle_{L_s'} \\
&= \sum_{n,k,l} b_{n,k,l}(\eta_i(t-1))^{k-l}\left\langle g_i(t-1)^l\left(-\mathrm{i}\sum_l J_{li}\hat{g}_l(t)\right)^{2n-1-k}\right\rangle_{L_s'} \\
&= \sum_{n,k,l} b_{n,k,l}\delta_{k,2n-1}(\eta_i(t-1))^{k-l}\langle g_i(t-1)^l\rangle_{L_s'} \\
&= \langle \tanh\left[g_i(t-1) + \eta_i(t-1)\right]\rangle_{L_s'}.
\end{aligned} \tag{45}$$

Now, note that the previous equality corresponds to setting $\hat{\eta}_i(t) = 0$ in (32).

4.5. The fully asymmetric limit

In [1] Mezard and Sakellariou derive equations for the magnetizations that are exact for fully asymmetric couplings:

$$m_i(t) = \int \frac{\mathrm{d}x}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \tanh\left[h_i(t-1) + \sum_j J_{ij} m_j(t-1) + x\sqrt{\gamma_{ii}(t-1)} \right], \tag{46}$$

where $\gamma_{ii}(t)$ has been defined in (39*a*).

In section 4.1 all entries of $\mathbf{S}$ were free to be optimized. However, we could have assumed that the blocks corresponding to $\boldsymbol{\lambda}(t-1, t)$ are set to zero *a priori*. By looking at (41) and (43) one easily realizes that with this constraint our optimization would have lead to (46), which is exact in the fully asymmetric limit. Notice that this prescription on $\boldsymbol{\lambda}$ would not affect the optimal value of any other variational parameter, since we optimized $\ln Z$ independently with respect of $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$ or $\mathbf{S}_{ij}(t, t')$.

5. Extended Plefka expansion

As mentioned in the Introduction, two particularly powerful approaches to studying disordered systems both in machine learning and statistical physics community are variational and weak coupling expansions. In the previous sections we reported some results regarding the variational approach. In this section we aim at developing a comprehensive weak coupling expansion for the disordered spin systems.

Weak coupling expansions in field theory and statistical physics of disordered systems take several forms. One of the most powerful amongst these, which has proven to be particularly useful for studying the equilibrium properties of glassy systems, is the Plefka expansion. The Plefka expansion was originally performed for the equilibrium Sherrington–Kirckpatrick model by expanding the Gibbs free energy at fixed magnetization, enforced via a Legendre transform, around the free energy of an uncoupled system. To the first order in J it yields the naive mean field results while to the second order the TAP equations are recovered. Although higher order terms vanish for the SK model, they can in general be computed [25].

In performing the Plefka expansion for the equilibrium model with binary spins it is sufficient to fix the magnetization and this has been the line taken by Roudi and Hertz in deriving Plefka expansion and dynamical TAP equations for the kinetic Ising model. However, in contrast to the equilibrium case, for the dynamics the magnetization is not the only relevant order parameter. Including other observables in the Plefka expansion, namely the correlation and response functions, is what we do in this section. As we will show with numerical results in the next section, this will lead to a significant improvement for predicting the dynamics of the system.

Instead of the generating functional in (4) and (5), let us now consider the following functional:

$$Z_\alpha[\boldsymbol{\psi}, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}] = \frac{1}{2^N (2\pi)^{NT}} \int D\mathcal{G}\, \mathrm{Tr}_{\mathbf{s}} \exp\left(L_\alpha[\boldsymbol{\psi}, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}, \mathcal{G}, \mathbf{s}]\right), \tag{47}$$

with

$$L_\alpha[\psi, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}, \mathcal{G}, \mathbf{s}] = \sum_{i,t}\Bigg\{ \mathrm{i}\hat{g}_i(t)\Bigg[g_i(t) - \alpha\sum_j J_{ij}s_j(t)\Bigg] + s_i(t+1)g_i(t) \\ - \ln 2\cosh g_i(t) - \mathrm{i}h_i(t)\hat{g}_i(t) + \psi_i(t)s_i(t) \\ + \frac{1}{2}\sum_{t'}\hat{C}_i(t, t')s_i(t)s_i(t') \\ + \frac{1}{2}\sum_{t'}\hat{B}_i(t, t')\hat{g}_i(t)\hat{g}_i(t') - \mathrm{i}\sum_{t'}\hat{R}_i(t', t)\hat{g}_i(t)s_i(t')\Bigg\}, \tag{48}$$

where we have introduced the parameter α to control the interaction strength. The introduction of the new auxiliary fields $\hat{C}$, $\hat{B}$ and $\hat{R}$ in the action (48) is related to the averages of the observables that we want to constrain when performing the Legendre transform. In particular, here we decide to fix all marginal first and second moments over time. One can find the moments and the physical meaning of these auxiliary fields by first derivatives of the generating functional with respect to the fields as follows:

$$\begin{aligned} -\mathrm{i}\hat{m}_i(t) &= \frac{\partial \ln Z_\alpha}{\partial h_i(t)} = -\mathrm{i}\langle\hat{g}_i(t)\rangle_\alpha \\ m_i(t) &= \frac{\partial \ln Z_\alpha}{\partial \psi_i(t)} = \langle s_i(t)\rangle_\alpha \\ C_i(t, t') &= \frac{\partial \ln Z_\alpha}{\partial \hat{C}_i(t, t')} = \langle s_i(t)s_i(t')\rangle_\alpha \quad \text{for } t' \neq t \\ \frac{1}{2}B_i(t, t') &= \frac{\partial \ln Z_\alpha}{\partial \hat{B}_i(t, t')} = \frac{1}{2}\langle\hat{g}_i(t)\hat{g}_i(t')\rangle_\alpha \\ R_i(t, t') &= \frac{\partial \ln Z_\alpha}{\partial \hat{R}_i(t, t')} = -\mathrm{i}\langle\hat{g}_i(t')s_i(t)\rangle_\alpha, \end{aligned} \tag{49}$$

where $\langle\cdots\rangle_\alpha$ denotes averaging over the distribution defined by the measure inside the functional (47). Namely, for any function $F(\mathbf{s})$ of the trajectory of spins $\mathbf{s}$ we define:

$$\langle F\rangle_\alpha = \frac{\int D\mathcal{G}\,\mathrm{Tr}_s\, F(\mathbf{s})\,\exp\left(L_\alpha[\psi, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}, \mathcal{G}, \mathbf{s}]\right)}{\int D\mathcal{G}\,\mathrm{Tr}_s \exp\left(L_\alpha[\psi, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}, \mathcal{G}, \mathbf{s}]\right)}. \tag{50}$$

The moments of the original dynamical system (2) can be found by setting the auxiliary fields to zero and $\alpha = 1$ at the end of the calculation. Note that $C(t, t) = 1$ and its conjugate field $\hat{C}(t, t) = 0$.

The Legendre transform of $\ln Z$ is given by

$$\Gamma_\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}] = \ln Z_\alpha[\psi, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}] - \sum_{it}\psi_i(t)m_i(t) + \mathrm{i}\sum_{it}h_i(t)\hat{m}_i(t) \\ - \frac{1}{2}\sum_{itt'}\hat{C}_i(t, t')C_i(t, t') - \frac{1}{2}\sum_{itt'}\hat{B}_i(t, t')B_i(t, t') - \sum_{itt'}\hat{R}_i(t', t)R_i(t', t), \tag{51}$$

where the fields ψ, $\boldsymbol{h}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{R}}$ in the above equation are to be considered as functions of the moments, and dependent on the α parameter, according to the following set of equations:

$$\frac{\partial \Gamma_\alpha}{\partial m_i(t)} = -\psi_i^\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}](t)$$
$$\frac{\partial \Gamma_\alpha}{\partial \hat{m}_i(t)} = \mathrm{i}h_i^\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}](t)$$
$$\frac{\partial \Gamma_\alpha}{\partial C_i(t, t')} = -\hat{C}_i^\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}](t, t') \quad \text{for } t' \neq t$$
$$\frac{\partial \Gamma_\alpha}{\partial B_i(t, t')} = -\frac{1}{2}\hat{B}_i^\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}](t, t')$$
$$\frac{\partial \Gamma_\alpha}{\partial R_i(t, t')} = -\hat{R}_i^\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}](t, t'). \tag{52}$$

We now perform a second order expansion of Γ_α around $\alpha = 0$ and consider the set of equations (52) within the expansion; the details of the calculation are reported in appendix D. Setting the auxiliary fields to zero, we can extract the value of the fields $\boldsymbol{\psi}^0$, $\boldsymbol{h}^0$, $\hat{\mathbf{C}}^0$, $\hat{\mathbf{B}}^0$, $\hat{\mathbf{R}}^0$ as functions of the correct (within the expansion) marginal first and second moments. Those fields thus represent effective external fields which have to be applied to the model *without* interactions ($\alpha = 0$) to obtain the same moments as the interacting model. Hence, we may consider $Z_0[\boldsymbol{\psi}^0, \boldsymbol{h}^0, \hat{\mathbf{C}}^0, \hat{\mathbf{B}}^0, \hat{\mathbf{R}}^0]$ as the generating functional for the true marginal distributions, giving us an effective non-interacting description of the true interacting dynamics. The explicit calculation (appendix D) yields $Z^0[h] = \prod_i Z_i^0[h_i]$, where

$$Z_i^0 \propto \left\langle \int \mathrm{d}\mathbf{g}_i \, \mathrm{Tr}_{s_i} \prod_t \frac{\mathrm{e}^{s_i(t+1)g_i(t)}}{2\cosh g_i(t)} \prod_t \delta\Big[g_i(t) - \phi_i(t) - \sum_j \Big(J_{ij} m_j(t) - \sum_{t'=0}^{t-1} J_{ij} J_{ji} R_j(t, t')[s_i(t') - m_i(t')] \Big) - h_i(t) \Big] \right\rangle_{\phi_i} \tag{53}$$

and where $\phi_i(t)$ is a gaussian random variables, drawn independetly for each i, with zero mean and covariance

$$\langle \phi_i(t)\phi_i(t') \rangle = \sum_j J_{ij}^2 [C_j(t, t') - m_j(t)m_j(t')]. \tag{54}$$

This corresponds to a stochastic equation for a single spin, where each spin i is subjected to an effective field

$$g_i(t) = \phi_i(t) + \sum_j \Big(J_{ij} m_j(t) - \sum_{t'=0}^{t-1} J_{ij} J_{ji} R_j(t, t')[s_i(t') - m_i(t')] \Big) + h_i(t). \tag{55}$$

The effective field in (55) is composed of a coloured gaussian noise (ϕ), a naive mean field (the second term), a retarded interaction with the past values of the spins (third term) and finally the external field ($h_i(t)$).

The retarded interactions and the noise covariance have to be computed as averages from the entire ensemble of independent spins. Luckily, this can be done in a causal fashion, i.e. the spin dynamics depends only on *past* spin history. However, this can not be done analytically, although one may proceed again with a perturbation expansions in order to get equation of motions for one and two time functions. The fact that the external noise is gaussian should be helpful. As an alternative, we have resorted to numerical simulations, where the necessary averages are estimated from a large number N_T of samples of trajectories. Sample averages

will be denoted by overbars; namely, for any function $F^k(\mathbf{s}^k)$ of the kth trajectory of spins $\mathbf{s}^k$ we define the following average:

$$\overline{F} = \frac{1}{N_T}\sum_{k=1}^{N_T} F^k. \tag{56}$$

In order to compute the retarded interaction $R_i(t, t')$, we recall that given a vector $\boldsymbol{\phi}$ with gaussian distributed components $\phi(t)$, with zero mean and covariance matrix $\langle \phi(t)\phi(t')\rangle = \mathcal{C}(t, t')$, and given a function $F(\boldsymbol{\phi})$ of the vector $\boldsymbol{\phi}$, the following relation holds

$$\langle F(\boldsymbol{\phi})\phi(t)\rangle = \sum_{t'} \mathcal{C}(t, t')\left\langle \frac{\partial}{\partial \phi(t')} F(\boldsymbol{\phi})\right\rangle, \tag{57}$$

as can be shown using integration by parts. By considering the function $F(\boldsymbol{\phi}) = s(t) \equiv s(t; \boldsymbol{\phi})$ and using (D.19) one finds the following equation relating the response and correlation functions:

$$\langle s_i(t)\phi_i(t')\rangle_{\phi_i} = \sum_{\tau=1}^{t-1} R_i(t, \tau)\sum_j J_{ij}^2 [C_j(\tau, t') - m_j(\tau)m_j(t')]. \tag{58}$$

The algorithm can be described as follows.

- Initial condition: set $s_i^k(0) = 1, \quad i = 1 \ldots N, k = 1 \ldots N_T$.
- For $t = 1 \ldots T$:
  - (i) Draw the spins at time t from

$$p(s_i^k(t)) = \frac{\mathrm{e}^{s_i^k(t) g_i^k(t-1)}}{2\cosh g_i^k(t-1)}, \quad \text{for } i = 1 \ldots N, k = 1 \ldots N_T,$$

    using the fields $g_i^k(t-1)$ calculated at the previous time step.
  - (ii) Compute the sample averages

$$C_i(t, t') = \overline{\tanh[g_i(t-1)]s_i(t')}, \quad \text{for } t' = 1 \ldots t-1, i = 1 \ldots N.$$

  - (iii) Draw the noise variables $\phi_i^k(t)$ for $i = 1 \ldots N, k = 1 \ldots N_T$, from the conditional probability $p(\phi_i^k(t)|\phi_i^k(0) \ldots \phi_i^k(t-1))$, which can be computed using the Yule Walker equations (appendix E).
  - (iv) Compute the sample averages that will be needed in (v):

$$\overline{s_i(t)\phi_i(t')} = \overline{\tanh[g_i(t-1)]\phi_i(t')}, \quad \text{for } t' = 1 \ldots t-1, i = 1 \ldots N.$$

  - (v) Compute $R_i(t, t')$, for $t' = 1 \ldots t-1$ using (58) by solving the system of linear equations:

$$\overline{s_i(t)\phi_i(t')} = \sum_{\tau=1}^{t-1} R_i(t, \tau)\sum_j J_{ij}^2 [C_j(\tau, t') - m_j(\tau)m_j(t')].$$

(vi) Compute the fields

$$g_i^k(t) = \phi_i^k(t) + \sum_j \left(J_{ij} m_j(t) - \sum_{t'=0}^{t-1} J_{ij} J_{ji} R_j(t, t') [s_i^k(t') - m_i(t')] \right) + h_i(t),$$
$$\text{for } i = 1 \dots N,\ k = 1 \dots N_T.$$

(vii) Compute the magnetizations at time $t + 1$:

$$m_i(t+1) = \overline{\tanh[g_i(t)]}, \quad \text{for } i = 1 \dots N.$$

To conclude this section, let us point out that the mean field result (46), which is exact for asymmetric networks in the thermodynamic limit for gaussian couplings with variance $1/N$, can be obtained in two ways. One either considers the result (54) and neglects the term $J_{ij}J_{ji}$ for an asymmetric network in the limit of large N, or one works with a simplified Plefka expansion where all two-time moments for different times are excluded from the beginning. Hence, from the second moments, one keeps only $B(t, t)$ in the expansion.

6. Numerical results

In the previous sections we studied analytically two approaches to improve on the saddle point approximation to the generating functional of the kinetic Ising model with synchronous update. In section 4.5 we have argued that the constrained gaussian average optimization leads to the mean field (MS-MF) equations of [1], whose performances was studied in [26]. One could wonder how this compares to the unconstrained gaussian average method, and so we iterated (41) and (43) to reconstruct the entire dynamics of magnetizations. In order to estimate the magnetizations for the extended Plefka expansion described in the previous section we designed the algorithm explained in section 5. Thus we can evaluate numerically the goodness of the two approximations in terms of magnetizations and compare them with existing algorithms. Specifically we investigate how they perform with respect to three mean field methods, namely Naive mean field, dynamical TAP (RH-TAP) equations of [2] and MS-MF equations of [1]. To recapitulate, Naïve mean field and TAP equations can be obtained via perturbative expansion in the magnitude of the couplings of the Legendre transform of the log generating functional at fixed magnetizations [2], without making any restriction on symmetry and distribution of the couplings. The first order expansion gives Naïve Mean Field, while second order terms lead to RH-TAP. MS-MF equations can be derived via central limit theorem arguments exploiting the fact that the couplings are independent identically distributed random variables with variance that scales as $1/N$ [1], without making any assumption on the couplings strength.

RH-TAP magnetizations under the kinetic Ising model with synchronous update are:

$$m_i(t) = \tanh\left[h_i(t-1) + \sum_j J_{ij} m_j(t-1) - m_i(t) \gamma_{ii}(t-1) \right], \tag{59}$$

where $\gamma_{ii}(t-1)$ has been defined in (39*a*). MS-MF equations correspond to (46) and Naïve mean field to (25).

In order to test the performances of our methods as a function of couplings asymmetry and strength we chose our couplings, following Crisanti and Sompolinsky [3]:

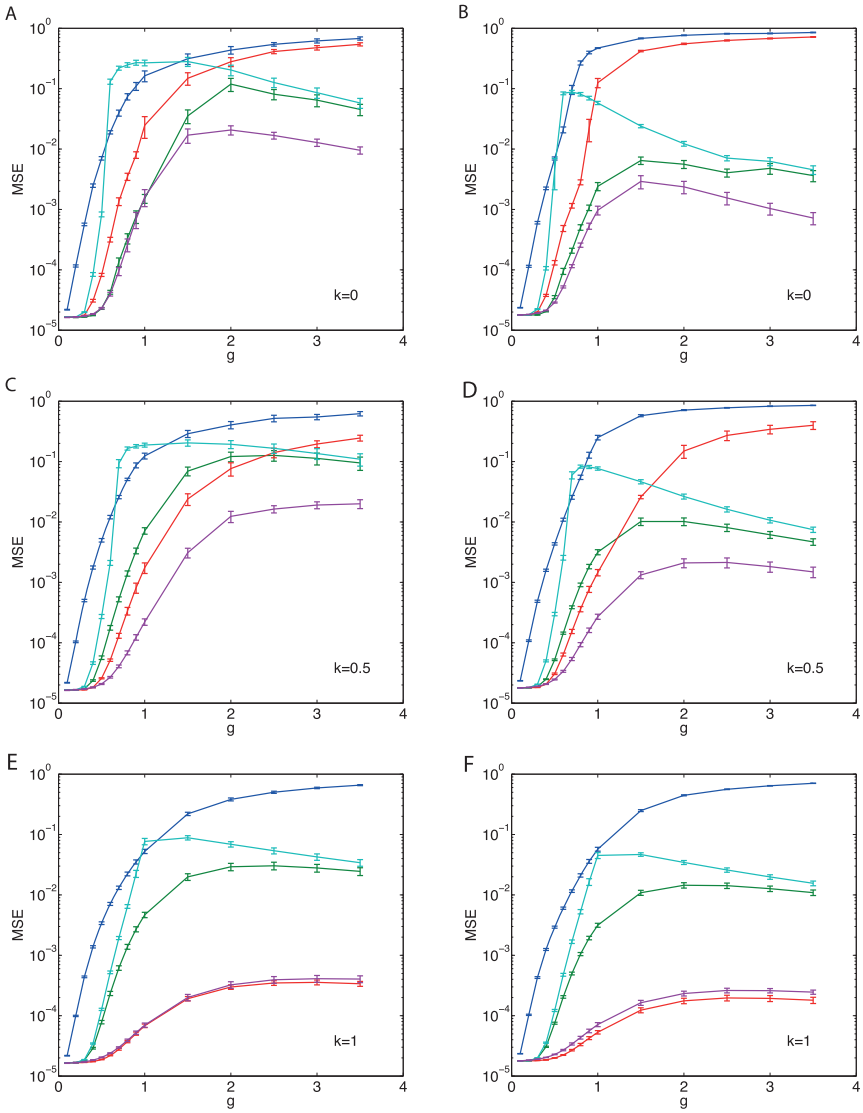

Figure 1. Mean squared error of Näive mean field (blue), RH-TAP (green), MS-MF (red), unconstrained gaussian average approach (light blue) and extended Plefka (magenta) for predicting entire dynamics of the magnetizations. The mean squared error is plotted as a function of the couplings strength g for a system of 50 spins. We have used 100 time steps and 50 000 repeats to calculate the experimental magnetizations and have averaged the errors over 10 realizations of the couplings. The error bars are standard deviations over these realizations. The number of sample trajectories used in the algorithm for the extended Plefka method is $N_T = 50\,000$. The different panels correspond to different values of the asymmetry parameter $k = 0, 0.5, 1$ from top to bottom. Left: stationary external field drawn independently for each spin from a normal distribution (zero mean, standard deviation 0.5). Right: sinusoidal external field with amplitude 0.5.

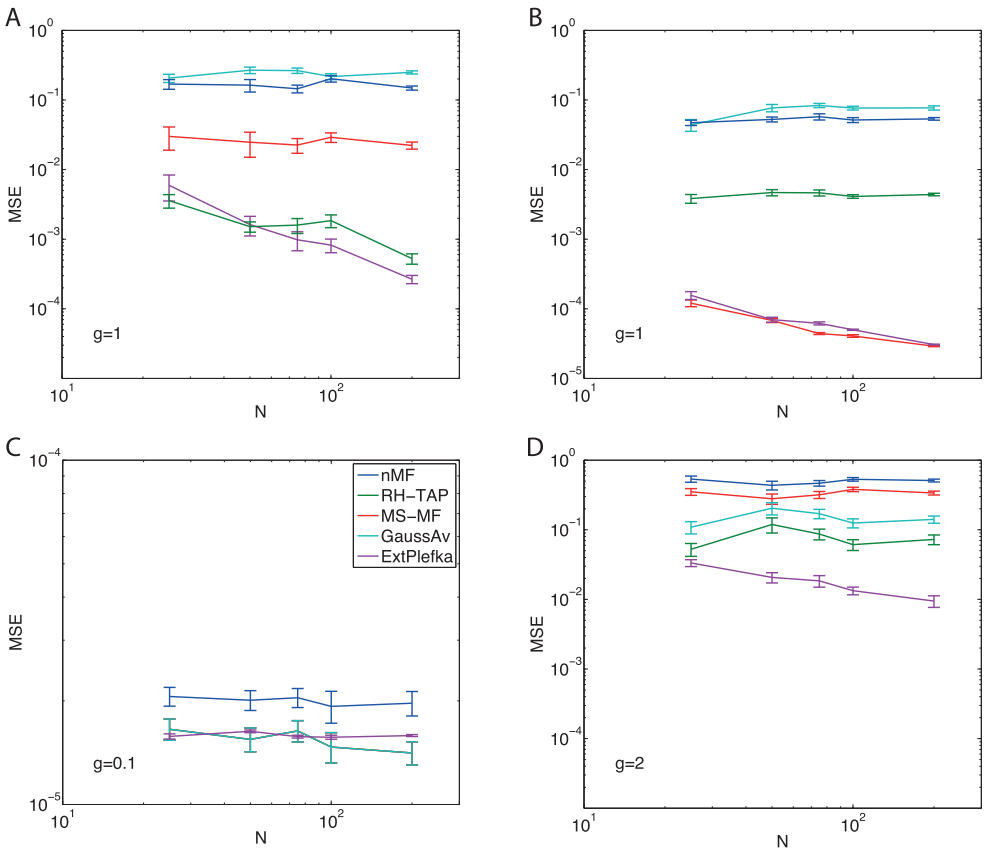

Figure 2. Mean squared error of Naive mean field (blue), RH-TAP (green), MS-MF (red), unconstrained gaussian average approach (light blue) and extended Plefka (magenta) for predicting entire dynamics of magnetizations. The mean squared error is plotted as a function of the system size N. We have used 100 time steps and 50 000 repeats to calculate the experimental magnetizations and have averaged the errors over 10 realizations of the couplings. The error bars are standard deviations over these realizations. The number of sample trajectories used in the algorithm for the extended Plefka method is $N_T = 50\,000$ for $N = 25, 75, 100$, while $N_T = 100\,000$ for $N = 200$. Stationary external field drawn independently for each spin from a normal distribution (zero mean, standard deviation 0.5). (A): symmetric case $k = 0$, couplings strength $g = 1$. (B): fully asymmetric case $k = 1$, $g = 1$. (C): $k = 0$, $g = 0.1$; unconstrained gaussian average, MS-MF and RH-TAP curves overlap. (D): $k = 0$, $g = 2$. Notice the different scale on the y-axis in (C) with respect to the other panels.

$$J_{ij} = J_{ij}^{\text{sym}} + kJ_{ij}^{\text{antisym}}, \tag{60}$$

where $J_{ij}^{\text{sym}} = J_{ji}^{\text{sym}}$ and $J_{ij}^{\text{antisym}} = -J_{ji}^{\text{antisym}}$, while k is the parameter that controls the asymmetry, interpolating between the fully asymmetric ($k = 1$) and the fully symmetric ($k = 0$) distributions. We draw all the couplings J_{ij}^{sym} and J_{ij}^{antisym} independently from a distribution with zero mean and variance:

$$\langle (J_{ij}^{\text{sym}})^2 \rangle = \langle (J_{ij}^{\text{antisym}})^2 \rangle = \frac{g^2}{N(1+k^2)}, \tag{61}$$

where g controls the strength of the couplings.

We initialize the algorithms with the same initial condition and then we iterate them for reconstructing the whole dynamics of magnetizations. We compare the predicted magnetizations with the experimental ones computing the mean square errors:

$$\text{MSE} = \frac{1}{T}\frac{1}{N}\left\langle \sum_{i=1}^{N}\sum_{t=1}^{T}(m_i(t) - m_i^{\text{exp}}(t))^2 \right\rangle_{J \text{ realizations}}, \tag{62}$$

where $m_i^{\text{exp}}(t)$ are obtained by sampling the kinetic Ising model distribution of (8).

The results are shown in figure 1. From these plots it is clear that, apart from Naïve Mean field, all methods that we considered are compatible in the high temperature limit. At lower temperatures the extended Plefka expansion is superior independently of the external field. Note, however, that for fully asymmetric couplings and sinusoidal external field the MS-MF method is performing slightly better than the extended Plefka approximation. This is likely due to the finite size effects, since the two approaches are equivalent for asymmetric networks with large N, as explained in section 5. Regardless the degree of symmetry of the couplings and the external field RH-TAP systematically improves on the unconstrained gaussian average approach, which fails at intermediate temperatures. The fact that the reconstruction is noisier with respect to [26] is due to error propagation during the dynamics.

The scaling of the MSE errors with N is shown in figure 2. Numerical simulations show that the error of the extended Plefka method decays with the system size N for every value of the parameters g and k, while the errors of the RH-TAP and MS-MF approximations decrease with N only in the range of the parameters for which the the approximations were developed, which corresponds respectively to a symmetric network with small couplings and to an asymmetric network. The error computed using Naïve mean field and unconstrained gaussian average approximations shows no scaling with N. This seems to suggest that the extended Plefka expansion provides an accurate mean field description of the dynamics. Notice however that evaluating the local moments with grater accuracy requires considering the whole history of the single spin trajectory and that the complexity of the algorithm described in section 5 scales with the degrees of freedom as $T^2N + TNN_T$. To speed up the algorithm one could argue that, when the couplings J_{ij} scale as $1/\sqrt{N}$, the two sums

$$\sum_j J_{ij}^2 C_j(t, t'); \quad \sum_j J_{ij}J_{ji}R_j(t, t')$$

appearing in (54) and (55) can be replaced by their self-averaging value

$$g^2 \sum_j C_j(t, t'); \quad g^2 \frac{1-k^2}{1+k^2}\sum_j R_j(t, t'),$$

where we considered the distribution (61) for the couplings. This would allow us to write a self-averaging version of (58) and the computational cost of the algorithm would reduce to $T^2 + TNN_T$. We postpone this analysis to future work.

7. Summary and discussion

In this paper we studied new approximations for predicting the dynamics of the kinetic Ising model with arbitrary couplings. First we distinguished between the variational and field theoretical approaches to the Naïve mean field theory for a generic Markov chain and pointed

out that the two do not coincide unless the transfer function is logistic, as is the case for the kinetic Ising model and there are no self-interactions in the system. (For an overview of the approaches to Naïve mean field theory for the equilibrium case see [16].) For the specific case of the kinetic Ising model with discrete time parallel updates, we then proposed two approximations based on generating functional integral technique: the gaussian average Variational method and the extended Plefka expansion. In the gaussian average variational method we expand the generating functional of the process to first order around a high dimensional complex gaussian integral and optimize the resulting expression. An unconstrained optimization of the parameters of this gaussian function, in which we assume no structure for the covariance matrix, provides equations of motion which, as our numerical analysis indicates, perform at the naive mean field level for small couplings while they get close to RH-TAP equations [2] for larger couplings. On the other hand making suitable assumptions on the covariance matrix, allows us to recover the MS-MF equations [1], known to be exact for fully asymmetric connectivities in the thermodynamic limit.

Although we numerically compared the dynamics of magnetizations predicted from our dynamical equations with those of simulating the system, we did not study the relaxation dynamics that our dynamical equations predict for such systems analytically. Such an analysis has been performed in the case of the p-spin spherical spin glass model in [10] where it is shown that the long term dynamics of the dynamical TAP equations for this system can be seen as descending through the free energy landscape. For symmetric couplings and constant external fields, the synchronous update model that we have considered here in the long time will equilibrate to a Boltzmann distribution determined by the Peretto's Hamiltonian [27]. The replica analysis for this model has not been performed and, therefore, we cannot make any statements as to what degree our extended Plefka and variational equations will be in agreement with such analysis. However, we would like to note that for the asynchronous update Glauber dynamics, once stationary magnetizations are assumed, the RH-TAP equations will coincide with the standard static TAP equations [28]. As noted before the equations derived here using the extended Plefka expansion are generalization of the RH-TAP equations [2] and reduce to those if correlations and response functions are not taken into account. Static TAP equations—which can be derived as the stationary limit of RH-TAP equations—in turn, describe the multitude of local minima observed in the low temperature phase of the SK model, and whose consistency with the replica approach has been formally established [29]. The situation regarding the variational method is less clear, for one reason because, besides that little is known of the low temperature properties of the Peretto's Hamiltonian [30], the resulting dynamical equations can change with the ansatz chosen for the covariance matrix of the fields and conjugate fields. If no ansatz is assumed, our numerical results show that at low temperatures (strong couplings), the error in predicting the magnetizations approaches those of the RH-TAP equations, which as stated before lead to static TAP equations in the stationary state. We will leave it to future studies to explore this similarity and the relaxation dynamics predicted by the variational approach in more detail and analytically.

In the extended Plefka approach, by expanding the log generating functional in the coupling strength, while fixing first and second order moments over time, we approximate the true interacting dynamics by an effective single site dynamics. Namely, within the approximated description, each spin is subjected to an effective local field (55) that contains a retarded interaction with its own past values and a coloured gaussian noise. The main difference with other mean field techniques is that the whole history of the single spin trajectory is taken into account in the equation for local order parameters. Numerical simulations show that considering this term leads to greater accuracy in predicting local magnetizations for all

values of couplings strength, coupling asymmetry and different choices of external fields. We find that this memory term is stronger for larger degree of symmetry of the network, and negligible when the couplings are uncorrelated: in this case the MS-MF approximation is retrieved.

The methods proposed in this paper are quite general in their scope and in theory can be used for studying the dynamics of other kinetic models. In particular, we find it interesting to see how these approximations perform for point process models from the generalized linear model family, from which the kinetic Ising model is just one simple example. Furthermore, these methods can also be applied for inverse problems: inferring the interactions and fields given spin trajectories [31]. In particular, given the fact that inference and learning in the presence of hidden nodes can be casted in a functional integral language [32], our methods can naturally lend themselves to developing novel approximations in this case for point processes. In fact, very recently, the extended Plefka approach has been used for learning and inference of the continuous variables in the presence of hidden nodes [33].

Acknowledgments

This work has been partially supported by the Marie Curie Initial Training Network NETADIS (FP7, grant 290038). YR and CB also acknowledge fundings from the Kavli Foundation and the Norwegian Research Council Centre of Excellence scheme. YR is also grateful to the Starr foundation for financing his membership at the IAS.

Appendix A. Determinant of S

As was mentioned in section 4.1 of the main text, in this appendix we demonstrate that the determinant of the matrix $\mathbf{S}$ that appears in (37) and (38) equals one. We are going to prove it irrespectively to the specific details of matrices $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ in (38). Consider a complex matrix $\mathbf{S}$ with the block structure defined in (37), where $\boldsymbol{\lambda}(t, t+1)$ and $\boldsymbol{\gamma}(t)$ are generic complex square matrices of order N.

In order to compute its determinant partition the matrix $\mathbf{S}$ as follows:

$$\mathbf{S} = \left[\begin{array}{c|cccccc} \mathbf{S}(0,0) & \mathbf{S}(0,1) & 0 & 0 & 0 & \cdots \\ \hline \mathbf{S}(1,0) & \mathbf{S}(1,1) & \mathbf{S}(1,2) & 0 & 0 & \cdots \\ 0 & \mathbf{S}(2,1) & \mathbf{S}(2,2) & \mathbf{S}(2,3) & 0 & \cdots \\ 0 & 0 & \mathbf{S}(3,2) & \mathbf{S}(3,3) & \mathbf{S}(3,4) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right]. \tag{A.1}$$

The determinant of this partitioned matrix can be formulated in terms of its blocks through the properties of Shur complements. Indeed for a generic matrix M:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad \det M = \det A \det [D - CA^{-1}B]. \tag{A.2}$$

Since the square matrix $\mathbf{S}(0,0)$ is invertible, as can be easily checked in (38) and (A.2) can be used to express the determinant of $\mathbf{S}$ as:

$$\det \mathbf{S} = \det \mathbf{S}(0,\,0) \det \left[\underbrace{\mathbf{S}^{\setminus 0} - \begin{pmatrix} \mathbf{S}(1,\,0) \\ 0 \\ \vdots \end{pmatrix} \mathbf{S}(0,0)^{-1} \begin{pmatrix} \mathbf{S}(1,\,0) & 0 & \ldots \end{pmatrix}}_{\tilde{\mathbf{S}}^{\setminus 0}} \right], \tag{A.3}$$

where we have denoted with $\mathbf{S}^{\setminus 0}$ the bottom right matrix in the partition (A.1).

Notice that second term in $\tilde{\mathbf{S}}^{\setminus 0}$—the Shur complement of $\mathbf{S}(0,\,0)$—has the form:

$$\begin{aligned} &\begin{pmatrix} \mathbf{S}(1,\,0) \\ 0 \\ \vdots \end{pmatrix} \mathbf{S}(0,0)^{-1} \begin{pmatrix} \mathbf{S}(1,\,0) & 0 & \ldots \end{pmatrix} \\ &= \left[\begin{array}{c|cccc} \hat{\mathbf{S}}(1,\,1) & 0 & 0 & 0 & \cdots \\ \hline 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right] \end{aligned} \tag{A.4}$$

with

$$\hat{\mathbf{S}}(1,\,1) = \left[\begin{array}{c|c} 0 & -\mathrm{i}\mathbb{I} \\ \hline -\mathrm{i}\mathbb{I} & \boldsymbol{\rho}(1) \end{array}\right], \qquad \boldsymbol{\rho}(1) = \boldsymbol{\lambda}(0,1)^{\top} \boldsymbol{\gamma}(0) \boldsymbol{\lambda}(0,\,1) \tag{A.5}$$

such that the matrix $\tilde{\mathbf{S}}^{\setminus 0}$ in (A.3) turns out having the same block form as $\mathbf{S}^{\setminus 0}$,

$$\tilde{\mathbf{S}}^{\setminus 0} = \left[\begin{array}{c|ccccc} \tilde{\mathbf{S}}(1,\,1) & \mathbf{S}(1,\,2) & 0 & 0 & 0 & \cdots \\ \hline \mathbf{S}(2,\,1) & \mathbf{S}(2,\,2) & \mathbf{S}(2,\,3) & 0 & 0 & \cdots \\ 0 & \mathbf{S}(3,\,2) & \mathbf{S}(3,\,3) & \mathbf{S}(3,\,4) & 0 & \cdots \\ 0 & 0 & \mathbf{S}(4,\,3) & \mathbf{S}(4,\,4) & \mathbf{S}(4,\,5) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right], \tag{A.6}$$

$$\tilde{\mathbf{S}}(t,\,t) = \left[\begin{array}{c|c} 0 & -\mathrm{i}\mathbb{I} \\ \hline -\mathrm{i}\mathbb{I} & \tilde{\boldsymbol{\gamma}}(t) \end{array}\right] \tag{A.7}$$

and

$$\tilde{\boldsymbol{\gamma}}(t) = \boldsymbol{\gamma}(t) - \boldsymbol{\lambda}(t-1,\,t)^{\top} \tilde{\boldsymbol{\gamma}}(t-1) \boldsymbol{\lambda}(t-1,\,t). \tag{A.8}$$

As a consequence $\tilde{\mathbf{S}}^{\setminus 0}$ is a block tridiagonal matrix, just like $\mathbf{S}$, and in order to compute its determinant one can apply (A.3) again, to express $\det \tilde{\mathbf{S}}^{\setminus 0}$ as a function of the determinant of $\tilde{\mathbf{S}}(1,\,1)$. By repeatedly applying (A.3) to the Shur complements $\tilde{\mathbf{S}}^{\setminus t}$ of $\tilde{\mathbf{S}}(t,\,t)$, one shows that the determinant of $\mathbf{S}$ can be factorized into determinants of $\tilde{\mathbf{S}}(t,\,t)$ s. As proven for $t = 0$ these matrices $\tilde{\mathbf{S}}(t,\,t)$ preserve the structure of $\mathbf{S}(t,\,t)$ and therefore their determinants are 1. Finally:

$$\det \mathbf{S} = \prod_{t} \det \tilde{\mathbf{S}}(t,\,t) = 1. \tag{A.9}$$

Appendix B. Inverse of S

In sections 4.2 and 4.4 we relate the optimal values of the variationl parameters and the magnetizations to the elements of the covariance matrix $\mathbf{S}^{-1}$ in the framework of the gaussian average method. In this appendix we derive expressions for these elements, namely the correlations between field and conjugate fields $\mathbf{S}^{-1}_{2Nt+i,2Nt+i} = \langle g_i(t)^2\rangle_{L_s'}$, $\mathbf{S}^{-1}_{2Nt+N+i,2Nt+N+j} = \langle \hat{g}_i(t)\hat{g}_j(t)\rangle_{L_s'}$ and $\mathbf{S}^{-1}_{2Nt+i,2N(t+1)+N+j} = \langle g_i(t)\hat{g}_j(t+1)\rangle_{L_s'}$—where $\langle\ldots\rangle_{L_s'}$ indicates averages under the gaussian measure $\mathrm{e}^{-L_s'}$, with $L_s' = \frac{1}{2}\boldsymbol{\mathcal{G}}\mathbf{S}\boldsymbol{\mathcal{G}}$.

Variance $\langle g_i(t)^2\rangle_{L_s'}$

Here we close the set of equation (41) for the magnetizations with equations for the variances $\mathbf{S}^{-1}_{2Nt+i,2Nt+i}$ in terms of the interaction matrix $\mathbf{S}$, whose entries are linked to the magnetizations through (37)–(39*b*).

Recall that the inverse of the non-singular matrix $\mathbf{S}$ can be computed as [34]

$$\mathbf{S}^{-1}_{ij} = \frac{(-1)^{i+j}\det([\mathbf{S}]_{ji})}{\det(\mathbf{S})}, \tag{B.1}$$

where $[\mathbf{S}]_{ji}$ is the ji minor of the matrix $\mathbf{S}$, obtained removing the jth row and the ith column from the matrix itself. In case of the matrix defined by (37), whose determinant equals 1, the problem of inverting the matrix corresponds to computing the determinant of these minors. We now aim to calculate the determinant of $[\mathbf{S}]_{2Nt+i,2Nt+i}$ following the derivation of the determinant of $\mathbf{S}$.

As in appendix A, we start with factorizing out the determinant of the diagonal blocks $\mathbf{S}(t', t')$ up to $\mathbf{S}(t-1, t-1)$, according to (A.3). Given that these all equal 1, we can rewrite the determinant of $[\mathbf{S}]_{2Nt+i,2Nt+i}$ as:

$$\det([\mathbf{S}]_{2Nt+i,2Nt+i}) = \det(I(i,t)), \tag{B.2}$$

where

$$I(i,t) \equiv [\mathbf{S}^{\backslash t-1}]_{ii} - \begin{pmatrix}[\mathbf{S}(t,t-1)]_{\backslash i}\\ 0\\ \vdots\end{pmatrix}\tilde{\mathbf{S}}(t-1,t-1)^{-1}\left([\mathbf{S}(t-1,t)]^{\backslash i} \quad 0 \quad \ldots\right) \tag{B.3}$$

and we have defined $\tilde{\mathbf{S}}(t,t)$ in (A.8). $[\mathbf{S}(t,t-1)]_{\backslash i}$ and $[\mathbf{S}(t,t-1)]^{\backslash i}$ have been obtained removing respectively the ith row and the ith column from $\mathbf{S}(t,t-1)$. $[\mathbf{S}^{\backslash t-1}]_{ii}$ is instead the ii minor of $\mathbf{S}^{\backslash t-1}$ defined analogously as $\mathbf{S}^{\backslash 0}$ in appendix A:

$$\mathbf{S}^{\backslash t-1} = \begin{bmatrix} \mathbf{S}(t,t) & \mathbf{S}(t,t+1) & 0 & 0 & \cdots\\ \mathbf{S}(t+1,t) & \mathbf{S}(t+1,t+1) & \mathbf{S}(t+1,t+2) & 0 & \cdots\\ 0 & \mathbf{S}(t+2,t+1) & \mathbf{S}(t+2,t+2) & \mathbf{S}(t+2,t+3) & \cdots\\ 0 & 0 & \mathbf{S}(t+3,t+2) & \mathbf{S}(t+3,t+3) & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{B.4}$$

One can easily see that $I(i, t)$ in (B.2) preserves the block form of $[\mathbf{S}^{\setminus t-1}]_{ii}$, namely

$$I(i, t) = \left[\begin{array}{c|cccc} [\tilde{\mathbf{S}}(t, t)]_{ii} & [\mathbf{S}(t, t+1)]_{\setminus i} & 0 & 0 & \cdots \\ \hline [\mathbf{S}(t+1, t)]^{\setminus i} & & & & \\ 0 & & \mathbf{S}^{\setminus t} & & \\ \cdots & & & & \end{array}\right] \tag{B.5}$$

and therefore one can apply the formula in (A.3) once more to factorize the determinant in (B.2) into a product of two determinants as follows:

$$\det([\mathbf{S}]_{2Nt+i,2Nt+i}) = \det([\tilde{\mathbf{S}}(t, t)]_{ii}) \det(I(i, t+1)), \tag{B.6}$$

where $I(i, t+1)$ has been defined in (B.3).

With a bit of algebra it is possible to show that the matrix $I(i, t+1)$ in (B.6), has the very same structure as $\mathbf{S}^{\setminus t}$ and consequently of $\mathbf{S}$. Thus the second factor in the above equation is 1 (appendix A) and what's left is to compute the determinant of the ii minor of the matrix $\tilde{\mathbf{S}}(t, t)$.

Given the structure of $[\tilde{\mathbf{S}}(t, t)]_{ii}$

$$[\tilde{\mathbf{S}}(t, t)]_{ii} \equiv \left[\begin{array}{c|c} 0 & -\mathrm{i}\mathbb{I}_{\setminus i} \\ \hline -\mathrm{i}\mathbb{I}^{\setminus i} & \tilde{\gamma}(t) \end{array}\right], \tag{B.7}$$

where $\tilde{\gamma}(t)$ has been defined in (A.8), its determinant reduces to:

$$\begin{aligned} \det([\tilde{\mathbf{S}}(t, t)]_{ii}) &= (-1)^{N-1} \det(\tilde{\gamma}(t)) \det(\mathrm{i}\mathbb{I}_{\setminus i} \tilde{\gamma}(t)^{-1} \mathrm{i}\mathbb{I}^{\setminus i}) \\ &= (-1)^{2N-2} \det(\tilde{\gamma}(t)) \det([\tilde{\gamma}(t)^{-1}]_{ii}) \\ &= \tilde{\gamma}_{ii}(t). \end{aligned} \tag{B.8}$$

Finally we will check that the diagonal elements of $\mathbf{S}^{-1}$ we've just obtained are well defined variances by proving that they can take only positive values. In order to do that we will show that the matrix $\tilde{\gamma}(t)$ is positive definite.

By substituting γ and $\boldsymbol{\lambda}$, using respectively (39*a*) and (39*b*), in (43) one can express $\tilde{\gamma}(t)$ in terms of $\tilde{\gamma}(t-1)$, the matrix of the couplings $\mathbf{J}$ and the matrix $M_{ij}(t) \equiv \delta_{ij}\sqrt{(1-m_i(t)^2)}$ (m are the magnetizations) as

$$\tilde{\gamma}(t) = (\mathbf{J}\mathbf{M}(t))(\mathbf{J}\mathbf{M}(t))^{\top} + \mathbf{J}\mathbf{M}^2(t)\tilde{\gamma}(t-1)\mathbf{M}^2(t)\mathbf{J}^{\top}. \tag{B.9}$$

The first matrix on the right-hand side of (B.9) is positive definite. Since the sum of two positive definite matrices is positive definite, it is left to show that the second term on the right-hand side of (B.9) is positive definite. We will prove it by induction. First of all given that $\tilde{\gamma}(0) = \gamma(0)$, from the definition of γ in (39*a*), we know that $\tilde{\gamma}(0)$ is positive definite. Then we assume that $\tilde{\gamma}(t-1)$ is positive definite and we prove that $\mathbf{J}\mathbf{M}^2(t)\tilde{\gamma}(t-1)\mathbf{M}^2(t)\mathbf{J}^{\top}$ is positive definite. If $\tilde{\gamma}(t-1)$ is positive definite, it exist a matrix A such that $\tilde{\gamma}(t-1) = AA^{\top}$. Exploiting the latter one can rewrite:

$$\mathbf{J}\mathbf{M}^2(t)\tilde{\gamma}(t-1)\mathbf{M}^2(t)\mathbf{J}^{\top} = (\mathbf{J}\mathbf{M}^2(t)A)(\mathbf{J}\mathbf{M}^2(t)A)^{\top} \tag{B.10}$$

proving that the second term on the right-hand side of (B.9) is positive definite. Consequently $\tilde{\gamma}(t)$ is a positive definite matrix and its diagonal entries take only positive values.

Correlations $\langle \widehat{g}_i(t)\widehat{g}_j(t)\rangle_{L'_s}$

Here we will prove that the two point correlation function between conjugate fields $\mathbf{S}^{-1}_{2Nt+N+i,2Nt+N+j}$ is zero, as claimed in 4.4, where it enters the proof of consistency of the optimal parameter $\hat{\eta} = 0$.

Similarly to the previous subsection we will use (B.1) to invert the matrix S and compute the determinant of the minor through Shur's complement formula (A.2):

$$\begin{aligned}\mathbf{S}^{-1}_{2Nt+N+i,2Nt+N+j} &= \frac{(-1)^{4Nt+2N+i+j}\det\left([\mathbf{S}]_{2Nt+N+i,2Nt+N+j}\right)}{\det(\mathbf{S})} \\ &= (-1)^{i+j}\det\left(Y(i,j,t)\right)\end{aligned} \tag{B.11}$$

with

$$\begin{aligned}Y(i,j,t) = [\mathbf{S}^{\backslash t-1}]_{N+i,N+j} &- \begin{pmatrix}[\mathbf{S}(t,t-1)]_{\backslash N+i}\\ 0\\ \vdots\end{pmatrix} \\ &\times \tilde{\mathbf{S}}(t-1,t-1)^{-1}\begin{pmatrix}[\mathbf{S}(t-1,t)]^{\backslash N+j} & 0 & \ldots\end{pmatrix}\end{aligned} \tag{B.12}$$

and we have defined $\tilde{\mathbf{S}}(t,t)$ in (A.7) and $\mathbf{S}^{\backslash t-1}$ in (B.4). $Y(i,j,t)$ in (B.12) preserves the block form of $[\mathbf{S}^{\backslash t-1}]_{N+i,N+j}$, namely

$$Y(i,j,t) = \left[\begin{array}{c|ccccc}[\tilde{\mathbf{S}}(t,t)]_{N+i,N+j} & [\mathbf{S}(t,t+1)]_{\backslash N+i} & 0 & 0 & 0 & \cdots \\ \hline [\mathbf{S}(t+1,t)]^{\backslash N+j} & & & & & \\ 0 & & & \mathbf{S}^{\backslash t} & & \\ \cdots & & & & & \end{array}\right]. \tag{B.13}$$

Conversely to the previous section we cannot express the determinant of $Y(i,j,t)$ in terms of the Shur's complement of $[\tilde{\mathbf{S}}(t,t)]_{N+i,N+j}$, since the latter is a singular matrix. One has instead to resort to the Shur's complement of the matrix $\mathbf{S}^{\backslash t}$ that we know is invertible and its determinant is 1, having the same structure of the matrix $\mathbf{S}$ (appendix A):

$$\begin{aligned}\det Y(i,j,t) = \det\mathbf{S}^{\backslash t}\det\left([\tilde{\mathbf{S}}(t,t)]_{N+i,N+j}\right) \\ - \begin{pmatrix}[\mathbf{S}(t,t+1)]_{\backslash N+i} & 0 & \ldots\end{pmatrix}[\mathbf{S}^{\backslash t}]^{-1}\begin{pmatrix}[\mathbf{S}(t+1,t)]^{\backslash N+j}\\ 0\\ \vdots\end{pmatrix}\Bigg).\end{aligned} \tag{B.14}$$

Just like $[\tilde{\mathbf{S}}(t,t)]_{N+i,N+j}$, the matrix whose determinant is the second factor on the right-hand side of (B.14) is singular: as can be easily checked its ith column is null, regardless of the elements of $[\mathbf{S}^{\backslash t}]^{-1}$. This completes the proof that $S_{2Nt+N+i,2Nt+N+j} = 0$ for all $i, j = 1,\ldots,N$ and $t = 0,\ldots,T-1$.

Correlations $\langle g_i(t)\widehat{g}_j(t+1)\rangle_{L'_s}$

Here we will prove that the two point correlation function between conjugate fields $\mathbf{S}^{-1}_{2Nt+i,2N(t+1)+N+j} = 0$ is zero, as claimed in 4.4, where it enters the proof of consistency of the optimal parameter $\hat{\eta} = 0$. The derivation is very similar to the one for $\mathbf{S}^{-1}_{2Nt+N+i,2Nt+N+j} = 0$ in the previous subsection.

We will use (B.1) to invert the matrix S and compute the determinant of the minor through Shur's complement formula (A.2):

$$\mathbf{S}^{-1}_{2Nt+i,2N(t+1)+N+j} = \frac{(-1)^{4Nt+3N+i+j}\det\left([\mathbf{S}]_{2Nt+i,2N(t+1)+N+j}\right)}{\det(\mathbf{S})} = (-1)^{3N+i+j}\det\left(Z(i,j,t)\right) \tag{B.15}$$

with

$$Z(i,j,t) \equiv [\mathbf{S}^{\setminus t-1}]_{i,3N+j} - \begin{pmatrix} [\mathbf{S}(t,t-1)]_{\setminus i} \\ 0 \\ \vdots \end{pmatrix} \times \tilde{\mathbf{S}}(t-1,t-1)^{-1}\left([\mathbf{S}(t-1,t)] \;\; 0^{\setminus N+j} \;\; \ldots\right) \tag{B.16}$$

and we have defined $\tilde{\mathbf{S}}(t,t)$ in (A.8) and $\mathbf{S}^{\setminus t-1}$ in (B.4). $Z(i,j,t)$ in (B.16) preserves the block form of $[\mathbf{S}^{\setminus t-1}]_{i,3N+j}$, namely

$$Z(i,j,t) = \left[\begin{array}{cc|cc} [\tilde{\mathbf{S}}(t,t)]_{\setminus i} & [\mathbf{S}(t,t+1)]_{i,N+j} & 0 & 0 \;\cdots \\ \mathbf{S}(t+1,t) & \mathbf{S}(t+1,t+1)^{\setminus N+j} & \mathbf{S}(t+1,t+2) & 0 \;\cdots \\ \hline 0 & \mathbf{S}(t+2,t+1) & & \\ \ldots & 0 & \mathbf{S}^{\setminus t+1} & \\ & \cdots & & \end{array}\right]. \tag{B.17}$$

Analogously to the previous section we will now express the determinant of $Z(i,j,t)$ using the Shur's complement of the matrix $\mathbf{S}^{\setminus t+1}$ that we know is invertible and its determinant is 1, just like the matrix $\mathbf{S}$, as shown in appendix A:

$$\det Z(i,j,t) = \det \mathbf{S}^{\setminus t+1} \det\left(\begin{bmatrix} [\tilde{\mathbf{S}}(t,t)]_{\setminus i} & [\mathbf{S}(t,t+1)]_{i,N+j} \\ \mathbf{S}(t+1,t) & \mathbf{S}(t+1,t+1)^{\setminus N+j} \end{bmatrix} - \begin{pmatrix} 0 & 0 \;\cdots \\ \mathbf{S}(t+1,t+2) & 0 \;\cdots \end{pmatrix}[\mathbf{S}^{\setminus t+1}]^{-1}\begin{pmatrix} 0 & \mathbf{S}(t+2,t+1) \\ 0 & 0 \\ \cdots & \cdots \end{pmatrix}\right). \tag{B.18}$$

The structure of $[\mathbf{S}^{\setminus t+1}]^{-1}$ reflects $\mathbf{S}^{-1}$ structure:

$$[\mathbf{S}^{\setminus t+1}]^{-1} = \begin{bmatrix} \Omega(t+2,t+2) & \Omega(t+2,t+3) & \Omega(t+2,t+3) & \cdots \\ \Omega(t+3,t+2) & \Omega(t+1,t+1) & \cdots & \cdots \\ \Omega(t+4,t+2) & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{B.19}$$

with

$$\Omega(t+2,t+2) = \begin{bmatrix} \tilde{\gamma}(t+2) & \Delta \\ \Gamma & 0 \end{bmatrix}, \tag{B.20}$$

where Δ and Γ are matrices of order $2N$. The block form of $\Omega(t+2,t+2)$ follows that of the diagonal blocks of $\mathbf{S}^{-1}$, that was proven to be such in the previous sections of this appendix.

Using (B.19) one can check that the matrix whose determinant is the second factor on the right-hand side of (B.18) is singular: its $(N+j)$-th row is null. This completes the proof that $\mathbf{S}^{-1}_{2Nt+i,2N(t+1)+N+j} = 0$ for all $i,j = 1,\ldots,N$ and $t = 0,\ldots,T-1$.

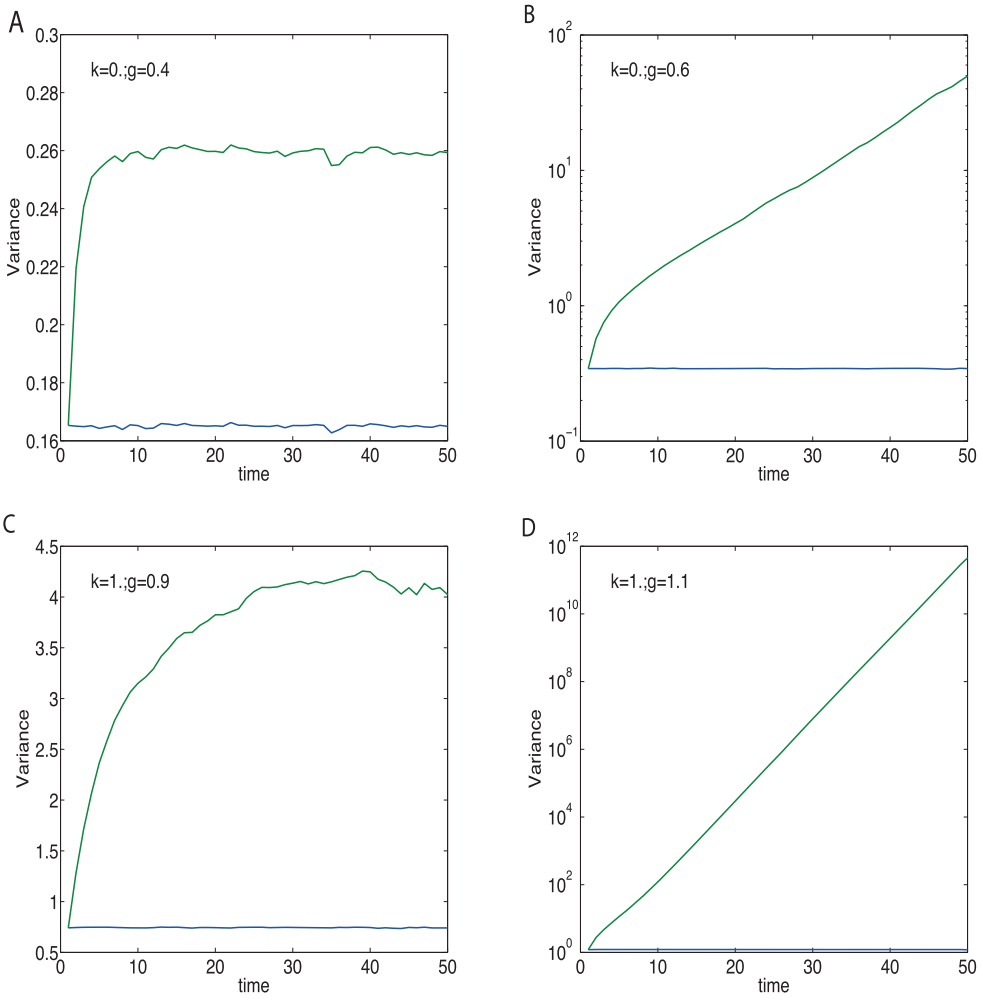

Figure C1. The mean variance in the gaussian integral for the magnetizations versus time of reconstruction, when the experimental history of magnetization is known. Green: the gaussian average method; blue: the variance in [1]. $N = 20$, single realization of the couplings with $g = 0.4$ (A), $g = 0.6$ (B), $g = 0.9$ (C) and $g = 1.1$ (D). The experimental magnetizations are computed using 10^4 samples of the dynamics. Zero external field. Top: asymmetry parameter $k = 0$. Bottom: asymmetry parameter $k = 1$.

Appendix C. Gaussian average method: variance

In this appendix we study the stability of the dynamical system for the matrix $\tilde{\gamma}$ defined in the main text by (43). In order to do that we first average (43) over the distribution of the couplings introduced in section 6 through (60) and (61). Consider then entries of $\tilde{\gamma}(1)$:

$$\begin{aligned} \overline{\tilde{\gamma}_{ij}(1)} &= \quad g^2(1 + g^2\overline{\tilde{\gamma}_{lm}(0)}) \quad \text{for} \qquad k = 0 \\ \overline{\tilde{\gamma}_{ij}(1)} &= g^2/2(1 + 8g^2\overline{\tilde{\gamma}_{lm}(0)}) \text{ for} \qquad k = 1, \end{aligned} \tag{C.1}$$

where the overbar indicates the average over the disorder, k is the parameter controlling the asymmetry of the couplings, while g is the coupling strength. Notice the different factors in (C.1) due to the correlations between the couplings. If we consider the univariate analogous to (C.1):

$$\begin{aligned} x(t) &= \quad g^2(1 + g^2 x(t-1)) \quad \text{for} \qquad k = 0 \\ x(t) &= g^2/2(1 + 8g^2 x(t-1)) \text{ for} \qquad k = 1 \end{aligned} \tag{C.2}$$

one can easily check that this dynamical system is characterized by a critical value for g, g_0 that discriminates between different stability classes for the system. Below g_0 x in (C.2) converges to a finite value, while it grows exponentially in time for $g > g_0$. The critical value for this chaotic behavior is respectively $g_0 = 1$ for fully asymmetric couplings ($k = 1$) and $0.7 < g_0 < 0.8$ for fully symmetric ones ($k = 0$).

We got numerical evidence to support our intuition. We found that the variance in the gaussian average method undergoes a chaotic behavior when the couplings strength reaches a certain critical value, $g_0 \sim 0.5$ for symmetric connectivities and $g_0 \sim 1$ for fully asymmetric ones. This value depends on the degree of symmetry of the couplings, on the presence of the external field and there are small fluctuation across different realizations of the couplings, but the phenomenon is qualitatively conserved. Figure C1 shows single realizations of the couplings below and above critical values and compares the mean of the gaussian average variances $\frac{1}{N}\sum_i \tilde{\gamma}_{ii}(t)$ with the mean of the MS-MF variances (mean of $\gamma_{ii}(t)$ (39*a*) in our notation) in the gaussian integral for fully asymmetric couplings [1].

Appendix D. Details on the extended Plefka expansion

We rewrite the functional Γ_α (51) as

$$\Gamma_\alpha = \ln \int \mathrm{d}\mathcal{G}\ \mathrm{Tr}_s \exp(\Xi_\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}, \mathcal{G}, \mathbf{s}]) - NT \ln 2\pi - N \ln 2, \tag{D.1}$$

where

$$\begin{aligned} \Xi_\alpha[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}, \mathcal{G}, \mathbf{s}] = \sum_{i,t} \Bigg\{ & \mathrm{i}\hat{g}_i(t)[g_i(t) - \alpha \sum_j J_{ij} s_j(t)] + s_i(t+1) g_i(t) \\ & - \ln 2 \cosh g_i(t) - \mathrm{i} h_i(t)[\hat{g}_i(t) - \hat{m}_i(t)] + \psi_i(t)[s_i(t) - m_i(t)] \\ & + \frac{1}{2}\sum_{t'} \hat{C}_i(t, t')[s_i(t) s_i(t') - C_i(t, t')] + \frac{1}{2}\sum_{t'} \hat{B}_i(t, t')[\hat{g}_i(t)\hat{g}_i(t') - B_i(t, t')] \\ & - \mathrm{i}\sum_{t'} \hat{R}_i(t', t)[\hat{g}_i(t) s_i(t') - \mathrm{i} R_i(t', t)] \Bigg\}, \end{aligned} \tag{D.2}$$

and proceed with the perturbation expansion of Γ_α around $\alpha = 0$ up to the second order:

$$\Gamma_\alpha = \Gamma^{(0)} + \alpha\Gamma^{(1)} + \frac{\alpha^2}{2}\Gamma^{(2)}, \tag{D.3}$$

where $\Gamma^{(k)} = \partial^k \Gamma_\alpha / \partial \alpha^k|_{\alpha=0}$. At the end of the calculation we will set $\alpha = 1$. The first term in the expansion is given by

$$\Gamma^{(0)}[\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}] = \ln Z_0[\boldsymbol{\psi}^0, \boldsymbol{h}^0, \hat{\mathbf{C}}^0, \hat{\mathbf{B}}^0, \hat{\mathbf{R}}^0] - \sum_{it}\psi_i^0(t)m_i(t) + \mathrm{i}\sum_{it}h_i^0(t)\hat{m}_i(t)$$
$$- \frac{1}{2}\sum_{itt'}\hat{C}_i^0(t, t')C_i(t, t') - \frac{1}{2}\sum_{itt'}\hat{B}_i^0(t, t')B_i(t, t') - \sum_{itt'}\hat{R}_i^0(t', t)R_i(t', t), \tag{D.4}$$

where

$$Z_0[\boldsymbol{\psi}, \boldsymbol{h}, \hat{\mathbf{C}}, \hat{\mathbf{B}}, \hat{\mathbf{R}}] = \frac{1}{2^N(2\pi)^{NT}}\int \mathrm{d}\mathcal{G}\,\mathrm{Tr}_s\prod_{it}\ \exp\{\mathrm{i}\hat{g}_i(t)g_i(t) + s_i(t+1)g_i(t)$$
$$- \ln\cosh g_i(t) - \mathrm{i}\hat{g}_i(t)h_i(t) + \psi_i(t)s_i(t) + \frac{1}{2}\sum_{t'}\hat{C}_i(t, t')s_i(t)s_i(t')$$
$$+ \frac{1}{2}\sum_{t'}\hat{B}_i(t, t')\hat{g}_i(t)\hat{g}_i(t') - \mathrm{i}\sum_{t'}\hat{R}_i(t', t)\hat{g}_i(t)s_i(t')\Big\} \tag{D.5}$$

and $\boldsymbol{\psi}^0, \boldsymbol{h}^0, \hat{\mathbf{C}}^0, \hat{\mathbf{B}}^0, \hat{\mathbf{R}}^0$ are the fields for which the set of equations (52) is satisfied for $\Gamma_\alpha = \Gamma^{(0)}$ for a given value of $\mathbf{m}, \hat{\mathbf{m}}, \mathbf{C}, \mathbf{B}, \mathbf{R}$. We compute $\Gamma^{(1)}$ as follows:

$$\Gamma^{(1)} = \frac{\partial\Gamma_\alpha}{\partial\alpha}\bigg|_{\alpha=0} = \left\langle\frac{\partial\Xi_\alpha}{\partial\alpha}\right\rangle\bigg|_{\alpha=0}. \tag{D.6}$$

Using (D.2) one finds:

$$\frac{\partial\Xi_\alpha}{\partial\alpha} = -\mathrm{i}\sum_{ijt}J_{ij}\hat{g}_i(t)s_j(t) - \mathrm{i}\sum_{it}\frac{\partial h_i(t)}{\partial\alpha}[\hat{g}_i(t) - \hat{m}_i(t)] + \sum_{it}\frac{\partial\psi_i(t)}{\partial\alpha}[s_i(t) - m_i(t)]$$
$$+ \sum_{itt'}\frac{1}{2}\sum_{t'}\frac{\partial\hat{C}_i(t, t')}{\partial\alpha}[s_i(t)s_i(t') - C_i(t, t')]$$
$$+ \sum_{itt'}\frac{1}{2}\sum_{t'}\frac{\partial\hat{B}_i(t, t')}{\partial\alpha}[\hat{g}_i(t)\hat{g}_i(t') - B_i(t, t')]$$
$$- \mathrm{i}\sum_{itt'}\sum_{t'}\frac{\partial\hat{R}_i(t', t)}{\partial\alpha}[\hat{g}_i(t)s_i(t') - iR_i(t', t)]. \tag{D.7}$$

When computing the average $\langle\partial\Xi_\alpha/\partial\alpha\rangle_\alpha$ as defined in (50), all the terms on the right-hand side of (D.7) except for the first one vanish because of the set of equations (49). Moreover at $\alpha = 0$ the spins are decoupled and the averages are trivial:

$$\Gamma^{(1)} = -\mathrm{i}\sum_{ijt}J_{ij}\langle\hat{g}_i(t)s_j(t)\rangle_0 = -\mathrm{i}\sum_{ijt}J_{ij}\hat{m}_i(t)m_j(t). \tag{D.8}$$

For the second derivative of Γ_α with respect to α we have

$$\frac{\partial^2\Gamma_\alpha}{\partial\alpha^2} = \left\langle\frac{\partial^2\Xi_\alpha}{\partial\alpha^2}\right\rangle_\alpha + \left\langle\left(\frac{\partial\Xi_\alpha}{\partial\alpha}\right)^2\right\rangle_\alpha - \left(\left\langle\frac{\partial\Xi_\alpha}{\partial\alpha}\right\rangle_\alpha\right)^2. \tag{D.9}$$

Using (D.7) and the set of equations (49), it is easy to show that the first term on the right-hand side of the above equation is zero. One thus finds

$$\Gamma^{(2)} = \left\langle\left(\frac{\partial\Xi_\alpha}{\partial\alpha} - \left\langle\frac{\partial\Xi_\alpha}{\partial\alpha}\right\rangle\right)^2\right\rangle\bigg|_{\alpha=0}, \tag{D.10}$$

which can be computed using (D.7) and the following Maxwell equations:

$$\frac{\partial \psi_i(t)}{\partial \alpha}\bigg|_{\alpha=0} = -\frac{\partial}{\partial m_i(t)}\frac{\partial \Gamma_\alpha}{\partial \alpha}\bigg|_{\alpha=0} = \mathrm{i}\sum_j J_{ji}\hat{m}_j(t)$$
$$\mathrm{i}\frac{\partial h_i(t)}{\partial \alpha}\bigg|_{\alpha=0} = \frac{\partial}{\partial \hat{m}_i(t)}\frac{\partial \Gamma_\alpha}{\partial \alpha}\bigg|_{\alpha=0} = -\mathrm{i}\sum_j J_{ij} m_j(t). \tag{D.11}$$

Note that the derivatives of the two-time conjugate fields with respect to α are zero, e.g.

$$\frac{\partial \hat{C}(t,t')}{\partial \alpha} = \frac{\partial}{\partial C_i(t,t')}\frac{\partial \Gamma_\alpha}{\partial \alpha} = 0. \tag{D.12}$$

We finally obtain

$$\Gamma^{(2)} = -\sum_{iji'j'tt'} \langle \delta\hat{g}_i(t) J_{ij} \delta s_j(t) \delta\hat{g}_{i'}(t') J_{i'j'} \delta s_{j'}(t')\rangle_0, \tag{D.13}$$

where we defined $\delta s_i(t) = s_i(t) - m_i(t)$ and $\delta\hat{g}_i(t) = \hat{g}_i(t) - \hat{m}_i(t)$. Since the averages are taken at $\alpha = 0$ spins at different sites are decoupled and the only non-vanishing terms in (D.13) correspond to the case $i' = i$, $j' = j$ and $i' = j$, $j' = i$:

$$\Gamma^{(2)} = -\sum_{ijtt'} [J_{ij}^2 \langle \delta\hat{g}_i(t)\delta\hat{g}_i(t')\rangle_0 \langle \delta s_j(t)\delta s_j(t')\rangle_0 + J_{ij}J_{ji}\langle \delta\hat{g}_i(t)\delta s_i(t')\rangle_0 \langle \hat{g}_j(t')\delta s_j(t)\rangle_0], \tag{D.14}$$

which can be written in terms of the moments as follows

$$\Gamma^{(2)} = -\sum_{ijtt'} [J_{ij}^2 (B_i(t,t') - \hat{m}_i(t)\hat{m}_i(t'))(C_j(t,t') - m_j(t)m_j(t')) + J_{ij}J_{ji}(\mathrm{i}R_i(t',t) - \hat{m}_i(t)m_i(t'))(\mathrm{i}R_j(t,t') - \hat{m}_j(t')m_j(t))]. \tag{D.15}$$

Inserting (D.4), (D.8) and (D.15) in (D.3) we find the explicit expression of the functional Γ_α expanded up to the second order. Considering the set of equations (52) within the second order expansion and setting the auxiliary fields to zero, we can extract the value of the fields $\boldsymbol{\psi}^0$, $\boldsymbol{h}^0$, $\hat{\mathbf{C}}^0$, $\hat{\mathbf{B}}^0$, $\hat{\mathbf{R}}^0$ as functions of the correct (within the expansion) marginal first and second moments:

$$\psi_i^0(t) = 0$$
$$h_i^0(t) = h_i(t) + \sum_j J_{ij} m_j(t) - \sum_{jt'} J_{ij}J_{ji}R_j(t,t')m_i(t')$$
$$\hat{C}_i^{\,0}(t,t') = 0$$
$$\hat{B}_i^{\,0}(t,t') = -\sum_j J_{ji}^2 (C_j(t,t') - m_j(t)m_j(t'))$$
$$\hat{R}_i^{\,0}(t,t') = \sum_j J_{ij}J_{ji}R_j(t',t). \tag{D.16}$$

From general results for generating functional analysis of spin systems [14] it can be shown that $\hat{m} = 0$, $B = 0$ and $R(t,t')$ has the meaning of a local response function and is non-vanishing only for $t > t'$. To get an explicit expression for Z_0 we insert (D.16) in (D.5). It yields $Z^0[h] = \prod Z_i^0[h_i]$, where

$$Z_i^0[h_i] = \frac{1}{2^N(2\pi)^{NT}} \int \mathrm{d}\mathcal{G}\, \mathrm{Tr}_{s_i} \prod_t \exp\{s_i(t+1)g_i(t) - \ln\cosh g_i(t) - \mathrm{i}\hat{g}_i(t)h_i(t) + \mathrm{i}\hat{g}_i(t)g_i(t) - \mathrm{i}\sum_j \hat{g}_i(t)J_{ij}m_j(t) - \frac{1}{2}\sum_{jt'}\hat{g}_i(t)J_{ij}^2[C_j(t,t') - m_j(t)m_j(t')]\hat{g}_i(t') - \mathrm{i}\sum_{jt'}\hat{g}_i(t)J_{ij}J_{ji}R_j(t',t)[s_i(t') - m_i(t')]\Big\}. \tag{D.17}$$

To linearize the quadratic terms in (D.17), we introduce the gaussian random variables $\phi_i(t)$, independently for each i, with zero mean and covariance $\langle\phi_i(t)\phi_i(t')\rangle = \sum_j J_{ij}^2(C_j(t,t') - m_j(t)m_j(t'))$, obtaining:

$$Z_i^0[h_i] = \frac{1}{2^N(2\pi)^{NT}} \left\langle \int \mathrm{d}\mathcal{G}\, \mathrm{Tr}_{s_i} \prod_t \exp\{s_i(t+1)g_i(t) - \ln\cosh g_i(t) + \mathrm{i}\hat{g}_i(t)[g_i(t) - \left(\phi_i(t) + \sum_j J_{ij}m_j(t) - \sum_j\sum_{t'}^{t-1} J_{ij}J_{ji}R_j(t,t')[s_i(t') - m_i(t')] + h_i(t)\right)]\Big\} \right\rangle_{\phi_i}. \tag{D.18}$$

From the above equation one can see that the moment $R_i(t,t')$ defined in (49) can be written as an average over the fields $\phi_i(t)$ as follows

$$R_i(t,t') = \left\langle \frac{\partial s_i(t)}{\partial \phi_i(t')} \right\rangle_{\phi_i}, \tag{D.19}$$

and can be interpreted as a response function.

Appendix E. The Yule–Walker equations

We want to generate the gaussian random field $\phi_i^k(t)$ for given trajectory k and spin i based on the past values of the field $\phi_i^k(0)$, $\phi_i^k(1) \ldots \phi_i^k(t-1)$. Since the random variables $\phi_i^k(0)$, $\phi_i^k(1) \ldots \phi_i^k(t)$ are jointly gaussian distributed with zero mean, we know that the conditional expectation $\widehat{\phi}_i^k(t) \equiv E\{\phi_i^k(t)|\phi_i^k(0), \phi_i^k(1) \ldots \phi_i^k(t-1)\}$ is given by the linear estimate

$$\widehat{\phi}_i^k(t) = \sum_{r=0}^{t-1} a(r)\phi_i^k(r), \tag{E.1}$$

which also happens to be the best mean square estimate of $\phi_i^k(t)$ given $\phi_i^k(r)$, $r = 0 \ldots t-1$. The coefficients $a(0)$, $a(1)$,$\ldots a(t-1)$ are such that the mean square value of the estimation error $E\{[\phi_i^k(t) - \widehat{\phi}_i^k(t)]^2\}$ is minimum. By the orthogonality principle, this condition holds if the following set of equations is satisfied

$$E\{[\phi_i^k(t) - \sum_{r=0}^{t-1} a(r)\phi_i^k(r)]\phi_i^k(r')\} = 0, \quad r' = 0 \ldots t-1, \tag{E.2}$$

which can be rewritten in matrix form as

$$\mathbf{A}\mathcal{C} = \mathcal{C}_t, \tag{E.3}$$

where $\mathbf{A} = [a(0) \ldots a(t-1)]$ is the vector of coefficients, $\mathcal{C}$ is the correlation matrix with elements $\mathcal{C}(r,r') = E\{\phi_i^k(r)\phi_i^k(r')\}$ for $r, r' = 0 \ldots t-1$ and $\mathcal{C}_t$ is the vector with elements

$C_t(r) = E\{\phi_i^k(t)\phi_i^k(r)\}$ for $r = 0 \ldots t-1$. Since $\phi_i^k(t) - \widehat{\phi}_i^k(t) \perp \phi_i^k(r)$ for every $r = 0, \ldots t-1$, from (E.1) we conclude that $\phi_i^k(t) - \widehat{\phi}_i^k(t) \perp \widehat{\phi}_i^k(t)$ and the error reduces to

$$E\{[\phi_i^k(t) - \widehat{\phi}_i^k(t)]^2\} = E\{[\phi_i^k(t) - \widehat{\phi}_i^k(t)]\phi_i^k(t)\} = 1 - \sum_j J_{ij}^2 m_j^2(t) - \mathbf{A}C_t. \tag{E.4}$$

Knowing that $E\{\phi_i^k(r)\phi_i^k(r')\} = \sum_j J_{ij}^2[C_j(r, r') - m_j(r)m_j(r')]$ $r, r' = 0 \ldots t$, we can compute the coefficients $a(r)$ from (E.3) and draw the gaussian random variable $\phi_i^k(t)$ with mean and covariance given, respectively, by (E.1) and (E.4).

References

[1] Mézard M and Sakellariou J 2011 *J. Stat. Mech.* L07001
[2] Roudi Y and Hertz J 2011 *J. Stat. Mech.* P03031
[3] Crisanti A and Sompolinsky H 1987 *Phys. Rev.* A **36** 4922
[4] Kappen H and Spanjers J 2000 *Phys. Rev.* E **61** 5658
[5] Plefka T 1982 *J. Phys. A: Math. Gen.* **15** 1971
[6] Aurell E and Mahmoudi H 2012 *Phys. Rev.* E **85** 031119
[7] Toyoizumi T, Rad K R and Paninski L 2009 *Neural Comput.* **21** 1203–43
[8] Huang H and Kabashima Y 2014 *J. Stat. Mech.: Theory and Experiment* **2014** P05020
[9] Mahmoudi H and Saad D 2014 *J. Stat. Mech.* P07001
[10] Biroli G 1999 *J. Phys. A: Math. Gen.* **32** 8365
[11] Bravi B, Sollich P and Opper M 2015 arXiv:1509.07066
[12] Martin P C, Siggia E and Rose H 1973 *Phys. Rev.* A **8** 423
[13] De Dominicis C and Peliti L 1978 *Phys. Rev.* B **18** 353
[14] Coolen A C C 2000 arXiv:cond-mat/0006011
[15] Hertz J A, Roudi Y and Sollich P 2016 arXiv:1604.05775
[16] Opper M and Saad D 2001 *Advanced Mean Field Methods: Theory and Practice* (Cambridge, MA: MIT Press)
[17] Kholodenko A 1990 *J. Stat. Phys.* **58** 355–70
[18] Weiss P 1907 *J. Phys. Theor. Appl.* **6** 661–90
[19] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Berlin: Springer)
[20] Müschlegel B and Zittarts H 1963 *Z. Phys.* **175** 553–73
[21] Sissakian A and Solovtsov I 1992 *Z. Phys.* C **54** 263–71
[22] Kleinert H 2009 *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets* 5th edn (Singapore: World Scientific)
[23] Stevenson P M 1981 *Phys. Rev.* D **23** 2916–44
[24] Altland A and Simons B D 2010 *Condensed Matter Field Theory* (Cambridge: Cambridge University Press)
[25] Georges A and Yedidia J S 1991 *J. Phys. A: Math. Gen.* **24** 2173
[26] Sakellariou J, Roudi Y, Mezard M and Hertz J 2012 *Phil. Mag.* **92** 272–9
[27] Peretto P 1984 *Biol. Cybern.* **50** 51–62
[28] Thouless D J, Anderson P W and Palmer R G 1977 *Phil. Mag.* **35** 593–601
[29] Cavagna A, Giardina I, Parisi G and Mézard M 2003 *J. Phys. A: Math. Gen.* **36** 1175
[30] Scharnagl A, Opper M and Kinzel W 1995 *J. Phys. A: Math. Gen.* **28** 5721
[31] Roudi Y and Hertz J 2011 *Phys. Rev. Lett.* **106** 048702
[32] Dunn B and Roudi Y 2013 *Phys. Rev.* E **87** 022127
[33] Bravi B and Sollich P 2016 arXiv:1603.05538
[34] Friedberg S, Insel A and Spence L 1996 *Linear Algebra* 3rd edn (Englewood Cliffs, NJ: Prentice-Hall)

3.3 Discussion

In this section, we present additional comparisons between different algorithms aimed at predicting single site magnetizations. The analysis of Paper 1 reveals that a novel approximation denoted as the Extended Plefka Expansion outperforms other existing methods, at the cost of keeping memory of the spin fluctuations at all past times. The required Monte Carlo algorithm can be speeded up from $T^2N + TNN_T$ to $T^2 + TNN_T$ computational time steps, where N is the size of the system, T the number of time steps and N_T the number of Monte Carlo trajectories. If the couplings J_{ij} have variance $\sim 1/N$, we expect local two-times moments to be self-averaging. Hence, we can rewrite the effective single site field of equation (55) as

$$g_i(t) = \phi_i(t) + \sum_j \Big(J_{ij} m_j(t) \\ -g^2 \frac{1-k^2}{1+k^2} \frac{1}{N} \sum_{t'=0}^{t-1} \sum_j R_j(t,t')[s_i(t') - m_i(t')] \Big) + h_i(t). \tag{3.1}$$

and the covariance of the Gaussian noise of equation (54) becomes

$$\langle \phi_i(t)\phi_i(t') \rangle = \frac{g^2}{N} \sum_j [C_j(t,t') - m_j(t)m_j(t')]. \tag{3.2}$$

The new version of the algorithm is written in Appendix 3.C. Figure 3.1 shows that - for networks of 50 spins- the accuracy of the prediction on the magnetization is reduced if we replace the local second order moments with their average value, especially for small couplings. However, the scaling of the mean squared error with the system size (Figure 3.2) suggests that the two algorithms perform equally well for very large networks: in both implementations of the extended Plefka method, error of the predicted magnetization goes to zero in the limit $N \to \infty$.

As a final comparison, we investigate the performance of the Extended Gaussian approximation of [MS14] (see section 3.B) and plot its error in Figure 3.1, in case of a symmetric network. We recall that, for completely asymmetric networks, this method agrees with the exact mean field theory of [MS11]. One observes that, for very small couplings (high temeratures), all the methods approach each other. For quite large values of the couplings (corresponding to values of g greater than $g \approx 1.5$), the extended Gaussian approximation performs equally well as the extended Plefka algorithm. To be more precise,

the extended Gaussian method outperforms the extended Plefka approximation for $N = 50$; however, we observe that the result of the extended Plefka method decreases for increasing N, contrary to the extended Gaussian method (Figure 3.2). There are intermediate values of the strength of the couplings (approximately corresponding to values $0.2 < g < 1.5$) for which the performance of the extended Gaussian approximation is poorer that both the Extended Plefka and the RH-TAP, where the latter method does not consider the effect of correlations and responses.

3.4 Conclusions

We have presented several approximate methods to study the transient dynamics of an Ising model with the parallel update Glauber rule, deriving equations for the time evolution of single-spin magnetisations for fixed values of the quenched couplings. We considered system with couplings that are weak, long-ranged and are allowed to have any degree of symmetry. While a lot of efforts have been devoted to the study of the dynamics of spin glasses with symmetric couplings (for a review see [BCKM98]), comparatively less attention has been devoted to networks with asymmetric couplings - for which an energy function cannot be defined. In the fully asymmetric case, correlations between spins at various times are negligible, and the fields acting on each spin can be simply described in terms of effective Gaussian fields [MS11] . However, if the couplings have a non-zero degree of symmetry, these correlations persist also at distant times, and studying the dynamics is less trivial. In this chapter, we introduced a novel approaches to this task, denoted as the extended Plefka expansion. It is based on a weak coupling expansion of the log-generating functional, where the local moments over time are constrained via a Legendre transform. The novelty of our formalism relies in including not only the first-order marginal moments, but also all the second-order marginal moments; the latter quantities are trivial for the equilibrium Ising model (since $\sigma^2 = 1$) but are non-negligible in the dynamic case for a correct mean-field description of the system. The result is an effective log-generating functional that factorizes over single-site trajectories and contains the correct first and second order moments within the approximation. Namely, within the approximate description, each spin is subjected to an effective local field which contains Gaussian noise - encoding the effect of correlations with past spin values - and a memory term where the response function is coupled to the whole history of spin fluctuations. The contribution of the memory term is stronger for larger degree of symmetry of the network, and negligible when the couplings are uncorrelated, in which case the exact mean field theory is retrieved. The

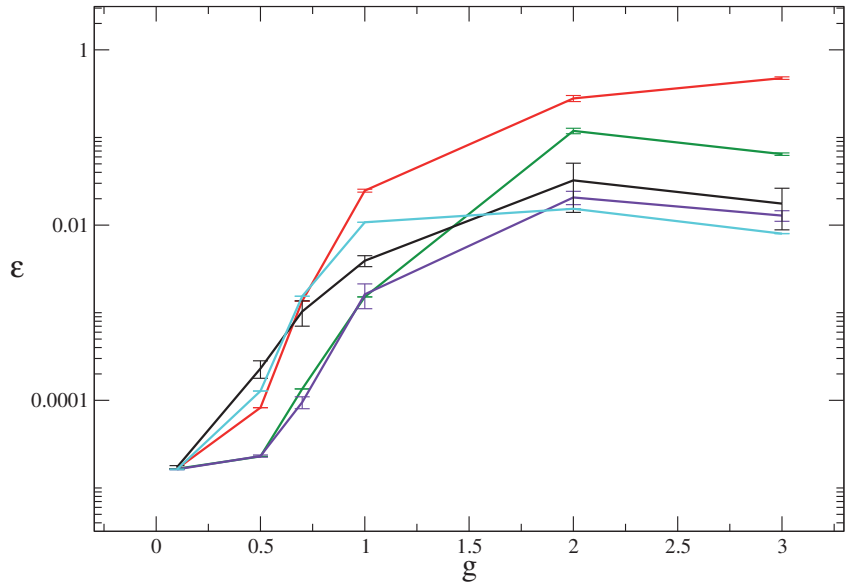

Figure 3.1: Mean squared error of the predicted magnetization averaged over spins and time: $\varepsilon = \langle m_i(t)^{\text{experimental}} - m_i(t)^{\text{predicted}} \rangle$, where the 'experimental' magnetization is the one we get from simulating the exact dynamics by using 50000 Monte Carlo repeats. The mean squared error is plotted as a function of the parameter g - representing the strength of the couplings- for a system of $N = 50$ spins with fully symmetric couplings. We have used 100 time steps and have averaged the errors over 10 realizations of the couplings. The error bars are standard deviations over these realizations. Different colors correspond to different methods to predict the magnetizations, some of them referring to Paper 1: RH-TAP (green), MS-MF (red), extended Gaussian average (light blue), extended Plefka (violet) and extended Plefka with averaged moments (black). The number of sample trajectories used in the algorithm for both the extended Plefka and extended Plefka with averaged moments methods is $N_T = 50000$.

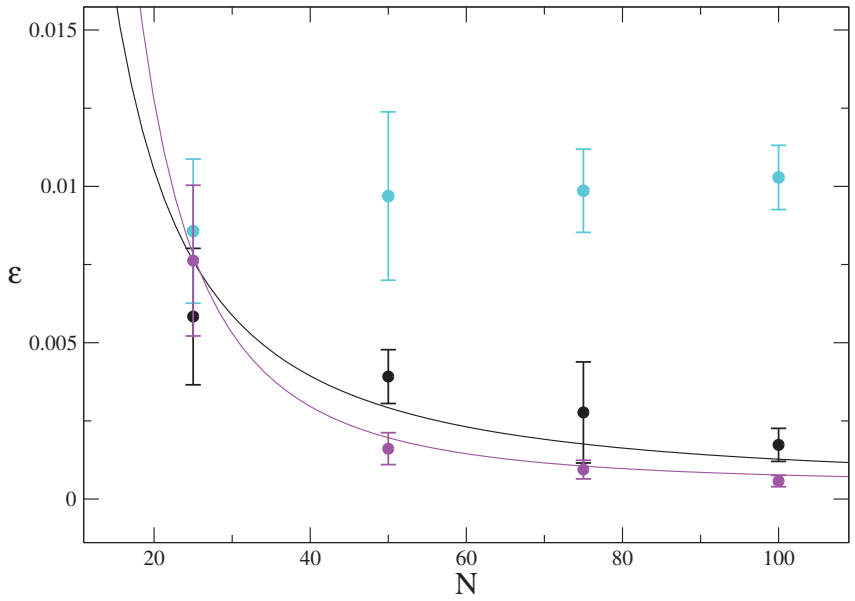

Figure 3.2: Mean squared error of the Extended Plefka method with averaged moments (black), Extended Gaussian approximation (blue) and Extended Plefka method of Paper 1 (magenta) for predicting entire dynamics of magnetizations. The mean squared error is computed for a fully symmetric network with fixed coupling strength $g = 1$, and plotted as a function of the system size N. We have used 100 time steps and 50,000 repeats to calculate the experimental magnetizations and have averaged the errors over 10 realizations of the couplings. The error bars are standard deviations over these realizations. The number of sample trajectories used in the algorithm for the algorithms based on the Extended Plefka approximation is $N_T = 50000$. Stationary external field drawn independently for each spin from a normal distribution (zero mean, standard deviation 0.5). The Extended Gaussian method shows no scaling with N, a behaviour that we also observed at $g = 0.5$ and $g = 3$. For the error of the two algorithms based on the Extended Plefka approximation, we fitted a shifted power law to the data, to infer the value of the asymptotic error for large networks. The fit seems to indicate that the mean squared error goes to zero in the limit $N \to \infty$ in both cases, with exponent 1.8 ± 0.2 for the Extended Plefka method and 0.89 ± 0.15 for the algorithm with averaged moments.

complex local fields require a Monte Carlo simulation to be computed, which makes the method numerically complex. A less complex version of the Monte Carlo algorithm, where the second-order moments are replaced by their self-averaging value, performs equally well only for very large systems. The result outperforms the other methods in predicting the single site magnetization, and the mean squared error of the predicted magnetization goes to zero with increasing system size as a power low. This result, together with the retrieval of the exact equations for the fully asymmetric case, seems to indicate that the approximation should become exact in the thermodynamic limit of a large network. A yardstick for comparison can be the results of [EO94], where the transient zero-temperature dynamics of an Ising model with arbitrary degree of symmetry was derived for the disordered average system. Our equation for the magnetisation shows strong analogies with their result; as a next step, we intend to perform the proper comparison by averaging our mean-field solution over the quenched couplings. The complexity of performing this average lies in the nonlinear dependence of the magnetisations on both the couplings and the magnetisations at previous times, which in turn depend on the couplings. We expect that the computation could be treated using the replica trick.

The mean-field equations that we discussed in this chapter can be also used as inference tools to compute the value of the couplings from data. A future direction of the present work consists in designing a mean-field estimator for the couplings starting from the extended Plefka expansion; for example, one could extend the approach of [MS11, MS14], where a relation between equal times and one-time delayed correlation matrices is derived, and then inverted to compute the couplings from correlations observed from data.

Along with developing more accurate mean-field inference techniques, we find it important to assess the quality of mean-field estimators against other types of estimators. In the next chapter, we carry out such analysis for fully asymmetric networks, for which the exact mean-field equations are known.

Appendix

3.A TAP equations for the SK model

The TAP equations were derived by Thouless Anderson and Palmer [TAP77] in 1977 and provided the mean field solution to the Sherrington-Kirkpatrick (SK) model of spin glasses [SK75].

$$H(\sigma) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i^{ex}\sigma_i, \tag{3.3}$$

where the couplings J_{ij} are independent Gaussian random variables with zero mean and variance J_0/N, and h_i^{ex} are external local fields. The TAP equations consists in a set of mean field equations for the local magnetization, valid for a given realization of the random couplings J_{ij}:

$$m_i = \tanh(\beta\tilde{h}_i) + \mathcal{O}(1/N) \tag{3.4}$$

where

$$\tilde{h}_i = h_i^{ex} + \sum_j J_{ij}\left(m_j - m_i\sum_j J_{ij}\chi_{jj}\right) \tag{3.5}$$

$$= h_i^{ex} + \sum_j J_{ij}m_j - m_i\sum_j J_{ij}^2\beta(1-m_j^2). \tag{3.6}$$

The last equality follows from the definition of the susceptibility χ_{jj}, which represents the reaction of the magnetization m_j to a small change of the field $\tilde{h}_j$:

$$\chi_{jj} = \frac{\partial m_j}{\partial\tilde{h}_k} = \beta(1-m_k^2). \tag{3.7}$$

In the following two sections, we present two methods to derive these equations.

3.A.1 The cavity approach

The TAP equations for the local magnetizations $m_i = \langle\sigma_i\rangle$ of the SK model can be derived from the cavity [MPV87] method (see [OS01, Nis01, OW01a]).

We start by deriving an set of approximate equations for the single site marginal distribution of spins $P_i(\sigma_i)$. The key observation is that the single spin σ_i depends on the other spins through the local fields $h_i = \sum_j J_{ij}\sigma_j$, and that the joint distribution of σ_i and h_i is

$$P(\sigma_i, h_i) \propto e^{\beta\sigma_i(h_i+h_i^{ex})} P(h_i \backslash \sigma_i). \tag{3.8}$$

$P(h_i\backslash\sigma_i)$ is the distribution of the local field h_i in an auxiliary system of $N-1$ spins, where the spin σ_i has been removed by setting $J_{ij} = 0$ for $j \neq i$. It is called the cavity distribution, and can be explicitly written as

$$P(h_i\backslash\sigma_i) \equiv \sum_{\boldsymbol{\sigma}\backslash\sigma_i} \delta(h_i - \sum_j J_{ij}\sigma_j) P(\boldsymbol{\sigma}\backslash\sigma_i), \tag{3.9}$$

where $P(\boldsymbol{\sigma}\backslash\sigma_i)$ is the distribution of the spin configuration in a system where the spin σ_i has been removed. Hence, the marginal distribution of single spins can be found once the cavity distribution of local field is specified:

$$P_i(\sigma_i) \propto \int dh_i e^{\beta\sigma_i(h_i+h_i^{ex})} P(h_i\backslash\sigma_i). \tag{3.10}$$

Let us consider the distribution (3.9). The SK model is a fully connected system, and the number of terms in the sum $\sum_j J_{ij}\sigma_j$ is $N-1$. If all the spins in the sum were independent and identically distributed, the central limit theorem would tell us that the cavity distribution (3.9) is Gaussian. We assume that this is the case for the SK model, where correlations of different sites σ_j are weak[4]:

$$P(h_i\backslash\sigma_i) \approx \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{(h_i - \langle h_i\rangle_{\backslash i})^2}{2V_i}\right), \tag{3.11}$$

where $\langle ...\rangle_{\backslash i}$ denotes averages with respect to the cavity distribution. By inserting this Gaussian distribution in (3.10) one finds the following equation for the single site magnetization

$$m_i = \tanh \beta \langle h_i\rangle_{\backslash i}. \tag{3.12}$$

We now have to evaluate the expectations $\langle h_i\rangle_{\backslash i}$ and the variances V_i.

The definition of full expectation

$$\langle h_i\rangle = \sum_{\sigma_i} \int dh_i\, h_i P(\sigma_i, h_i), \tag{3.13}$$

together with (3.8) and the Gaussian cavity field (3.11) yields

$$\langle h_i \rangle = \langle h_i \rangle_{\setminus i} + V_i \langle \sigma_i \rangle. \tag{3.14}$$

The variance of the cavity field is defined as

$$V_i = \sum_{ij} J_{ij} J_{ik} (\langle \sigma_j \sigma_k \rangle_{\setminus i} - \langle \sigma_j \rangle_{\setminus i} \langle \sigma_k \rangle_{\setminus i}). \tag{3.15}$$

It can be shown [MPV87] that, in the above equation, only diagonal terms $j = k$ contribute to the sum; we do not prove the validity of this result, but only mention that it is due to the independence of J_{ij} and J_{ik} $(i \neq k)$ and to the property called clustering

$$\lim_{N \to \infty} \frac{1}{N^2} (\langle \sigma_j \sigma_k \rangle - \langle \sigma_j \rangle \langle \sigma_k \rangle) = 0, \tag{3.16}$$

that was shown to be valid when the temperature is sufficiently high and there is only one solution to (3.4-3.6). We finally get

$$V_i \approx \sum_j J_{ij}^2 (1 - \langle \sigma_j \rangle_{\setminus i}^2) \approx \sum_j J_{ij}^2 (1 - \langle \sigma_j \rangle^2) = \sum_j J_{ij}^2 (1 - m_j^2), \tag{3.17}$$

by assuming that the average in the original system and the average in the auxiliary system where one spin has been removed are approximately the same. From (3.12), (3.14) and (3.17) one gets the TAP equations (3.6).

3.A.2 Plefka's expansion

Another method to derive TAP equations is the Plefka [Ple82] expansion. It consists in a weak coupling expansion of the free energy at fixed magnetizations, which are constrained through a Lagrange transform (i.e., the Gibbs potential):

$$-\beta G_\alpha(\beta, \boldsymbol{m}) = Extr_h \left(\log \mathrm{Tr} e^{-\beta H_\alpha} - \beta \sum_i h_i m_i \right), \tag{3.18}$$

where

$$H_\alpha = -\alpha \sum_{i<j} J_{ij} \sigma_i \sigma_j - \sum_i \sigma_i (h_i^{\mathrm{ex}} + h_i). \tag{3.19}$$

[4] For a more detailed discussion of the validity of the mentioned approximation, see [MPV87].

the parameter α is introduced to control the strength of the couplings, $\alpha = 0$ corresponding to the non-interacting system, while a $\alpha = 1$ to the the SK model. The auxiliary fields h_i are Lagrange multipliers that enforce the constraint

$$m_i = \langle \sigma_i \rangle_\alpha, \tag{3.20}$$

where $\langle ... \rangle_\alpha$ is the expectation with respect to the Hamiltonian (3.19). These auxiliary fields are to be considered as functions of the moments, according to the following set of equations

$$h_i[\boldsymbol{m}] = \frac{\partial G_\alpha}{\partial m_i}, \tag{3.21}$$

where we have used the condition (3.20). Note that when the auxiliary fields h_i are set to zero, the Gibbs free energy (3.18) is equivalent to the unconstrained equilibrium free energy, so that the condition for the equilibrium magnetization is

$$\frac{\partial G_\alpha}{\partial m_i} = 0. \tag{3.22}$$

The idea is to expand (3.18) in powers of α and set $\alpha = 1$ at the end of the calculations to recover the result for the SK model:

$$G_\alpha = G_0 + \left.\frac{\partial G}{\partial \alpha}\right|_{\alpha=0} + \frac{1}{2}\left.\frac{\partial^2 G}{\partial \alpha^2}\right|_{\alpha=0} \alpha^2 + \mathcal{O}(\alpha^3). \tag{3.23}$$

Let us compute it term by term. The Gibbs potential of the non-interacting Ising system is computed to

$$\beta G_0 = \sum_i h_i^{ex} m_i + \frac{1}{2}\sum_i [(1+m_i)\log\frac{1}{2}(1+m_i) + (1-m_i)\log\frac{1}{2}(1-m_i)], \tag{3.24}$$

while the first and second derivatives give, respectively,

$$\begin{aligned} \frac{\partial G}{\partial \alpha} &= \langle H^{\text{int}} \rangle_\alpha, \\ \frac{\partial^2 G}{\partial \alpha^2} &= -\beta \left\langle H^{\text{int}} \left(H^{\text{int}} - \langle H^{\text{int}} \rangle_\alpha - \sum_i \frac{\partial h}{\partial \alpha}(\sigma_i - m_i) \right) \right\rangle_\alpha , \end{aligned} \tag{3.25}$$

where $H^{\text{int}} = \partial H / \partial \alpha$ is the interacting part of the Hamiltonian. By evaluating the above equations at $\alpha = 0$ we find

$$\begin{aligned} \left.\frac{\partial G}{\partial \alpha}\right|_{\alpha=0} &= -\frac{1}{2}\sum_{i\neq j} J_{ij} m_i m_j, \\ \left.\frac{\partial^2 G}{\partial \alpha^2}\right|_{\alpha=0} &= -\frac{1}{2}\beta \sum_{i\neq j} J_{ij}^2 (1-m_i^2)(1-m_j^2), \end{aligned} \tag{3.26}$$

where we have used the condition (3.20) and where the latter equation follows from the relation:

$$\frac{\partial h_i}{\partial \alpha} = \frac{\partial}{\partial m_i}\frac{\partial G}{\partial \alpha}\bigg|_{\alpha=0} = -\sum_{j,j\neq i} J_{ij} m_j, \tag{3.27}$$

which, in turn, follows from (3.21). Inserting (3.24) and (3.26) in (3.23) we get the final equation for the Gibbs potential:

$$\begin{aligned}\beta G_\alpha(\beta, \boldsymbol{m}) = &\sum_i h_i^{ex} m_i + \frac{1}{2}\sum_i [(1+m_i)\log\frac{1}{2}(1+m_i) + (1-m_i)\log\frac{1}{2}(1-m_i)] \\ &- \frac{\beta\alpha}{2}\sum_{i\neq j} J_{ij} m_i m_j - \left(\frac{\beta\alpha}{2}\right)^2 \sum_{i\neq j} J_{ij}^2 (1-m_i^2)(1-m_j^2) + \mathcal{O}(\alpha^3).\end{aligned} \tag{3.28}$$

For $\alpha = 1$, if terms $\mathcal{O}(\alpha^3)$ are neglected, the TAP equations for the magnetization are then recovered by extremization of (3.28) with respect to m_i, according to (3.22). In [Ple82], Plefka shows that these higher order terms in (3.28) can be neglected in the $N \to \infty$ limit, as long as the system is not in the spin glass phase.

3.B Mean field approaches to the kinetic Ising model: previous results

Let us now review the first generalization to dynamics of Plefka's expansion, as proposed in [RH11b]. We consider the kinetic Ising model introduced in section 1.2, composed of N Ising spins $s_i(t)$ interacting through couplings J_{ij} and with local external fields $h_i(t)$. It evolves in time according to a Glauber dynamics with parallel update rule. The distribution of spin trajectories has the following Markovian form

$$\mathrm{P}(\sigma_{0:T}) = \prod_{t=0}^{T-1} \mathrm{P}(\sigma(t+1)|\sigma(t))\mathrm{P}\left(\sigma(0)\right), \tag{3.29}$$

where the transition probability is given by

$$\mathrm{P}(\sigma(t+1)|\sigma(t)) = \prod_{i=1}^{N} \frac{\exp \sigma_i(t+1)[h_i^{\mathrm{ex}}(t) + \sum_{j=1}^N J_{ij}\sigma_j(t)]}{2\cosh[h_i^{\mathrm{ex}}(t) + \sum_{j=1}^N J_{ij}\sigma_j(t)]}. \tag{3.30}$$

Standard field theoretical manipulations [ZJ02] lead to define the following Martin-Siggia-Rose generating functional [MSR73, DDP78]:

$$Z_\alpha[\boldsymbol{\psi},\boldsymbol{h}] = \int \prod_{i,t} (dg_i(t)d\hat{g}_i(t)) \operatorname{Tr}_\sigma \prod_{it} \exp\left\{ i\hat{g}_i(t)\left[h_i(t) - \alpha \sum_j J_{ij}\sigma_j(t)\right]\right.$$
$$\left. +\sigma_i(t+1)g_i(t) - \log 2\cosh g_i(t) - h_i(t)\hat{g}_i(t) + \psi_i(t)\sigma_i(t)\right\} \tag{3.31}$$

where α is introduced to control the coupling strength. By derivatives of the generating functional we find the moments

$$-i\hat{\mu}_i(t) \doteq \frac{\partial \log Z_\alpha[\boldsymbol{\psi},\boldsymbol{h}]}{\partial H_i(t)} = -i\langle \hat{g}_i(t)\rangle_\alpha \tag{3.32}$$

$$\mu_i(t) \doteq \frac{\partial \log Z_\alpha[\boldsymbol{\psi},\boldsymbol{h}]}{\partial \psi_i(t)} = \langle \sigma(t)\rangle_\alpha \tag{3.33}$$

where $\langle \dots \rangle_\alpha$ denotes expectation under the measure inside the integral in (3.31). Note that by taking the limits $\psi \to 0$, $h \to h^{\text{ex}}$ and setting $\alpha = 1$ at the end of the calculation, (3.33) reduces to the magnetization

$$m_i(t) \doteq \langle \sigma(t)\rangle_P = \lim_{\psi\to 0} \mu_i(t) \tag{3.34}$$

averaged over the distribution (3.29). In this framework, Hertz and Roudi derive a set of mean field equations by extending the Plefka expansion for the SK model [Ple82] to the dynamical case. To this end, they draw a parallel between the logarithm of the generating functional and the Helmholtz free energy in the equilibrium statistical mechanics; accordingly, the Legendre transform of $\log Z_\alpha$ corresponds to the Gibbs free energy. While the original Plefka expansion consisted in a weak coupling expansion of the Gibbs free energy, now the Legendre transform of $\log Z_\alpha$ at fixed moments (3.32-3.33) is Taylor expanded in powers of α, around $\alpha = 0$. At the end of the calculation one sets $\alpha = 1$. The Legendre transform of $\log Z$ with respect to the real and auxiliary fields is

$$\Gamma_\alpha[\boldsymbol{\mu},\hat{\boldsymbol{\mu}},] = \log Z_\alpha[\boldsymbol{\psi},\boldsymbol{h},] - \sum_{it} \psi_i(t)\mu_i(t) + ih_i(t)\hat{\mu}_i(t) \tag{3.35}$$

where the fields $\boldsymbol{\psi},\boldsymbol{h}$ are functions of the moments (3.32-3.33) according to the following equations:

$$\begin{aligned} \frac{\partial \Gamma[\boldsymbol{\mu},\hat{\boldsymbol{\mu}}]}{\partial m_i(t)} &= -\psi_i(t) \\ \frac{\partial \Gamma[\boldsymbol{\mu},\hat{\boldsymbol{\mu}}]}{\partial \hat{m}_i(t)} &= ih_i(t). \end{aligned} \tag{3.36}$$

By expanding (3.35) in powers of α and considering the equations (3.36) within expansion, one finds the following 'naive' mean field equation

$$m_i(t+1) = \tanh\left[h_i^{\text{ex}}(t) + \sum_j J_{ij} m_j(t)\right] \tag{3.37}$$

at the first order; a second order expansion yields TAP-like equations,

$$m_i(t+1) = \tanh\left[h_i^{\text{ex}}(t) + \sum_j J_{ij} m_j(t) - m_i(t+1) J_{ij}^2 [1 - m_j(t)^2]\right], \tag{3.38}$$

that should be solved self-consistenly for $m_i(t+1)$ at each time step. In this formulation, the generating functional (3.31) is defined in terms of fields that act linearly on the degrees of freedom and the Legendre transform (3.35) is performed by fixing the first order statistics over time. In Paper 1 we will observe that, in contrast to the equilibrium case, the mean field description is considerably improved if we take into account also the second order statistics -namely correlations and responses. In particular, we will introduce these quadratic terms in the generating functional, so that the second order moments can be easily found by first derivatives of the generating functional.

In the case of a completely asymmetric network, where the couplings have variance scaling as $1/N$, the correlations between spins at different times is small [CS88] and the central limit theorem can be used to describe the statistics of local fields. This the approach followed in [MS11], where the field acting on each spin at site i and time-step t:

$$\sum_j J_{ij} \sigma_j(t-1)$$

is treated as a Gaussian distributed field with mean

$$g_i(t-1) = \sum_j J_{ij} m_j(t-1) \tag{3.39}$$

and variance

$$\Delta(t-1) = \sum_j (1 - m_j(t-1)^2). \tag{3.40}$$

From the definition of the dynamics (3.29), one retrieves the following equation for the magnetization:

$$m_i(t) = \int \mathcal{D}x \tanh\left[g_i(t-1) + h_i^{\text{ex}}(t-1) + x\sqrt{\Delta_i(t-1)}\right], \tag{3.41}$$

where the integral is over the Gaussian noise $\mathcal{D}x = dx\, e^{-x^2/2}/(2\sqrt{\pi})$.

A comparison between the approximations (3.38) and (3.41) is presented in [SRMH12], where the limit of validity of the two methods are compared. An extension of the latter method to the case of arbitrary degree of symmetry is derived in [MS14]. The effective local fields are still assumed to be Gaussian distributed, but a non-zero covariance between spins at different times is also considered. The resulting magnetization is

$$m_i(t) = \int \mathcal{D}x \tanh\left[g_i(t-1) + h_i^{\text{ex}}(t-1) + x\sqrt{V_{ii}(t-1,t-1)}\right], \quad (3.42)$$

where $g_i(t)$ was defined in (3.39), and

$$V_{ij}(t,s) = \langle \delta g_i(t) \delta g_j(s) \rangle \quad (3.43)$$

is the covariance of the field fluctuations $\delta g_i(t) = \sum_j J_{ij}\delta\sigma_j(t-1)$,where $\delta s_i(t-1) = s_i(t) - m_i(t)$. Starting from cavity arguments, the authors analytically derive a set of recursive equations for calculating covariances at different times in terms of covariances at previous times. For time indices $s \leq t$ the result is

$$\boldsymbol{V}(t,s) = \boldsymbol{J}^\top \boldsymbol{A}(t-1)\boldsymbol{V}(t-1,s) + \boldsymbol{J}^\top \boldsymbol{C}(t-1,s-1)\boldsymbol{J}, \quad (3.44)$$

where $C_{ij}(t,s) = \delta_{ij}\langle \delta s_i(t)\delta s_i(s)\rangle$ is the auto-covariance function and

$$A_{ij}(t) = \delta_{ij}\int \mathcal{D}x \left[1 - \tanh^2\left(g_i(t) + h_i^{\text{ex}}(t) + x\sqrt{V_{ii}(t,t)}\right)\right].$$

Note that to compute the magnetization (3.42) one needs $V_{ii}(t-1,t-1)$, which is obtained from (3.44) in terms of past values of $\boldsymbol{V}$ and m, and in terms of the auto-covariance function at two successive time steps; the latter quantity is compute to be

$$\begin{aligned} C_{ii}(t-1,t) = \int dh_i(t) \int dh_i(t-1)\, p(h_i(t), h_i(t-1)) \\ \tanh\left[h_i(t) + H_i(t)\right] \tanh\left[h_i(t-1) + h_i^{\text{ex}}(t-1)\right], \end{aligned} \quad (3.45)$$

where $p(h_i(t), h_i(t-1))$ is a bivariate Gaussian distribution with mean vector

$$\left(g_i(t)\,,\, g_i(t-1)\right)^\top$$

and covariance matrix given by

$$\begin{pmatrix} V_{ii}(t,t) & V_{ii}(t,t-1)\sqrt{V_{ii}(t,t)V_{ii}(t-1,t-1)} \\ V_{ii}(t,t-1)\sqrt{V_{ii}(t,t)V_{ii}(t-1,t-1)} & V_{ii}(t-1,t-1) \end{pmatrix}.$$

3.C Algorithm with averaged moments

Keeping the same notation, the algorithm of (Paper 1, section 5) is modified as follows.

- Initial condition: set $s_i^k(0) = 1, \quad i = 1...N,\ k = 1...N_T$.
- For $t = 1...T$:
  1. Draw the spins at time t from
  $$p(s_i^k(t)) = \frac{e^{s_i^k(t) g_i^k(t-1)}}{2\cosh g_i^k(t-1)}, \quad \text{for } i = 1...N,\ k = 1...N_T,$$
  using the fields $g_i^k(t-1)$ calculated at the previous time step.
  2. Compute correlations
  $$C_i(t,t') = \overline{\tanh[g_i(t-1)] s_i(t')}, \quad \text{for } t' = 1...t-1,\ i = 1...N,$$
  and their averaged value $\langle C(t,t')\rangle \doteq \frac{1}{N}\sum_i C_i(t,t')$.
  3. Draw the Gaussian random variable $\phi_i^k(t)$ with mean
  $$\widehat{\phi}_i^k(t) = \sum_{r=0}^{t-1} a(r)\phi_i^k(r), \tag{3.46}$$
  and covariance
  $$E\{[\phi_i^k(t) - \widehat{\phi}_i^k(t)]^2\} = 1 - \frac{g^2}{N}\sum_j m_j^2(t) - \sum_{r=0}^{t-1} a(r)\langle C(t,r)\rangle, \tag{3.47}$$
  where the coefficient $\{a(0)\dots a(t-1)\}$ are computed from
  $$\sum_{r=0}^{t-1} a(r)\langle C(r,t')\rangle = \langle C(t,t')\rangle, \quad t' = 0\dots t-1.$$
  4. Compute the sample averages that will be needed in (5):
  $$\overline{s_i(t)\phi_i(t')} = \overline{\tanh[g_i(t-1)]\phi_i(t')}, \quad \text{for } t' = 1...t-1,\ i = 1...N.$$
  5. Compute $\langle R(t,t')\rangle \doteq \frac{1}{N}\sum_i R_i(t,t')$, for $t' = 1...t-1$ by solving the system of linear equations:
  $$\frac{1}{N}\sum_i \overline{s_i(t)\phi_i(t')} = \sum_{\tau=1}^{t-1}\langle R(t,\tau)\rangle g^2[\langle C(\tau,t')\rangle - \frac{1}{N}\sum_j m_j(\tau)m_j(t')].$$

6. Compute the fields

$$g_i^k(t) = \phi_i^k(t) + \sum_j \Big(J_{ij} m_j(t) - g^2 \frac{1-k^2}{1+k^2} \sum_{t'=0}^{t-1} J_{ij} J_{ji} \langle R(t,t') \rangle [s_i^k(t') - m_i(t')] \Big) + h_i(t),$$
$$\text{for } i = 1...N,\ k = 1...N_T.$$

7. Compute the magnetizations at time $t+1$:

$$m_i(t+1) = \overline{\tanh[g_i(t)]}, \quad \text{for } i = 1...N.$$

4 Learning in kinetic Ising models

4.1 Introduction

In Chapter 3 , we discussed mean-field approaches to the forward problem for the kinetic Ising model. Given a specific set of model parameters, we described the time evolution of system observables, such as magnetisations and correlations.

Here, we focus on the inverse problem: based on a set of measurements from the system, we want to infer the model parameters (i.e., couplings between the spins and external fields). The amount of information encoded in the data will affect the quality of parameter estimation and it is important to quantify how the performance of the inference algorithm depends on the size of the data set. This question is particularly relevant in the context of the new high-throughput data collection techniques, where the number of variables that can be simultaneously recorded is almost as large as the number of possible trials [AG16].

In this Chapter, we assume to have access to time series data of length T for a system of N spins, specifying the value of each spin at successive time points. A widely used estimator for the parameters is the maximum likelihood estimator, which converges in probability to the true value of the parameters when the size of the dataset (rescaled by the size of the system) tends to infinity, with the lowest possible asymptotic mean squared error [Cra16]. The likelihood can be computed in polynomial time in T and N, which makes the computation much faster with respect to the equilibrium case (see section 5.A). Still, maximum likelihood conditions must be computed at every step of the iteration, based on the current value of the parameters, and the iteration can take a long time to converge, also depending on the choice of initial conditions and learning rate. A much faster method is provided by approximate techniques, such as the mean-field methods discussed in chapter 3.

The first focus of this chapter is to analyse the theoretical performance of estimators based on a mean-field approximation. In a mean-field framework, inference in kinetic Ising models was initially studied based on data from their non-equilibrium steady state. The TAP equations (3.6) for the magnetisation derived at equilibrium for the SK model were argued to be valid for the

asynchronously [KS00] and synchronously [RH11b] updated Glauber dynamics. Based on those equations, a linear relation between the one-time-delayed and equal-time correlation matrix, respectively denoted by $\boldsymbol{D}$ and $\boldsymbol{C}$, can be found:

$$\boldsymbol{J} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{C}^{-1}, \tag{4.1}$$

where the matrix A encodes for the details of the considered approximation (Appendix 4.C). If the true correlation matrices are replaced by the empirical ones, this relation provides a linear estimator for the couplings, which can be simply computed via a matrix inversion. More recently, these results have been extended to transient dynamics. In chapter 3, we saw that the mean-field description is particularly simple in the case of an asymmetric network, where a linear relation between one-time-delayed and equal-time correlation matrices provides the exact solution in the thermodynamic limit. This relation can be easily inverted to infer the couplings in the same form as (4.1), where now the correlation matrices depend on time (see section 4.C). Ideally, correlations are computed from multiple trials of spin trajectories. However, it is hard to have access to such data, and averages over trials are replaced by averages over time [MS11]. Hence, the quality of the estimator will depend on the length of the observed spin trajectories.

In this framework, we aim to compute the error associated with the linear mean-field estimator and study how it scales with the length of the observed trajectories.

The theoretical framework for our analysis is given by the statistical mechanics of learning [EVdB01], where the phase space consists of the couplings to be inferred, while the spin values are considered to be fixed observations. We work in the so called student-teacher scenario, where the data are generated independently from a teacher network and a learning algorithm adapts the couplings of a student network as estimator for the teacher. The error associated to the algorithm is given by the average mean squared error between the teacher coupling vector and the student one, where averages are computed by using the replica method of statistical physics [MPV87, Nis01] [1].

In this chapter, we extend the replica formalism used for learning of perceptrons to the kinetic Ising model. In the large N limit, two-times correlations can be neglected in asymmetric networks, for which the memory of the system

[1] Replica calculations have been widely applied to problems related to learning in feed-forward neural networks [WRB93, SST92, OK96], following the seminal work of Gardner [Gar87, Gar88], who exploited it to compute the critical capacity of the perceptron with continuous synaptic weights; more recent applications also include communication theory [GV05], compressed sensing [GBS09, RGF09, GS10, KMS+12],matrix factorization [KKM+16] and and high-dimensional regression [AG16].

is lost after one time step (see, e.g., [CS87]). This allows us to use a central limit theorem argument to treat the probability distribution underlying the Markovian dynamics as the distribution of T independent perceptrons, in the thermodynamic limit of a large system. In each perceptron (corresponding to each time step), the inputs are not independent but spatially correlated through the equal-time correlation matrix. Surprisingly, we find that the equal-time spatial correlation matrix has a non-negligible influence on the estimation error. We compute the statistics of this random correlation matrix and obtain an explicit result for the estimation error as a function of the growing length of observed trajectories.

In section 4.3, we study the performance of other two approximate algorithms. First, within the class of estimators that minimize a local cost function which is quadratic in the couplings, we consider the optimal one, minimising the mean square error of estimated parameters. Then, we turn to a Bayesian probabilistic formulation. Introducing a prior distribution, the Bayes optimal estimator of the parameters is given by their posterior expectation. Since computing posterior averages exactly is intractable, we propose an analytic approximation to the posterior expectations based on cavity arguments and design an efficient algorithm to numerically implement our solution. Finally, we use an analogous formalism to the one developed in Paper 2 to compute the error of the Bayes optimal estimator and compare it with the mean-field and linear optimal estimators.

An introduction to the replica method, applied to the physics of spin glasses and to the problem of learning in neural networks is presented in section 4.A.1; the general framework for the statistical mechanics of learning is also briefly explained. The derivation of the maximum likelihood estimator and of the linear mean field estimator - both for the stationary and for the transient dynamics - is given in sections 4.B, 4.C, 4.D, respectively.

4.2 Paper 2.

Author's contribution: I performed the analytical and numerical calculations, prepared the figures and contributed to writing the paper.

Learning of couplings for random asymmetric kinetic Ising models revisited: random correlation matrices and learning curves

J. Stat. Mech. (2015) P09016

(http://iopscience.iop.org/1742-5468/2015/9/P09016)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Learning of couplings for random asymmetric kinetic Ising models revisited: random correlation matrices and learning curves

Ludovica Bachschmid-Romano and Manfred Opper

Department of Artificial Intelligence, Technische Universität Berlin, Marchstraße 23, Berlin 10587, Germany
E-mail: ludovica.bachschmidromano@tu-berlin.de and manfred.opper@tu-berlin.de

Received 13 March 2015
Accepted for publication 21 July 2015
Published 16 September 2015

Online at stacks.iop.org/JSTAT/2015/P09016
doi:10.1088/1742-5468/2015/09/P09016

Abstract. We study analytically the performance of a recently proposed algorithm for learning the couplings of a random asymmetric kinetic Ising model from finite length trajectories of the spin dynamics. Our analysis shows the importance of the nontrivial equal time correlations between spins induced by the dynamics for the speed of learning. These correlations become more important as the spin's stochasticity is decreased. We also analyse the deviation of the estimation error

Keywords: network reconstruction, learning theory, statistical inference, kinetic Ising model

J. Stat. Mech. (2015) P09016

1742-5468/15/P09016+19$33.00

Contents

1. Introduction 2
2. Estimators 3
3. Learning curves from the replica approach 5
4. Statistics of correlation matrices 8
5. Results 11
6. Outlook 14

Acknowledgments 14

Appendix A. Details of the replica calculation of the free energy 15

Appendix B. Derivation of the generating function 16

Appendix C. Independence of the J and $B(t)$ matrices: an example 16

Appendix D. Padè Approximant 17

Appendix E. Details on the statistics of the correlation matrix 17

Appendix F. Asymptotic order parameters for ML estimator 18

References 18

J. Stat. Mech. (2015) P09016

1. Introduction

Recently, the learning of synaptic couplings for a recurrent neural network modelled by a kinetic Ising model with random couplings has attracted attention in the statistical physics community, see e.g. [1–10]. The model is defined by a system of N Ising spins σ_i connected through couplings J_{ij}. We assume throughout the paper that the interactions are non–symmetric, i.e. we have $J_{ij} \neq J_{ji}$ and $J_{ii} = 0$. The system evolves in discrete time according to a synchronous parallel dynamics, where spins at time $t+1$ are updated independently with transition probability (specialised on the case of no external fields)

$$P(\sigma_i(t)|\{\sigma_j(t-1)\}_{j=1}^{N}) = \frac{\mathrm{e}^{\beta\sigma_i(t)\sum_j J_{ij}\sigma_j(t-1)}}{2\cosh(\beta\sum_j J_{ij}\sigma_j(t-1))}. \tag{1}$$

We are interested in learning the spin couplings J_{ij}, assuming that a complete trajectory $\{\boldsymbol{\sigma}\}_{0:T} = \{\sigma_i(t)\}_{i=1,\ldots,N,t=1,\ldots,T}$ of length T for all spins is observed. A well known solution to this problem is given by the method of maximum likelihood, which leads

to a set of coupled nonlinear equations which have to be solved by iteration. A computationally much simpler and elegant solution valid for large networks with random couplings which avoids an iterative solution was recently presented in [1]. This solution is based on an exact mean field (EMF) expression for spin correlations which can be explicitly solved for the couplings. The EMF estimator replaces exact correlations by empirical correlations which can e.g. be computed from a single spin trajectory. Simulations have shown good agreement between true and estimated couplings [1].

Of course, if there is only a limited number of observations available there will be a nonzero estimation error for the EMF method. One may then ask how much one has to pay for the numerical efficiency of the algorithm in terms of a loss in statistical efficiency. Hence, we would like to investigate at what rate the error decreases with growing length of trajectories and if the decrease is slower than that of a statistically efficient estimator such as the maximum likelihood estimator which has an optimal asymptotic rate [11]. Using the replica method we will compute the estimation error of the EMF method in the thermodynamic limit $N \to \infty$ assuming that the data are generated from a kinetic Ising model with true couplings drawn at random from a Gaussian distribution. The analysis of the statistical properties is significantly simplified by the fact that kinetic Ising models with non–symmetric random couplings have spin correlations which decay after a single time step (see for example [12]) and computations of learning curves resemble those for temporally independent data. A nontrivial aspect however is the occurrence of equal time spin correlations of the spin dynamics. We compute an exact result for the statistics of the random correlation matrix. From this it is possible to obtain an explicit expression for the learning curve for the EMF algorithm and the asymptotics of the ML estimator.

2. Estimators

The EMF estimator [1] is based on a linear relation between the time-delayed and the equal time correlator matrices,

$$C_{ij} = \langle \delta\sigma_i(t)\delta\sigma_j(t)\rangle, \qquad D_{ij} = \langle \delta\sigma_i(t+1)\delta\sigma_j(t)\rangle, \tag{2}$$

for the spin fluctuations $\delta\sigma_j(t) \doteq \sigma_j(t) - m_j(t)$, where $m_j(t)$ denotes the local magnetisation at time t and the brackets $\langle \ldots \rangle$ denote expectation with respect to the spin dynamics (1). Here we assume stationarity for which the matrices are time independent. If the couplings J_{ij} are assumed to be mutually independent Gaussian random variables, with zero meand and variance $1/N$, the following mean field relation is found to be exact in the thermodynamic limit $N \to \infty$:

$$D_{ij} = a_i \sum_k J_{ik} C_{kj}, \tag{3}$$

where

$$a_i = \beta \int \mathcal{D}x \Big[1 - \tanh^2\Big[\beta\Big(H^{\text{ext}} + x\sqrt{\Delta_i}\Big)\Big]\Big], \qquad \Delta_i = \sum_j J_{ij}^2 (1 - m_j^2) \tag{4}$$

J. Stat. Mech. (2015) P09016

and $\mathcal{D}x$ is the normal Gaussian measure. Throughout the paper we will specialise to the case of zero external field and vanishing initial magnetisations. In this case we have $m_i(t) = 0$, $H^{\text{ext}} = 0$, $\Delta_i = 1$ and

$$a_i = a = \beta \int \mathcal{D}x[1 - \tanh^2(\beta x)] \tag{5}$$

is independent of time. For the estimator the exact correlation matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ are approximated by empirical averages using a long trajectory of spins (assuming zero magnetisations):

$$C_{ij} \to \hat{C}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \sigma_i(t)\sigma_j(t), \qquad D_{ij} \to \hat{D}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \sigma_i(t+1)\sigma_j(t). \tag{6}$$

One can then obtain the couplings by inverting (3) as follows:

$$J_{ij} = \frac{1}{a}\sum_k \hat{D}_{ik}\hat{C}^{-1}_{kj}. \tag{7}$$

It is easy to see that the EMF estimator can be rephrased as the minimiser of the following cost function

$$\mathrm{E}^{\mathrm{i}}_{\mathrm{MF}} = \frac{1}{2}\sum_{t=1}^{T}\left(\sigma_i(t) - a\sum_j J_{ij}\sigma_j(t-1)\right)^2 \tag{8}$$

with respect to the couplings $\{J_{ij}\}_{j=1}^N$. Note that the estimation of the ingoing couplings $\{J_{ij}\}_{j=1}^N$ for each spin i can be treated separately for the coupling distribution we are considering. The EMF estimator is based on simple explicit computation (inversion of the correlation matrix in (7), which is possible if the parameter $\alpha = T/N$ is grater than 1) which makes the method fast. Other estimators such as the well known maximum likelihood method (ML) have to resort to numerical optimisations using iterative algorithms which could become computationally involved for large system sizes N and a large number of data T. The ML estimator maximises the probability of spin histories $\{\boldsymbol{\sigma}\}_{0:T}$ given by

$$P(\{\boldsymbol{\sigma}\}_{0:T}|\boldsymbol{J}) = \prod_{i=1}^{N}\prod_{t=1}^{T} P(\sigma_i(t)|\{\sigma_j(t-1)\}_{j=1}^N)\; P(\sigma(0)), \tag{9}$$

where $P(\sigma(0))$ is the initial probability of spins. Since this probability factorises in the spins i and J_{ij} are assumed independent, the ML estimator for all couplings $\{J_{ij}\}_{j=1}^N$ pointing into spin i minimises the cost function

$$\mathrm{E}^{\mathrm{i}}_{\mathrm{ML}} = \sum_{t=1}^{T}\left(-\beta\sigma_i(t)\sum_j J_{ij}\sigma_j(t-1) + \ln 2\cosh\left(\beta\sum_j J_{ij}\sigma_j(t-1)\right)\right). \tag{10}$$

While minimizing the cost function (8) just requires the computation of the empirical averages $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{D}}$, in order to minimize (10) with respect to J_{ij} one needs to compute the quantity $\sum_t \sigma_j(t)\tanh\left(\beta\sum_j J_{ij}\sigma_j(t)\right)$ that explicitely depends on the current

value of J_{ij} and has to be recomputed at each step of the algorithm, adding a $N_{\rm step} \cdot T$ operation to the calculation. We observe that in order to avoid second order methods in the solution we need a fine tuning of the step size which makes the algorithm fairly slow for large N. Although it is more computationally expensive, the ML estimator has the important property that it is asymptotically (i.e. for $T \to \infty$) *efficient*. This means that the asymptotic convergence of the mean squared estimation error to zero (assuming the model is correct) happens at a rate which is minimal for any (asymptotically) unbiased estimator [11]. In the following we will compute the error of the EMF algorithm in the thermodynamic limit $N, T \to \infty$, keeping α fixed and compare with the asymptotic $\alpha \to \infty$ optimal error rate of the ML estimator.

3. Learning curves from the replica approach

In this section we will introduce the replica method for computing the EMF prediction error as a function of the scaled number of observed data. We will work in a teacher–student scenario [13, 14], where the data are assumed to be generated at random from the dynamics of a teacher network with random couplings J_{ij}^*. We will use the scaling $J_{ij}^* = W_{ij}^*/\sqrt{N}$ and assume that the W_{ij}^* are independent Gaussian random variables with $W_{ij}^* \sim \mathcal{N}(0,1)$. We can treat the estimation of the ingoing couplings $\boldsymbol{W}^* \equiv \{W_{ij}^*\}_{j=1}^N$ for each spin i separately. For the sake of simplicity, in the following we will drop the index i and define $W_j \doteq W_{ij}$. The average square prediction error for any estimator of the couplings given by $\boldsymbol{W}$ is defined as

$$\varepsilon = \frac{1}{N}\overline{\| \boldsymbol{W}^* - \boldsymbol{W} \|^2} = 1 - 2\rho + Q, \tag{11}$$

where we defined

$$\rho = N^{-1}\overline{\boldsymbol{W}^* \cdot \boldsymbol{W}}, \qquad Q = N^{-1}\,\overline{\|\boldsymbol{W}\|^2}. \tag{12}$$

The bar denotes an average over the spin trajectories $\{\boldsymbol{\sigma}\}_{0:T}$ generated with couplings $\boldsymbol{W}^*$ and over the teacher couplings. We will now analyse the performance of algorithms which minimise a cost function of the type

$$E = \sum_{t=1}^{T} \mathcal{E}(\sigma(t), h_t), \qquad h_t = \frac{1}{\sqrt{N}} \sum_j W_j \sigma_j(t-1),$$

such as (8) and (10), on a random finite set of spin trajectories of size T. One can compute average properties such as the order parameters ρ and Q by introducing an auxiliary probability density of couplings,

$$q(\boldsymbol{W}) = \frac{1}{Z}\, \mathrm{e}^{-\nu E(\boldsymbol{W})}, \tag{13}$$

with a formal inverse 'temperature' parameter ν and the partition function

$$Z(\boldsymbol{\sigma}) = \int \mathrm{d}\boldsymbol{W} \mathrm{e}^{-\nu E(\boldsymbol{W})}. \tag{14}$$

For any ν, we can compute disorder averages of 'thermal averages' of variables such as ρ and Q from the quenched average of the free energy per coupling, defined by

$$F = -N^{-1}\nu^{-1}\overline{\log Z(\boldsymbol{\sigma})} = -\nu^{-1}\lim_{n\to 0}\frac{\partial}{\partial n}N^{-1}\log \overline{Z^n(\boldsymbol{\sigma})}. \tag{15}$$

By taking finally the limit $\nu \to \infty$ (zero 'temperature'), the probability density (13) concentrates at the minimum of $E(\boldsymbol{W})$ and we can extract the desired order parameters. To compute the average, we will make the following assumptions. While the spins $\sigma_i(t)$ are still treated as binary random variables, in computing expectations over $\sigma_j(t)$ for $j \neq i$ we assume a central limit theorem to be valid for the fields h_t as sums of a large number of weakly dependent random variables. Hence, we consider only the second order statistics of these variables and treat them as Gaussian random variables. For equal times the corresponding Gaussian density would be $p(\{\sigma_j(t)\}_{j\neq i}) = \mathcal{N}(0, \boldsymbol{C})$, where the stationary covariance matrix $\boldsymbol{C}$ is a random matrix which itself depends on the random matrix of teacher couplings $\boldsymbol{W}^*$ of the entire network. For different times $t \neq t'$, dependencies between spins $\sigma_j(t)$ and $\sigma_k(t')$ are neglected. This is in accordance with our previous assumptions for $|t - t'| > 1$, but we need an extra argument to justify neglecting D_{jk} giving the correlations at times t and $t + 1$. In principle, $\boldsymbol{D}$ might enter the computation of order parameters as well. Equation (3) shows a relation between the $\boldsymbol{D}$ and $\boldsymbol{C}$ matrices involving the teacher couplings linearly. The arguments presented later in section 4 indicate that for the asymptotic random matrix calculations involving similar relations we can treat teacher couplings and random matrices $\boldsymbol{C}$ as asymptotically independent. Hence, we argue that in an expectation over teacher couplings the contributions due to $\boldsymbol{D}$ vanish. We will see later that the statistical properties of the matrix $\boldsymbol{C}$ will enter the final result of the learning curve through the self averaging moment $C_{-1} \doteq \frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}}$. We will then show in section 4 how this and other moments can be computed. Thus we will include the average over the teacher couplings W^*_{kj} for $k \neq i$ in the statistics of $\boldsymbol{C}$, but we need to perform the average over the teacher couplings $W^*_j \equiv W^*_{ij}$ pointing to spin i explicitly. Finally, the dependencies between random correlation matrices $\boldsymbol{C}$ at different times are also neglected for $N \to \infty$. This results in an effective statistical weight over spin histories given by

$$P(\boldsymbol{\sigma}) \simeq \int \mathrm{d}\boldsymbol{W}^* \mathrm{e}^{-\frac{1}{2}\boldsymbol{W}^*\cdot\boldsymbol{W}^*} \prod_{t=1}^{T}\left\{\frac{\mathrm{e}^{\beta\sigma_i(t)\frac{1}{\sqrt{N}}\sum_j W^*_j\sigma_j(t-1)}}{2\cosh\left[\frac{\beta}{\sqrt{N}}\sum_j W^*_j\sigma_j(t-1)\right]}\, p(\{\sigma_j(t)\}_{j\neq i})\right\}, \tag{16}$$

where the Gaussian measure accounts for our prior knowledge on the teacher couplings distribution. Hence, for large N, we are effectively dealing with the statistical mechanics of a learning problem for a binary classifier neural network (aka logistic regression), where the 'input' data $\sigma_j(t-1)$ are used to predict the 'outputs' $\sigma_i(t)$; the input variables are independent for different t, but have nontrivial 'spatial' correlations given by the matrix $\boldsymbol{C}$. The calculation of the free energy follows the steps of replica calculations for

perceptron learning problems [13–15]. Averages over $\sigma_j(t)$ factorize over time and can be expressed through Gaussian fields h_a for each replicated coupling variable W_a, and fields $u = \frac{1}{\sqrt{N}}\sum_j W_j^*\sigma_j(t-1)$ for the teacher. Under the replica symmetry assumption, which is plausible to be correct for convex cost functions, the covariances are expressed by order parameters

$$\langle u^2\rangle = \frac{1}{N}\sum_{ij} W_i^* C_{ij} W_j^* = 1, \tag{17}$$

$$\langle h_a u\rangle = \frac{1}{N}\sum_{ij} W_i^a C_{ij} W_j^* \doteq R, \tag{18}$$

$$\left\langle h_a^2\right\rangle = \frac{1}{N}\sum_{ij} W_i^a C_{ij} W_j^a \doteq q_0, \tag{19}$$

$$\langle h_a h_b\rangle = \frac{1}{N}\sum_{ij} W_i^a C_{ij} W_j^b \doteq q \qquad a \neq b \tag{20}$$

and the free energy (15) is computed as (appendix A):

$$\begin{aligned} F = -\mathrm{Extr}_{q,R,q_0}\frac{1}{\nu}\Bigg\{ & \frac{1}{2}\frac{q_0 - R^2}{q - q_0} - \frac{1}{2}\log(q - q_0) - \frac{1}{2N}\mathrm{Tr}\log C \\ & + \alpha\sum_{\sigma_0}\int \mathcal{D}t\mathcal{D}y \frac{\mathrm{e}^{\beta\sigma_0\left(\sqrt{1-\frac{R^2}{q}}\,t + \frac{R}{\sqrt{q}}y\right)}}{2\cosh\left[\beta\left(\sqrt{1-\frac{R^2}{q}}\,t + \frac{R}{\sqrt{q}}y\right)\right]} \\ & \log\int \mathcal{D}z \mathrm{e}^{-\nu\mathcal{E}(\sigma_0, \sqrt{q_0 - q}\,z + \sqrt{q}\,y)}\Bigg\}. \end{aligned} \tag{21}$$

The limit $\nu \to \infty$ will occur with $q_0 \to q$, since the different solutions $\boldsymbol{W}$ have to converge to the same minimum. In this limit, keeping the quantity $x \doteq (q_0 - q)\nu$ finite, we finally get

$$\begin{aligned} F = -\mathrm{Extr}_{q,R,x,z}\Bigg\{ & \frac{q - R^2}{2x} + \alpha\sum_{\sigma_0}\int \mathcal{D}t\mathcal{D}y \frac{\mathrm{e}^{\beta\sigma_0\left(\sqrt{1-\frac{R^2}{q}}\,t + \frac{R}{\sqrt{q}}y\right)}}{2\cosh\left[\beta\left(\sqrt{1-\frac{R^2}{q}}\,t + \frac{R}{\sqrt{q}}y\right)\right]} \\ & \left[-\frac{z^2}{2} - \mathcal{E}(\sigma_0, \sqrt{x}\,z + \sqrt{q}\,y)\right]\Bigg\}. \end{aligned} \tag{22}$$

Remarkably, the explicit dependence of F on the correlation matrix (last term in the first line of equation (21)) drops when taking the limit $\nu \to \infty$. Hence, the result we get for F and for the order parameters extremizing F is the same that we would get if the

J. Stat. Mech. (2015) P09016

spins over which we are computing the expectations were independent and the matrix C was not included in the calculation. Still, the correlation matrix affects the error through the parameters ρ and Q defined in (11), which are found to be (appendix A)

$$\rho = R, \tag{23}$$

$$Q = R^2 + (q - R^2)\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}}, \tag{24}$$

where R and q are the order parametrs extremizing the free energy (22). Inserting the above equations in (11) we find the following result for the error:

$$\varepsilon = 1 - 2R + q + (q - R^2)\left(\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}} - 1\right). \tag{25}$$

The last term represents the effect of the correlations of the data on the error and vanishes when $\boldsymbol{C}$ equals the unit matrix. This term can be shown to be positive and leads to an increase in error. In section 5 we will give explicit results for the error of the EMF algorithm.

4. Statistics of correlation matrices

In this section we show how one can compute the stationary value of the negative integer moment of the spin correlations

$$C_{-1} \equiv \lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}(t)}, \tag{26}$$

necessary for the estimation error (25). Here the bar denotes expectation with respect to independent random Gaussian couplings with zero mean and variance $1/N$. Our analysis begins with the time evolution for the correlation matrix $\boldsymbol{C}(t)$ assuming zero magnetisations $m_j(t) = 0$. Following [1], we can assume that in the limit of large N the random variables g_i and g_j, where $g_i = \sum_k J_{ik}\sigma_k(t)$, are zero mean Gaussian random variables with $\langle g_i g_j\rangle = \sum_{kl} J_{ik}C_{kl}(t)J_{lj}$ and $\langle g_i^2\rangle = 1$. An expansion with respect to weak correlations similar to equations (15)–(16) in [1] yields the time evolution

$$\boldsymbol{C}(t+1) = \boldsymbol{I}\gamma(t) + a^2\boldsymbol{J}\boldsymbol{C}(t)\boldsymbol{J}^\top, \tag{27}$$

where $\boldsymbol{I}$ is the unit matrix, $\boldsymbol{C}(0) = \boldsymbol{I}$ and $\boldsymbol{J}$ is the $N\times N$ coupling matrix. The self-averaging quantity γ must be determined such that $C_{ii}(t) = 1$ yielding the condition that $\gamma(t) = 1 - a^2\mathrm{Tr}\overline{\boldsymbol{J}\boldsymbol{C}(t)\boldsymbol{J}^\top}$. Defining $\boldsymbol{B}(t) = \frac{1}{\gamma(t)}\boldsymbol{C}(t)$ and assuming $\gamma(t+1)\approx\gamma(t)$ one finds the simplified iteration

$$\boldsymbol{B}(t+1) = \boldsymbol{I} + a^2\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^\top, \qquad \text{having the solution} \qquad \boldsymbol{B}(t) = \sum_{k=0}^{t} a^{2k}\boldsymbol{J}^k(\boldsymbol{J}^\top)^k. \tag{28}$$

J. Stat. Mech. (2015) P09016

Note that in the limit of small β (small a) one could choose to truncate the sum in (28) to the first order in a (corresponding to $k = 0$) and thus approximatig $\boldsymbol{B}$ by the unit matrix, or to keep the first two orders in a (up to $k = 1$) and thus getting the sum of the unit matrix and a Wishart matrix. From the above equations we get $\gamma \doteq \lim_{t\to\infty}\gamma(t) = \frac{1}{1+a^2}$. We can use (28) to derive an iteration for the generating function of integer moments. In the thermodynamic limit the calculation simplifies remarkably. Consider e.g. the computation of $\lim_{N\to\infty}\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{B}^k(t+1)}$ for some integer k. One would have to deal with terms of the form

$$\frac{1}{N}\mathrm{Tr}(\overline{\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^\top\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^\top\cdots\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^\top}). \tag{29}$$

Writing $\boldsymbol{B}(t)$ as the sum in (28) one is left with a sum of averages involving only the $\boldsymbol{J}$ and $\boldsymbol{J}^\top$ matrices. Given the Gaussian form of the $\boldsymbol{J}$ random matrix, Wick's theorem applies and the expectation in (29) can be computed using diagrammatic techniques. As is well known [16], for $N\to\infty$ only the planar diagrams, i.e. the ones for which lines are not crossing, will contribute to the limit. Besides, note that in the evaluation of (29) the terms containig $\overbrace{\boldsymbol{J}\ldots\boldsymbol{J}}$ and $\overbrace{\boldsymbol{J}^\top\ldots\boldsymbol{J}^\top}$ pairings will vanish because of the asymmetry of the $\boldsymbol{J}$ matrix. It is easy to see (an example is given in appendix C) that this implies that also pairings of the kind $\overbrace{\boldsymbol{B}(t)\ldots\boldsymbol{J}}$ and $\overbrace{\boldsymbol{B}(t)\ldots\boldsymbol{J}^\top}$ are forbidden, where $\overbrace{\boldsymbol{B}(t)\ldots\boldsymbol{J}}$ is a shortcut to indicate the pairing between $\boldsymbol{J}$ and any of the $\boldsymbol{J}$s contained in $\boldsymbol{B}(t)$. Hence, in computing moments by iteration over time, we can formally treat $\boldsymbol{B}(t)$ as independent from $\boldsymbol{J}^k$. We will not pursue the diagrammatic approach further but use this independence directly in the selfconsistent computation of the generating function $S(x)$ of the asymptotic integer moments. This is given by

$$S(x) = \lim_{t\to\infty} S_t(x) = \sum_{k=0}^{\infty}(-x)^k B_k,$$

where

$$S_t(x) \doteq \lim_{N\to\infty}\frac{1}{N}\mathrm{Tr}\overline{(I + x\boldsymbol{B}(t))^{-1}}, \tag{30}$$

$$B_k = \lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{B}^k(t)}, \tag{31}$$

Finally, from $S(x)$ we can also deduce (26)

$$C_{-1} = \frac{1}{\gamma}\lim_{x\to\infty} xS(x). \tag{32}$$

We use an expression for $S_t(x)$ based on the Gaussian ensemble of auxiliary N-dimensional vectors $\boldsymbol{y}$. This is defined by the partition function

J. Stat. Mech. (2015) P09016

$$Z_{t+1}(x) = \int \prod_i \mathrm{d}y_i \exp\left[-\frac{1}{2}\boldsymbol{y}^\top (I + x\boldsymbol{B}(t+1))\boldsymbol{y}\right]$$
$$= \int \prod_i \mathrm{d}y_i \exp\left[-\frac{1}{2}(1+x)\boldsymbol{y}^\top\boldsymbol{y} - \frac{a^2 x}{2}\boldsymbol{y}^\top \boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^\top\boldsymbol{y}\right], \tag{33}$$

from which the generating function is obtained as

$$S_{t+1}(x) = \lim_{N\to\infty} \frac{1}{N} \overline{\langle \boldsymbol{y}^\top\boldsymbol{y}\rangle}_{t+1}, \tag{34}$$

where the brackets denote expectation wrt to (33). We compute the average over random matrices $\boldsymbol{J}$, using the fact that we can neglect the dependency between the random matrices $\boldsymbol{J}$ and $\boldsymbol{B}(t)$ in the partition function (33). An annealed average of (33) and the limit $t \to \infty$ (appendix B) yields the self consistent equation

$$S(x) = \frac{1}{1+x} S(a^2 x S(x)). \tag{35}$$

The explicit computation of moments is facilitated by introducing an auxiliary function $\boldsymbol{\phi}$, its power series expansion (whose coefficients are denoted by M_k) and its inverse by

$$\phi(x) = \frac{a^2 x}{a^2 - x} S\left(\frac{x}{a^2 - x}\right) = x \sum_{k=0}^{\infty} (-1)^k x^k M_k, \tag{36}$$

$$a^2 y S(y) = \phi\left(\frac{a^2 y}{1+y}\right). \tag{37}$$

From (30), (36) and taking the limit $y \to \infty$ in (37), we obtain

$$C_{-1} = \frac{1}{\gamma a^2}\phi(a^2) = \frac{1}{\gamma}\sum_{k=0}^{\infty} (-a^2)^k M_k. \tag{38}$$

We will next see how to obtain closed form expressions for the B_k and M_k recursively. Let us first show that for known values of $B_1, \ldots, B_n$, we can compute M_n. From (35) and (36) we get the expression

$$\phi(x) = xS(\phi(x)). \tag{39}$$

Applying Lagrange's inversion formula [17] to (39) one can express the coefficients of the power series expansion of $\phi(x)$ in terms of those of S:

$$M_n = \frac{(-1)^n}{n+1}[\phi^n]\{(S(\phi))^{n+1}\} = \frac{(-1)^n}{n+1}[\phi^n]\left\{\left(\sum_{k=0}^{\infty}(-1)^k \phi^k B_k\right)^n\right\}, \tag{40}$$

where $[\phi^n]$ denotes the coefficient of ϕ^n in a power series expansion of the mathematical expression in the brackets $\{\ldots\}$. Finally, we insert in (40) the expansion of S (30). One can see that the coefficients are of the form

$$M_n = B_n + f_n(B_1, \ldots, B_{n-1}), \tag{41}$$

where the functions f_n can be computed in closed form for any n with a computer algebra programme such as Mathematica. To obtain a relation for B_n, we expand both sides of (36) into powers of y. Using elementary properties of binomial coefficients and comparing coefficients of y^n yields the second explicit relation

$$B_n = \sum_{l=0}^{n} a^{2l} \binom{n}{l} M_l = a^{2n} M_n + \sum_{l=0}^{n-1} a^{2l} \binom{n}{l} M_l. \tag{42}$$

Hence, inserting (41) into (42), we obtain

$$B_n = \frac{1}{1-a^{2n}} \left(a^{2n} f_n(B_1, \ldots, B_{n-1}) + \sum_{l=0}^{n-1} a^{2l} \binom{n}{l} M_l \right). \tag{43}$$

Unfortunately, the series (38) turns out to be an asymptotic one. Coefficients M_n diverge for $n \to \infty$ and one has to use a regularisation method such as the Borel summation or the Padè approximation in order to extract a useful result out of a finite number of coefficients. We have resorted to the latter method (appendix D). Our results obtained in this way are in excellent agreement with simulations of the kinetic Ising model for $N = 200$ and $T = 1000$. Figure 1 shows that for small values of a, i.e. small β, the matrix $\mathbf{C} \approx \mathbf{I}$. For increasing β also $\mathbf{C}_{-1}$ increases but remains finite. Note, that for $\beta \to \infty$, the parameter a converges to the value $a = \sqrt{2/\pi}$.

5. Results

In the case of the EMF estimator (8) the free energy (22) becomes:

$$F = -\operatorname{Extr}_{q,R,x,z} \left\{ \frac{q-R^2}{2x} + \alpha \sum_{\sigma_0} \int \mathcal{D}u \mathcal{D}v \frac{\mathrm{e}^{\beta\sigma_0 u}}{2\cosh(\beta u)} \left[-\frac{z^2}{2} - \frac{1}{2} \left(\sigma_0 - a \left(\sqrt{x}\, z + Ru + \sqrt{q-R^2}\, v \right) \right)^2 \right] \right\}. \tag{44}$$

Integration by part shows that $\int \mathcal{D}u\, u \tanh(\beta u) = a$, thus the above equation reduces to

$$F = \operatorname{Extr}_{R,q,x} \left\{ \frac{q-R^2}{2x} - \frac{\alpha}{1+a^2 x} (1 + a^2 q - 2a^2 R) \right\}, \tag{45}$$

and the extremum conditions yield the following equations for the order parameters:

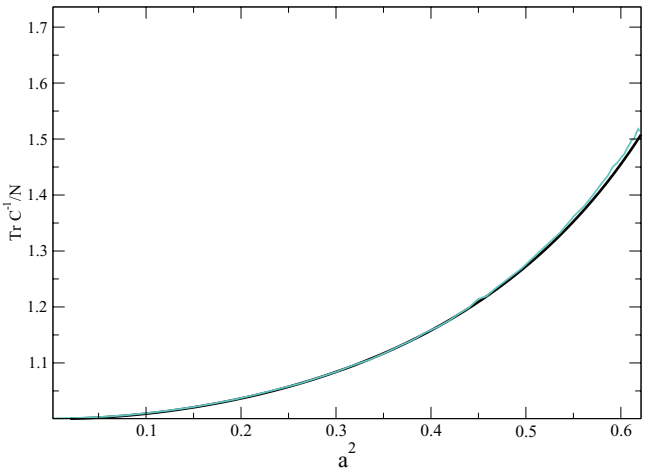

Figure 1. The analytic result(black line) for $C_{-1} = \frac{1}{N}\mathrm{Tr}\overline{C^{-1}}$ is compared with the values obtained from simulation (blue line) for $N = 200$ and $T = 1000$. Results are averaged over 50 istances of the network and error bars are negligible.

$$R = 1 \tag{46}$$

$$q = \frac{a^2(\alpha - 2) + 1}{a^2(\alpha - 1)} \tag{47}$$

$$x = \frac{1}{a^2(\alpha - 1)}. \tag{48}$$

Inserting the above equations in (25) the error is computed as follows:

$$\varepsilon_{\mathrm{EMF}} = \frac{1}{\alpha - 1}\frac{1 - a^2}{a^2}\frac{1}{N}\mathrm{Tr}\overline{C^{-1}}. \tag{49}$$

We defer a detailed analysis of the finite α performance of the ML estimator to a future publication. Here we are interested in the leading behaviour of the decay of the prediction error as $\alpha \to \infty$. It is well known that ML estimators are asymptotically efficient, i.e. the errors decay at an optimal speed. Hence, our asymptotic result should be a yardstick that allows for a comparison of algorithms. The calculation in appendix F shows that for large values of the α parameter this optimal error decays as

$$\epsilon_{\mathrm{opt}} \simeq \frac{1}{\beta a \alpha}\frac{1}{N}\mathrm{Tr}\, C^{-1}. \tag{50}$$

Hence, for $\alpha \to \infty$, we have

$$\lim_{\alpha\to\infty}\frac{\epsilon_{\mathrm{opt}}}{\epsilon_{\mathrm{EMF}}} = \frac{a}{\beta(1 - a^2)}. \tag{51}$$

For small β, i.e. large stochasticity of the spins, we have $a \simeq \beta$ and both algorithms decay at the same rate. This can still be seen in figure 2 for $\beta = 1$, where the EMF

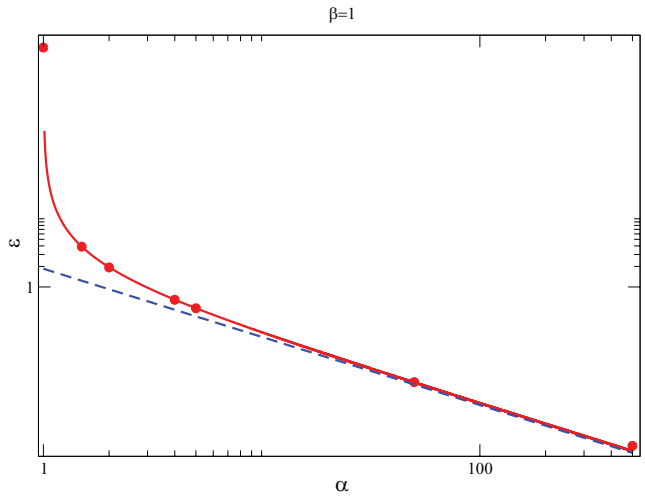

Figure 2. Mean squared error of the couplings inferred with the EMF method (red dots) for a system of size $N = 200$ with $\beta = 1$. Results are averaged over 25 istances of the network. Error bars are negligible. The red line corresponds to the replica result for the EMF prediction error, the blue line to the replica result for the asymptotic optimal prediction error.

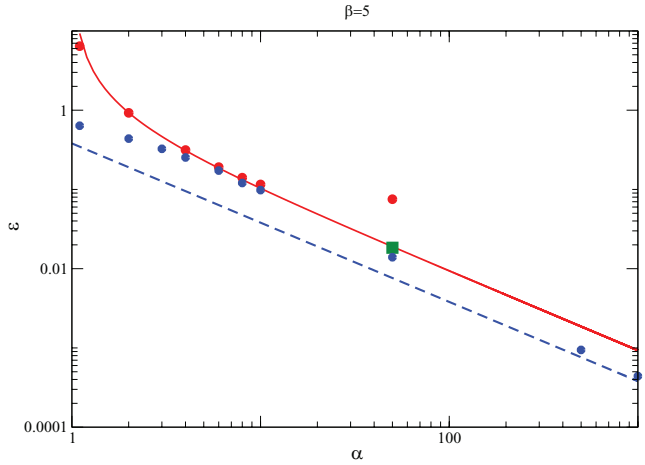

Figure 3. Mean squared error of the couplings inferred with the EMF method (red dots) for a system of size $N = 200$ with $\beta = 5$. Results are averaged over 25 instances of the network. The red line corresponds to the replica result for the EMF prediction error, the blue line to the replica result for the optimal prediction error. The blue dots are results from simulations of a penalised ML algorithm. Error bars are negligible. For large values of α, the EMF method displays finite-size effects (see the red dot at $\alpha = 50$), which are stronger for larger β. The green dot takes into accout finite-size corrections, and it is obtained as explained in figure 4.

algorithms performs close to optimal. For larger β, the spins behave more deterministically and as shown in figure 3 the EMF algorithm deviates significantly from optimality. We have also included data points from a simulation of a penalised ML estimator, where we have minimised the cost function $E_{\rm ML} + \frac{\boldsymbol{W}^{\top}\boldsymbol{W}}{2}$ numerically by a gradient

J. Stat. Mech. (2015) P09016

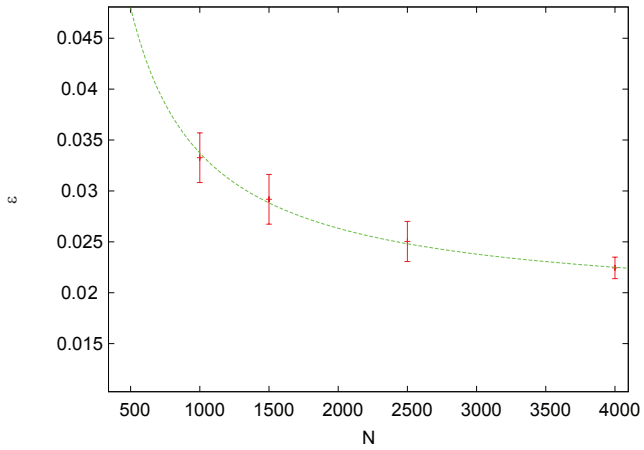

Figure 4. EMF prediction error for fixed $\alpha = 50$ and $\beta = 5$ as a function of N. Fitting a power law to the data we find the asymptotic value valid for large N, which corresponds to the green dot in figure 3.

descent algorithm. Note that the penalty term we chose is equivalent to the prior and we are thus maximizing the log-posteror. One can see that this type of algorithm achieves asymptotic optimality. Finally, with increasing β the ratio (51) decays to zero. While the decay rate of the EMF algorithm converges to a nonzero value (note that for $\beta \to \infty$, we have $a \to \sqrt{2/\pi}$), the optimal asymptotic error rate converges to zero indicating a transition to a faster decay than $1/\alpha$ in the limit. It is also interesting to note that for larger β simulations of the EMF algorithms show strong finite size effects in N and the error reaches a plateau for increasing α. Hence, we had to apply a finite scaling for the last simulation point in figure 3 .

6. Outlook

It will be interesting to develop and study algorithms which include prior knowledge about the couplings to be learnt. This could be done within a Bayesian approach where a prior probability density over couplings is specified. In this way one may e.g. introduce sparsity. Using a similar replica approach, one could compare the performance of different algorithms to that of the Bayes estimator, which is optimal on average over teacher networks drawn at random from the prior. A nontrivial question is that of an algorithmic realisation of the Bayes predictor. We expect that cavity approaches (TAP equations) could be applied to get a tractable approximation which becomes exact in the thermodynamic limit. We also expect that one should include explicit knowledge of the statistics of the spin correlations into such an approach in order to get optimal performance.

Acknowledgments

This work is supported by the Marie Curie Training Network NETADIS (FP7, grant 290038).

Appendix A. Details of the replica calculation of the free energy

After some standard manipulations [13–15], the quenched free energy (15) is computed as

$$F = -\mathrm{Extr}_{q,R,q_0}\frac{1}{\nu}\left\{ G(R,q,q_0) + \alpha\sum_{\sigma_0}\int \mathcal{D}t\mathcal{D}y \frac{\mathrm{e}^{\beta\sigma_0\left(\sqrt{1-\frac{R^2}{q}}\,t+\frac{R}{\sqrt{q}}y\right)}}{2\cosh\left[\beta\left(\sqrt{1-\frac{R^2}{q}}\,t+\frac{R}{\sqrt{q}}y\right)\right]} \right.$$
$$\left. \log\int \mathcal{D}z\mathrm{e}^{-\nu\mathcal{E}(\sigma_0,\sqrt{q_0-q}\,z+\sqrt{q}\,y)} \right\}, \tag{A.1}$$

where $G(R,\,q,\,q_0)$ is the weight of the coupling vectors $\boldsymbol{W}$ which are constrained by the order parameters:

$$G(R,q,q_0) = \lim_{n\to 0}\frac{\partial}{\partial n}\frac{1}{N}\ln Z_{\mathrm{coup}}, \tag{A.2}$$

with

$$Z_{\mathrm{coup}} = \int \mathrm{d}\boldsymbol{W}^* \prod_a \mathrm{d}\boldsymbol{W}^a \mathrm{e}^{-\frac{1}{2}\boldsymbol{W}^*\cdot\boldsymbol{W}^*} \prod_a \delta\left(\sum_{ij} W_i^a C_{ij} W_j^* - Nq_0\right)$$
$$\prod_a \delta\left(\sum_{ij} W_i^a C_{ij} W_j^a - NR\right) \prod_{a<b} \delta\left(\sum_{ij} W_i^a C_{ij} W_j^b - Nq\right). \tag{A.3}$$

We can decouple the integrals over different spins by diagonalising $C = U\Lambda U^\top$ and transforming to new variables $U^\top W^a \to W^a$, $U^\top W^* \to W^*$ which we give just the same name:

$$Z_{\mathrm{coup}} = \int \mathrm{d}\boldsymbol{W}^* \prod_a \mathrm{d}\boldsymbol{W}^a \mathrm{e}^{-\frac{1}{2}\boldsymbol{W}^*\cdot\boldsymbol{W}^*} \prod_a \delta\left(\sum_{i} W_i^a \Lambda_i W_i^* - Nq_0\right)$$
$$\prod_a \delta\left(\sum_{i} W_i^a \Lambda_i W_i^a - NR\right) \prod_{a<b} \delta\left(\sum_{i} W_i^a \Lambda_i W_i^b - Nq\right). \tag{A.4}$$

The integration over the couplings and the auxiliary parameters gives rise to the following equation for G:

$$G(R,q,q_0) = \frac{1}{2}\frac{q_0 - R^2}{q - q_0} - \frac{1}{2}\log(q-q_0) - \frac{1}{2N}\mathrm{Tr}\log C. \tag{A.5}$$

In order to compute the parameters ρ and Q from the free energy F, we introduce the auxiliary variables $\{\eta_1, \eta_2\}$ in the partition function Z_{coup} (A.4) as follows:

$$Z_{\rm coup} = \int \mathrm{d}\boldsymbol{W}^* \prod_a \mathrm{d}\boldsymbol{W}^a \; d\hat{q}_0 \; d\hat{R} \; d\hat{q} \; \mathrm{e}^{-\frac{1}{2}\boldsymbol{W}^*\cdot\boldsymbol{W}^*} \prod_a \mathrm{e}^{\mathrm{i}\hat{q}_0\left(\sum_i W_i^a \Lambda_i W_i^* - N q_0\right)}$$
$$\prod_a \mathrm{e}^{\mathrm{i}\hat{R}\left(\sum_i W_i^a(\Lambda_i+\eta_1) W_i^a - NR\right)} \prod_{a<b} \mathrm{e}^{\mathrm{i}\hat{q}\left(\sum_i W_i^a(\Lambda_i+\eta_2) W_i^b - Nq\right)}. \tag{A.6}$$

By derivatives with respect to $\{\eta_1, \eta_2\}$ and taking the limit $\eta_1 \to 0, \eta_2 \to 0$ one recovers (24).

Appendix B. Derivation of the generating function

For a Gaussian model without external field we have $\langle y_i \rangle = 0$, hence $q = \frac{1}{N}\sum_i \langle y_i \rangle^2 = 0$ and there is no need to introduce replicas, (absence of spin–glass ordering) and we can restrict ourselves to an annealed average. Decoupling the quadratic form in the exponent of (33) using correlated Gaussian random vectors with covariance $\langle \boldsymbol{z}\boldsymbol{z}^\top \rangle_c = \boldsymbol{B}(t)$, we get

$$\begin{aligned}
\overline{Z_{t+1}(x)} &= \int \prod_i \mathrm{d}y_i \exp\left[-\frac{1}{2}(1+x)\boldsymbol{y}^\top\boldsymbol{y}\right] \left\langle \exp\left(-\frac{a^2 x}{2N}(\boldsymbol{z}^\top\boldsymbol{z})(\boldsymbol{y}^\top\boldsymbol{y})\right)\right\rangle_z \\
&\propto \int_0^\infty \mathrm{d}s \; s^{\frac{N+1}{2}} \exp\left[-\frac{N}{2}(1+x)s\right] \left\langle \exp\left(-\frac{a^2 x}{2N}(\boldsymbol{z}^\top\boldsymbol{z})s\right)\right\rangle_z \\
&\propto \int_0^\infty \mathrm{d}s \; s^{\frac{N+1}{2}} \exp\left[-\frac{N}{2}(1+x)s\right] \; |I + a^2 x s \boldsymbol{B}(t)|^{-1/2} \\
&= \int_0^\infty \mathrm{d}s \; s^{\frac{N+1}{2}} \exp\left[-\frac{N}{2}(1+x)s - \frac{1}{2}\mathrm{Tr}\ln(I + a^2 x s \boldsymbol{B}(t))\right],
\end{aligned} \tag{B.1}$$

where in the second line we have introduced polar coordinates $s = \frac{1}{N}\boldsymbol{y}^\top\boldsymbol{y}$. We compute the final integral for $N \to \infty$ by Laplace's method, and use the fact that from (34) the maximiser of the integral gives $s = \frac{1}{N}\langle \boldsymbol{y}^\top\boldsymbol{y}\rangle = S_{t+1}(x)$. Finally from $-\frac{1}{2}\mathrm{Tr}\ln(I + a^2 x s\boldsymbol{B}(t)) = \mathrm{const} + \ln Z_t(a^2 x s)$ we get the recursion

$$S_{t+1}(x) = \frac{1}{1+x} S_t(a^2 x S_{t+1}(x)). \tag{B.2}$$

Taking the limit $t \to \infty$ yields (35).

Appendix C. Independence of the J and $B(t)$ matrices: an example

To better illustrate the independence of the $\boldsymbol{J}$ and $\boldsymbol{B}(t)$ matrices, let us give an example and consider the evaluation of one of the terms needed for the computation of $\lim_{N\to\infty}\frac{1}{N}\mathrm{Tr}\,\overline{\boldsymbol{B}^k(t+1)}$ (see (29)):

$$\frac{1}{N}\mathrm{Tr}(\overline{\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^{\top}\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^{\top}}). \tag{C.1}$$

The only sets of contractions giving nonzero contribution in the large N limit are the following two:

$$\begin{aligned}\frac{1}{N}\mathrm{Tr}(\overbrace{\boldsymbol{J}\overline{\boldsymbol{B}}(t)\boldsymbol{J}^{\top}}\overbrace{\boldsymbol{J}\overline{\boldsymbol{B}}(t)\boldsymbol{J}^{\top}}) &= \frac{1}{N}\mathrm{Tr}\left(\overline{\boldsymbol{B}}(t)\right)^2,\\ \frac{1}{N}\mathrm{Tr}(\overbrace{\boldsymbol{J}\boldsymbol{B}(t)\overbrace{\boldsymbol{J}^{\top}\boldsymbol{J}}\boldsymbol{B}(t)\boldsymbol{J}^{\top}}) &= \frac{1}{N}\mathrm{Tr}\overline{(\boldsymbol{B}(t)^2)}.\end{aligned} \tag{C.2}$$

The contractions involving the pairing of a $\boldsymbol{J}$ with a $\boldsymbol{B}(t)$ vanish, since they involve either $\overbrace{\boldsymbol{J}\ldots\boldsymbol{J}}$ ($\overbrace{\boldsymbol{J}^{\top}\ldots\boldsymbol{J}^{\top}}$) pairings or crossing lines (resulting in non planar diagrams), as shown in the two examples below:

$$\frac{1}{N}\mathrm{Tr}(\overbrace{\boldsymbol{J}\boldsymbol{B}(t)\overbrace{\boldsymbol{J}^{\top}\boldsymbol{J}}\boldsymbol{B}(t)\boldsymbol{J}^{\top}}) = 0, \qquad \frac{1}{N}\mathrm{Tr}(\overbrace{\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^{\top}\boldsymbol{J}\boldsymbol{B}(t)\boldsymbol{J}^{\top}}) = 0. \tag{C.3}$$

Appendix D. Padè Approximant

The so called Padè approximant [18], is a rational function (of a specified order) whose power series expansion agrees with a given power series to the highest possible order. Given a rational function of the form

$$R(x) \equiv \sum_{k=0}^{M} a_k x^k \Bigg/ \left(1 + \sum_{k=1}^{N} b_k x^k\right), \tag{D.1}$$

then R is said to be the Padè approximant to the series

$$f(x) = \sum_{k=0}^{\infty} c_k x^k \tag{D.2}$$

if the following set equations is satisfied:

$$R(0) = f(0) \tag{D.3}$$

$$\frac{\mathrm{d}^k}{\mathrm{d}\,x^k}R(x)\bigg|_{x=0} = \frac{\mathrm{d}^k}{\mathrm{d}\,x^k}f(x)\bigg|_{x=0} \qquad k = 1, \ldots, M+N, \tag{D.4}$$

which gives $M + N + 1$ equations for the unknowns $a_0, \ldots, a_M$ and $b_0, \ldots, b_N$.

Appendix E. Details on the statistics of the correlation matrix

The iterative methods explained is section 4 allows us to calculate the moments B_k and M_k, defined respectively in (31) and (36), for any given k. As an example, in the following we will enumerate the first three moments.

$$B_1 = (1-a^2)^{-1} \tag{E.1}$$

$$B_2 = (1-a^4)^{-1}(1-a^2)^{-2} \tag{E.2}$$

$$B_3 = (1+2a^4)(1-a^6)^{-1}(1-a^4)^{-1}(1-a^2)^{-3} \tag{E.3}$$

$$M_1 = (1-a^2)^{-1} \tag{E.4}$$

$$M_2 = (2-a^4)(1-a^2)^{-2}(1-a^4)^{-1} \tag{E.5}$$

$$M_3 = (5+a^4-4a^6+a^{10})(1-a^2)^{-4}(1-a^4)^{-1}(1+a^2+a^4). \tag{E.6}$$

Appendix F. Asymptotic order parameters for ML estimator

The free energy for the ML estimator is given by

$$F = -\mathrm{Extr}_{q,R,x,z}\left\{\frac{q-R^2}{2x} + \alpha\sum_{\sigma}\int \mathcal{D}t\mathcal{D}y \frac{\mathrm{e}^{\beta\sigma\left(\sqrt{1-\frac{R^2}{q}}\,t+\frac{R}{\sqrt{q}}y\right)}}{2\cosh\left[\beta\left(\sqrt{1-\frac{R^2}{q}}\,t+\frac{R}{\sqrt{q}}y\right)\right]}\right.$$
$$\left.\left[-\frac{z^2}{2}+\beta\sigma(\sqrt{x}\,z+\sqrt{q}\,y)-\log 2\cosh[\beta(\sqrt{x}\,z+\sqrt{q}\,y)]\right]\right\}. \tag{F.1}$$

It is possible to show that for $\alpha\to\infty$ one can assume that $q-R^2\to 0$, $x\to 0$ and $q\to 1$. Expanding the α dependent part of (F.1) for small $\sqrt{x}$, solving for z and finally taking the limit $q\to R^2$, we obtain

$$F \simeq -\mathrm{Extr}_{q,R,x}\left\{\frac{q-R^2}{2x}+\alpha\left(\frac{\beta a x}{2}+Rb+\int \mathcal{D}y \log 2\cosh(\beta\sqrt{q}\,y)\right)\right\}. \tag{F.2}$$

This yields the following asymptotic scaling of order parameters:

$$R\simeq 1, \quad x\simeq\frac{1}{\alpha b}, \quad q-R^2\simeq\frac{1}{\alpha b}. \tag{F.3}$$

Inserting the above expressions in the definition (25) one obtains (50).

References

[1] Mézard M and Sakellariou J 2011 Exact mean-field inference in asymmetric kinetic Ising systems *J. Stat. Mech.* L07001

[2] Roudi Y and Hertz J 2011 Mean field theory for nonequilibrium network reconstruction *Phys. Rev. Lett.* **106** 048702

J. Stat. Mech. (2015) P09016

[3] Roudi Y and Hertz J 2011 Dynamical TAP equations for non-equilibrium Ising spin glasses *J. Stat. Mech.* P03031
[4] Aurell E and Mahmoudi H 2012 Dynamic mean-field and cavity methods for diluted Ising systems *Phys. Rev.* E **85** 031119
[5] Huang H and Kabashima Y 2013 Dynamics of asymmetric kinetic Ising systems revisited arXiv:13105003
[6] Dunn B and Roudi Y 2013 Learning and inference in a nonequilibrium Ising model with hidden nodes *Phys. Rev.* E **87** 022127
[7] Tyrcha J and Hertz J 2014 Network inference with hidden units *Math. Biosci. Eng.* **11** 149
[8] Bachschmid-Romano L and Opper M 2014 Inferring hidden states in a random kinetic Ising model: replica analysis *J. Stat. Mech.* P06013
[9] Mahmoudi H and Saad D 2014 Generalized mean field approximation for parallel dynamics of the Ising model *J. Stat. Mech.* P07001
[10] Battistin C, Hertz J, Tyrcha J and Roudi Y 2015 Belief propagation and replicas for inference and learning in a kinetic Ising model with hidden spins *J. Stat. Mech.* P05021
[11] Schervish M J 1995 *Theory of Statistics* (*Springer Series in Statistics*) (New York: Springer)
[12] Eissfeller H and Opper M 1994 Mean-field Monte Carlo approach to the Sherrington-Kirkpatrick model with asymmetric couplings *Phys. Rev.* E **50** 709–20
[13] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[14] Opper M and Kinzel W 1996 Statistical mechanics of generalization *Models of Neural Networks III* ed E Domany *et al* (New York: Springer)
[15] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: an Introduction* (Oxford: Oxford University Press)
[16] Hooft G 1974 A planar diagram theory for strong interactions *Nucl. Phys.* B **72** 461–73
[17] Wilf H S 2006 *Generatingfunctionology* (Wellesley, MA: A K Peters)
[18] Press W, Teukolsky S, Vetterling W and Flannery B 2007 *Numerical Recipes* (Cambridge: Cambridge University Press)

J. Stat. Mech. (2015) P09016

ERRATA CORRIGE: Equation (22) should be replaced by

$$F = -\mathrm{Extr}_{q,R,x}\left\{\frac{q-R^2}{2x} + \alpha\sum_{\sigma_0}\int \mathcal{D}t\mathcal{D}y \frac{e^{\beta\sigma_0(\sqrt{1-\frac{R^2}{q}}t+\frac{R}{\sqrt{q}}y)}}{2\cosh[\beta(\sqrt{1-\frac{R^2}{q}}t+\frac{R}{\sqrt{q}}y)]} \right.$$
$$\left. \mathrm{Max}_z\left[-\frac{z^2}{2} + \mathcal{E}(\sigma_0, \sqrt{x}z+\sqrt{q}y)\right]\right\}.$$

4.3 Further results

4.3.1 The optimal linear estimator

We discuss the performance of another estimator, obtained by minimizing a cost function of the same form as the one of the linear mean-field estimator:

$$E^i = \frac{1}{2}\sum_{t=1}^{T}\left(\sigma_i(t) - a\sum_j J_{ij}\sigma_j(t-1)\right)^2, \tag{4.2}$$

where now a is a free parameter. This allows us to derive the optimal linear estimator.

The replica calculation explained in Paper 2 is used to find the estimation error. The order parameters of the model are found from equations (Paper 2, 43). However, for the mean field estimator, equations (Paper 2, 44-46) follow from the explicit definition of the parameter a (Paper 2, 4). If we now consider a as a free parameter, the saddle point equations become:

$$R = \frac{\int \mathcal{D}x\; x\; \tanh(\beta x)}{a}, \tag{4.3}$$

$$q = \frac{\left(\int \mathcal{D}x\; x\; \tanh(\beta x)\right)^2(\alpha-2)+1}{(\alpha-1)a^2}, \tag{4.4}$$

$$x = \frac{1}{2(\alpha-1)a^2} \tag{4.5}$$

and the error (Paper 2, 25) is

$$\varepsilon = 1 - \frac{2\int \mathcal{D}x\; x\; \tanh(\beta x)}{a} \tag{4.6}$$

$$+ \frac{(\alpha - 1 - \frac{1}{N}\mathrm{Tr}\overline{C^{-1}})\left(\int \mathcal{D}x\; x\; \tanh(\beta x)\right)^2 + \frac{1}{N}\mathrm{Tr}\overline{C^{-1}}}{(\alpha-1)a^2}. \tag{4.7}$$

The value of a which minimizes (4.7) is

$$a^{\text{opt}} = \int \mathcal{D}x\ x\ \tanh(\beta x) + \frac{\frac{1}{N}\text{Tr}\overline{C^{-1}}\left[1 - \left(\int \mathcal{D}x\ x\ \tanh(\beta x)\right)^2\right]}{(\alpha - 1)\int \mathcal{D}x\ x\ \tanh(\beta x)}, \tag{4.8}$$

with the corresponding minimal error

$$\varepsilon^{\text{opt}} = \frac{\frac{1}{N}\text{Tr}\overline{C^{-1}}\left[1 - \left(\int \mathcal{D}x\ x\ \tanh(\beta x)\right)^2\right]}{(\alpha - 1 - \frac{1}{N}\text{Tr}\overline{C^{-1}})\int \mathcal{D}x\ x\ \tanh(\beta x) + \frac{1}{N}\text{Tr}\overline{C^{-1}}}. \tag{4.9}$$

The error shows no divergence for $\alpha = 1$ (Figure 4.1) and reaches the same asymptotic value as the mean field estimator. Since a^{opt} is independent on the couplings and can be directly estimated from data, the associated linear estimator

$$J_{ij} = \frac{1}{a^{\text{opt}}} \sum_k \hat{D}_{ik}\hat{C}^{-1}_{kj} \tag{4.10}$$

only relies on one matrix inversion and it is very fast to compute. Also, with respect to the linear mean field estimator, the optimal linear estimator shows weaker finite size effects, providing a faster and better algorithm. Still, comparison with the asymptotic error of maximum likelihood shows that linear estimators are suboptimal for large values of β.

4.3.2 Bayesian inference

In a Bayesian setting, if the correct prior knowledge on the distribution of the parameters is introduced, one can design an algorithm that is asymptotically optimal: it is the Bayes optimal estimator given by the posterior expectation of the parameters. However, posterior averages require high dimensional integrals to be computed exactly. Here, we propose an analytic approximation to the posterior expectations based on cavity arguments.

Let us recall the likelihood of a spin sequence
$\boldsymbol{\sigma} = \{\sigma(0) \dots \sigma(T)\}$ for given couplings W:

$$P(\boldsymbol{\sigma}|W) = \prod_{t=1}^{T}\prod_{i=1}^{N} \frac{e^{\frac{\beta}{\sqrt{N}}\sigma_i(t)\sum_j W_{ij}\sigma_j(t-1)}}{2\cosh\frac{\beta}{\sqrt{N}}\sum_j W_{ij}\sigma_j(t-1)} P(\sigma_0), \tag{4.11}$$

where $P(\sigma_0)$ is the initial distribution of spins. As prior distribution over the couplings, we consider a univariate Gaussian distribution:

$$W_{ij} \sim \mathcal{N}(0, 1) \tag{4.12}$$

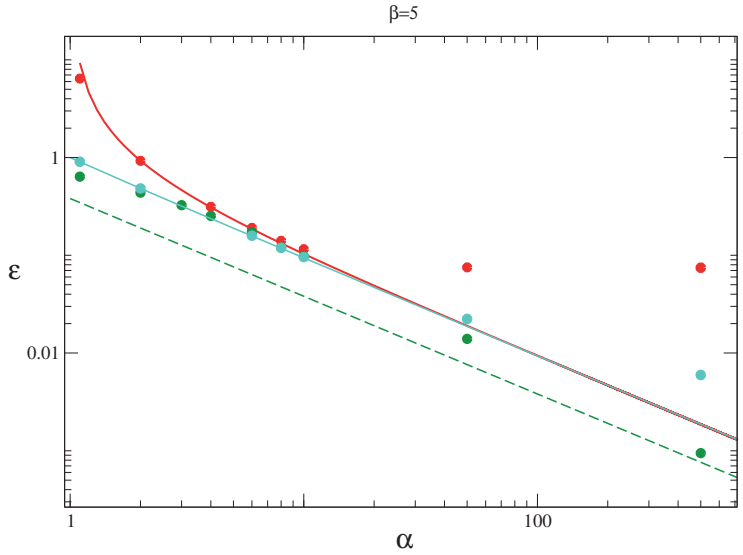

Figure 4.1: Mean squared error of the couplings inferred with different algorithms as a function of α. Light blue dots correpond to the linear optimal algorithm, red dots correspond to the mean field algorithm of Paper 2, green dots to penalized maximum likelihood. We consider a system of $N = 200$ spins with $\beta = 5$. Results are averaged over 25 istances of the network. Continuous line refer to the average error from the replica calculation, light blue for the linear optimal and red for the mean field algorithm. The green dotted line shows the asymptotic error for the maximum likelihood algorithm.

independently for any $i = 1, \ldots, N$ and $j = 1, \ldots, N$. The posterior distribution

$$P(W|\boldsymbol{\sigma}) \propto P(\boldsymbol{\sigma}|W)P(W) \tag{4.13}$$

represents the information about likely couplings, when a spin trajectory $\boldsymbol{\sigma}$ is observed. The *Bayes optimal* prediction for the couplings is given by the posterior mean $\langle W \rangle$, where the brackets denote an expectation over the posterior. Since the couplings are non-symmetric, coupling vectors $W^{(i)}$ at different neurons i are neither interacting in the likelihood (4.11) nor in the prior (4.12). Hence, inference for different coupling vectors can be done independently for each $W^{(i)}$.

We write the posterior distribution of the coupling vector $W^{(i)}$ as the product of factors:

$$p(W^{(i)}|\boldsymbol{\sigma}) = \frac{1}{p(\boldsymbol{\sigma})} \prod_{t=0}^{T} f_t(W^{(i)}), \tag{4.14}$$

where, for $t = 0$, the factor $f_0(W^{(i)})$ coincides to the prior

$$f_0(W^{(i)}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_j (W_j^{(i)})^2}, \tag{4.15}$$

while the other factors correspond to the likelihood:

$$f_t(W^{(i)}) = \frac{e^{\beta \sigma_i(t) \frac{1}{\sqrt{N}} \sum_j W_j^{(i)} \sigma_j(t-1)}}{2\cosh \frac{\beta}{\sqrt{N}} \sum_j W_j^{(i)} \sigma_j(t-1)} \qquad \text{for } t = 1...T. \tag{4.16}$$

The normaliser is given by the partition function

$$Z(\boldsymbol{\sigma}) = \int dW^{(i)} \prod_{t=0}^{T} f_t(W^{(i)}). \tag{4.17}$$

To lighten notation, in the following we will drop the superscript i of the and consider the inference problem for a singular spin vector.

4.3.3 Approximating the posterior by cavity arguments

We are interested in computing an approximation to the posterior statistics of W_j. Since the prior is Gaussian, the following exact representation for the first two moments is derived using integration by parts:

$$\langle W_j \rangle = \beta \sum_{t=1}^{T} \frac{\sigma_j(t-1)}{\sqrt{N}} \left\{ \sigma(t) - \langle \tanh(\beta h_t) \rangle \right\} \tag{4.18}$$

$$\frac{1}{N} \sum_j \left(\langle W_j^2 \rangle - \langle W_j \rangle^2 \right) = 1 - \frac{\beta^2}{N} \sum_{t=1}^{T} \left\{ \langle (h_t - \langle h_t \rangle) \tanh(\beta h_t) \rangle \right\}. \tag{4.19}$$

where the brackets denote expectation with respect to the posterior distribution (4.14) of the field h_t:

$$h_t = \frac{1}{\sqrt{N}} \sum_j W_j \sigma_j(t-1). \tag{4.20}$$

We will now derive an approximation to the distribution $p(h_t)$ of h_t using a cavity argument. We first write

$$p(h_t) \propto f_t(h_t) p^{\setminus t}(h_t), \tag{4.21}$$

where $p^{\setminus t}(h_t)$ is the 'cavity distribution' of h_t, i.e the distribution over a system where the term f_t was left out of the posterior. Using standard arguments (see

section 3.A.1) based on central limit theorem, in the large N limit, $p^{\setminus t}(h_t)$ is assumed to be a Gaussian:

$$p^{\setminus t}(h_t) = \frac{1}{\sqrt{2\pi\lambda^{\setminus t}}}\, e^{-\frac{1}{2\lambda^{\setminus t}}(h_t-\gamma^{\setminus t})^2}. \tag{4.22}$$

To close the system of equations, we express the mean $\gamma^{\setminus t}$ and the variances $\lambda^{\setminus t}$ in terms of the 'full' expectations $\langle W_j\rangle$ and $\frac{1}{N}\sum_j \left(\langle W_j^2\rangle - \langle W_j\rangle^2\right)$. This yields the following relations:

$$\lambda^{\setminus t} = \lambda = \frac{1}{N}\sum_j \left(\langle W_j^2\rangle - \langle W_j\rangle^2\right), \tag{4.23}$$

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_j \langle W_j\rangle\sigma_j(t-1) = \langle h_t\rangle &= \frac{\langle h_t f_t(h_t)\rangle_{\setminus t}}{\langle f_t(h_t)\rangle_{\setminus t}} \\ &= \gamma_{\setminus t} + \beta\lambda^{\setminus t}\left\{\sigma(t) - \langle\tanh(\beta h_t)\rangle\right\}.\end{aligned} \tag{4.24}$$

In the first equation, we have neglected correlations between couplings and assumed that total cavity variance and 'full' variance of couplings are equal in the thermodynamic limit. To derive the second equation, we used the Gaussian form of the cavity distribution and an integration by parts.

4.3.4 A simple expectation propagation algorithm

It is not a priori clear how the sets of coupled nonlinear equations (4.18), (4.19), (4.23) and (4.24) can be solved in an efficient way to get explicit predictions. We have resorted to the so–called *Expectation Propagation* (EP) algorithm, an approximate inference techniques widely used in machine learning [OW00, Min01]. We will state the algorithm first and then show that its fixed points agree with the solution of the cavity equations (4.23) and (4.24).

The algorithm is based on an auxiliary Gaussian approximation $q(W)$ to the posterior, which is used for book–keeping of the first and second order moments of the W_j and their respective cavity statistics. This pseudo–posterior $q(W)$ can be written as a product of factors:

$$q(W) = \frac{1}{\tilde{Z}}\prod_{t=0}^{T}\tilde{f}_t(W), \tag{4.25}$$

where $\tilde{Z}$ is a normalizing term, $\tilde{f}_0(W) = f_0(W)$ and for $t = 1,\ldots,T$

$$\tilde{f}_t(W) = \left(2\pi\tilde{\lambda}(t)\right)^{-N/2}\exp\left[-\frac{1}{2\tilde{\lambda}(t)}\sum_j (W_j - \tilde{\mu}_j(t))^2\right]. \tag{4.26}$$

We are approximating each factor $f_t(W)$ of the true posterior (4.14) with one Gaussian factor $\tilde{f}_t(W)$. In this version of the algorithm, which we call 'Naive EP', we are making the simplifying assumption that the covariance matrix of (4.26) is diagonal. We will later discuss the case where the full covariance matrix is considered. In order to determine the set of parameters of the approximate posterior, the factors $\tilde{f}_t(W)$ of the overall approximation are optimized sequentially. Suppose we wish to refine the factor $\tilde{f}_t(W)$. We first remove it from the current approximation of the posterior to get the unnormalized 'cavity' distribution, which is also Gaussian:

$$q^{\setminus t}(W) = \frac{q(W)}{\tilde{f}_t(W)}. \tag{4.27}$$

A new approximate posterior $q^{\text{new}}(W)$ is then computed by introducing the following distribution,

$$\frac{1}{Z_t} f_t(W) q^{\setminus t}(W), \tag{4.28}$$

which corresponds to the old $q(W)$ where one Gaussian factor $\tilde{f}_t(W)$ has been replaced by one factor $f_t(W)$ of the true posterior, and Z_t is normalizing the distribution to 1. In particular, the approximate posterior is updated by minimizing the Kullback-Leiber divergence

$$KL\left(\frac{1}{Z_t} f_t(W) q^{\setminus t}(W) \middle| q^{\text{new}}(W)\right). \tag{4.29}$$

Since the approximating distribution $q^{\text{new}}(W)$ is Gaussian, it is easy to prove [Bis06] that minimizing (4.29) is equivalent to matching the expected sufficient statistics of $q^{\text{new}}(W)$ to the corresponding moments of (4.28). Finally the revised form of the factor $\tilde{f}_t(W)$ is obtained as

$$\tilde{f}_t(W) = Z_t \frac{q^{\text{new}}(W)}{q^{\setminus t}(W)}. \tag{4.30}$$

The algorithm involves the computation of three Gaussian distributions, (4.27) (4.28) and $q(W)$ by sequentially updating their sufficient statistics. We denote the mean of the approximate posterior $q(W)$ by μ and its covariance matrix by $\Lambda = \mathbb{I}\lambda$. A summary of the algorithm is described as follows.

- Set $\tilde{f}_0$ equal to the prior
- Initialize all the factors $\tilde{f}_t$ to 1 for $t = 1...T$
- Iterate until convergence:

1. For t=1...T

   a) Update the moments of the cavity distribution (4.27):

$$\begin{aligned} \mu_j^{\setminus t} &= \frac{\tilde{\lambda}(t)\mu_j - \lambda\tilde{\mu}_j(t)}{\tilde{\lambda}(t) - \lambda} \\ \Lambda^{\setminus t} &= \mathbb{I}\lambda^{\setminus t}, \quad \lambda^{\setminus t} = \frac{\tilde{\lambda}(t)\lambda}{\tilde{\lambda}(t) - \lambda}. \end{aligned} \tag{4.31}$$

   b) Match the first and second moments $\{\mu_j, \lambda_j\}$ of the approximate posterior with the ones of the distribution (4.28). The latter moments can be computed as derivatives of the generating function:

$$Z_t(\psi) = \int dW q^{\setminus t}(W) f_t(W) e^{\sum_j W_j \psi_j}, \tag{4.32}$$

   in the limit $\psi \to 0$. For the first moment we obtain the following condition (see 4.E for details):

$$\mu_j \doteq \mu_j^{\setminus t} + \beta \frac{\lambda^{\setminus t}}{\sqrt{N}} \sigma_j(t-1) \left[\sigma(t) - \frac{\langle \tanh(\beta h^{\setminus t}) f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right], \tag{4.33}$$

   where the average is over a gaussian field with variance $\lambda^{\setminus t}$ and mean $\gamma^{\setminus t}$,

$$\gamma^{\setminus t} = \frac{1}{\sqrt{N}} \sum_j \sigma_j(t-1) \mu_j^{\setminus t}. \tag{4.34}$$

   The second moments λ_j turn out to be independent of j, due to the property $\sigma_j^2 = 1$ of Ising spins. One gets (4.E)

$$\begin{aligned} \lambda \doteq \lambda^{\setminus t} + \beta^2 \frac{(\lambda^{\setminus t})^2}{N} \Bigg\{ & 2 \frac{\langle \tanh(\beta h^{\setminus t}) [\tanh(\beta h^{\setminus t}) - \sigma(t)] f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \\ & - \left[\sigma(t) - \frac{\langle \tanh(\beta h^{\setminus t}) f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right]^2 \Bigg\}. \end{aligned} \tag{4.35}$$

   c) Evaluate and store the new factors $\tilde{f}_t(W)$ using (4.30). Its moments are:

$$\begin{aligned} \tilde{\mu}_j &= \frac{\lambda^{\setminus t}\mu_j - \lambda\mu_j^{\setminus t}}{\lambda^{\setminus t} - \lambda}, \\ \tilde{\lambda} &= \frac{\lambda^{\setminus t}\lambda}{\lambda^{\setminus t} - \lambda}. \end{aligned} \tag{4.36}$$

From (4.25) we see that, after convergence, we can compute the the moments of the posterior distribution as:

$$\mu_j = \lambda \sum_{t=1}^{T} \frac{\tilde{\mu}_j(t)}{\tilde{\lambda}(t)}, \tag{4.37}$$

$$\lambda = \left(1 + \sum_{t=1}^{T} \frac{1}{\tilde{\lambda}(t)}\right)^{-1}. \tag{4.38}$$

In 4.F we show that those fixed point equations are equivalent to the expected moments (4.23) and (4.24) of the couplings obtained from cavity arguments.

4.3.5 Average case: a replica analysis

The average prediction error for the Bayes optimal estimator can be computed in a student-teacher setting with a replica analysis, analogously to the analysis of Paper 2. We now work in a Bayesian framework, where the student has prior knowledge about the teacher. The distribution of the student couplings is given by the posterior distribution (4.14) and the partition function $Z(\boldsymbol{\sigma})$ is the normalizer (4.17) of the posterior distribution:

$$Z(\boldsymbol{\sigma}) = \int d\boldsymbol{W}\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_j (W_j)^2} \prod_t \frac{e^{\beta\sigma_i(t)\frac{1}{\sqrt{N}}\sum_j W_j\sigma_j(t-1)}}{2\cosh\frac{\beta}{\sqrt{N}}\sum_j W_j\sigma_j(t-1)} \tag{4.39}$$

Since (4.14) also represents the posterior distribution corresponding to a prior distribution of random teachers, $Z(\boldsymbol{\sigma})$ will be proportional to the total probability $P(\boldsymbol{\sigma})$:

$$P(\boldsymbol{\sigma}) = \frac{Z(\boldsymbol{\sigma})}{\mathcal{C}}, \tag{4.40}$$

where $\mathcal{C} = \sum_{\boldsymbol{\sigma}} Z(\boldsymbol{\sigma})$ is the normalization factor. Hence, the teacher and the student network enter the calculation in a completely symmetric way: the average student-teacher overlap equals the average student self-overlap, and the error is

$$\varepsilon = \frac{1}{N}(\boldsymbol{W}^* - \langle \boldsymbol{W}\rangle)^2 = 1 - \langle \boldsymbol{W}^a \cdot \boldsymbol{W}^b\rangle, \tag{4.41}$$

where $\langle \dots \rangle$ denotes averaging with respect to the distribution of couplings. The error can be computed from the free energy using the replica trick as follows:

$$F = -N^{-1}\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log Z(\boldsymbol{\sigma}) = -\lim_{n\to 1}\frac{\partial}{\partial n} N^{-1} \log \sum_{\boldsymbol{\sigma}} Z^n(\boldsymbol{\sigma}). \tag{4.42}$$

In order to compute the average over the spin trajectories, we consider the same approximation described in the paper, that we argue to be correct in the limit $N \to \infty$. For the central spin σ_0, the variables $\sigma_0(t)$ are binary at all times t. The other spins $\{\sigma_j(t-1)\}_{j\neq 0}$ enter the partition function through the fields $\sum_{j\neq 0} W_j^a \sigma_j(t-1)$. Since the system is weakly coupled, the central limit theorem tells us that such fields are Gaussian distributed and we treat the spins $\{\sigma_j(t-1)\}_{j\neq 0}$ themselves as Gaussian random variables: $p(\{\sigma_j(t)\}_{j\neq 0}) = \mathcal{N}(0, \boldsymbol{C})$. The stationary covariance matrix $\boldsymbol{C}$ takes into account equal time spatial correlations among the spins $\{\sigma_j(t-1)\}$ for $j \neq 0$, while dependences at different time steps are neglected. The partition function is:

$$\sum_{\boldsymbol{\sigma(t)}} Z^n(\boldsymbol{\sigma}) = \int \prod_{a=1}^{n} d\boldsymbol{W}^a \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_a \boldsymbol{W}^a.\boldsymbol{W}^a}$$
$$\left\{ \sum_{\sigma_0} \prod_{j\neq 0} d\sigma_j \frac{1}{\sqrt{|2\pi C|}} e^{-\frac{1}{2}\sum_{i,j\neq 0}\sigma_i C_{ij}^{-1}\sigma_j} \prod_{a=1}^{n} \frac{e^{\beta\sigma_0 \frac{1}{\sqrt{N}}\sum_{j\neq 0} W_j^a \sigma_j}}{2\cosh \frac{\beta}{\sqrt{N}}\sum_{j\neq 0} W_j^a \sigma_j} \right\}^T . \tag{4.43}$$

We have remapped the kinetic Ising model in a number T of logistic regression models, whose inputs are not independent but correlated through the matrix C. The stationary value of the matrix C depends on the (teacher) couplings and encodes for non-trivial equal time correlations among spins. We computed its statistics in Paper 2. Averages over the spins σ_j can be expressed through Gaussian fields

$$h_a \doteq \frac{1}{\sqrt{N}} \sum_j W_j^a \sigma_j \tag{4.44}$$

whose covariances in the limit $N \to \infty$ will become self averaging order parameters. Under the assumption of replica symmetry, correct for convex cost functions, we have:

$$\left\langle h_a^2 \right\rangle = \frac{1}{N} \sum_{ij} W_i^a C_{ij} W_j^a = 1, \tag{4.45}$$

$$\langle h_a h_b \rangle = \frac{1}{N} \sum_{ij} W_i^a C_{ij} W_j^b \doteq q \qquad a \neq b. \tag{4.46}$$

The first equality follows from the fact that the matrix C represents the correlation between all the spins but σ_0; hence, it is independent of W_{0j}. The

calculation, detailed in 4.H, yields:

$$F = -\mathrm{Extr}_{q,\hat{q}} \frac{1}{2} \left\{ \hat{q}(q-1) - \frac{1}{N}\mathrm{Tr}(1-\hat{q}C) \right. \\ \left. + \alpha \sum_{\sigma} \int \mathcal{D}y\, A(\sigma, y, q) \log A(\sigma, y, q) \right\}, \tag{4.47}$$

where

$$A(\sigma, y, q) = \int \mathcal{D}x\, \frac{e^{\beta\sigma(x\sqrt{1-q}+y\sqrt{q})}}{2\cosh[\beta(x\sqrt{1-q}+y\sqrt{q})]}. \tag{4.48}$$

Here, $\mathcal{D}x = (dx/\sqrt{2\pi})e^{-x^2/2}$, $\mathcal{D}y = (dy/\sqrt{2\pi})e^{-y^2/2}$ and the parameter $\alpha = T/N$ represents the rescaled length of the trajectories. The saddle point equations for the order parameters, extremising the free energy (4.47) are:

$$\begin{aligned} q &= 1 + \frac{1}{\hat{q}} \left(1 - \frac{1}{N}\mathrm{Tr}(1-(\hat{q}C)^{-1})^{-1} \right), \\ \hat{q} &= -\alpha \sum_{\sigma} \int \mathcal{D}y\, B(\sigma, y, q)/A(\sigma, y, q), \end{aligned} \tag{4.49}$$

where

$$B(\sigma, y, q) = \left[\beta \int \mathcal{D}x \left(\sigma - \tanh(x\sqrt{1-q} + y\sqrt{q}) \right) \frac{e^{\beta\sigma(x\sqrt{1-q}+y\sqrt{q})}}{2\cosh[\beta(x\sqrt{1-q}+y\sqrt{q})]} \right]^2.$$

In order to compute the error (4.41) we need the typical overlap between two student networks, that is different from the parameter q (4.46). It can be computed from (4.109), by noticing that

$$\frac{1}{\hat{q}} \sum_i \frac{\partial}{\partial \Lambda_i} \frac{1}{N} \ln Z_{\mathrm{c}}^n = -\frac{1}{2N} \sum_{a \neq b} \sum_i \langle \boldsymbol{W}_i^a \cdot \boldsymbol{W}_i^b \rangle \tag{4.50}$$

From (4.41) and (4.50) one gets the final result for the mean square error:

$$\varepsilon = 1 - \frac{2}{\hat{q}} \sum_i \frac{\partial}{\partial \Lambda_i} F_0(q, \hat{q}) = \frac{1}{N}\mathrm{Tr}(I - \hat{q}C)^{-1}. \tag{4.51}$$

4.3.6 Results

We evaluate the analytic expression of the error (4.51) from the system of equations (4.49) and from the statistics of the C matrix (Paper 2). Figure (4.2) compares the results with the mean square error of the couplings inferred by

using the Naive Expectation Propagation algorithm of section 4.3.4. The data are generated from a kinetic Ising model with independent Gaussian couplings with variance $1/N$. The Naive Expectation Propagation algorithm, where the true posterior distribution is approximated by a Gaussian distribution with diagonal covariance matrix, is in good agreement with the theoretical predictions of the Bayes estimator. For large values of α, the error of Expectation Propagation deviates from the replica result due to finite size effects. Figure (4.3) shows that it converges to the replica result for large N. In particular, we fix $\alpha = 500$ and show how the error decay as a function of N. Fitting a shifted power low to the data we obtain an asymptotic value $\varepsilon_{N\to\infty} = 0.0007 \pm 0.0004$, which is in good agreement with the replica value $\varepsilon = 0.000746$. For small values of α, Expectation Propagation outperforms all other algorithms. For completeness, in Appendix 4.G we design a 'Complete' Expectation Propagation algorithm, where the posterior distribution is approximated by a Gaussian with full covariance matrix. We tested it for values of α up to 10: the error is not significantly lower than the one of Naive EP while the required time for convergence is much higher. Regarding computational complexity, each iteration of both the Expectation Propagation and Maximum Likelihood algorithms to estimate one coupling vector $\{W_j\}$ requires a computation of the order TN. Expectation Propagation, though, converges in much fewer steps and shorter time. For instance, for a system of $N = 100$ spins at $\alpha = 10$, we needed approximately 284 updates of the learning rates for maximum likelihood (25 seconds) and 5 iterations over time for EP (2 seconds).

4.4 Conclusions

In this chapter we considered a kinetic Ising model where the couplings are independent Gaussian random variables with variance scaling as $1/N$, and computed the error of three different estimators for the couplings, working in a a student-teacher scenario. We analysed a linear mean field estimator, which can be rephrased as the minimizer of a local quadratic cost function; an estimator based on an anlogous quadratic cost function, which contains a free parameter that is optimized to minimize the estimation error; the optimal Bayes estimator, where a prior distribution is introduced and the couplings are estimated as their posterior averages.

The replica calculation revealed the importance of equal-time correlations between spins at different sites: despite being of the order $1/\sqrt{N}$ [MS11], they significantly affect the estimation error, especially when the stochasticity of the spin dynamics is decreased. By computing an exact result for the statistics of the random correlation matrix, we find an explicit expression for the learning

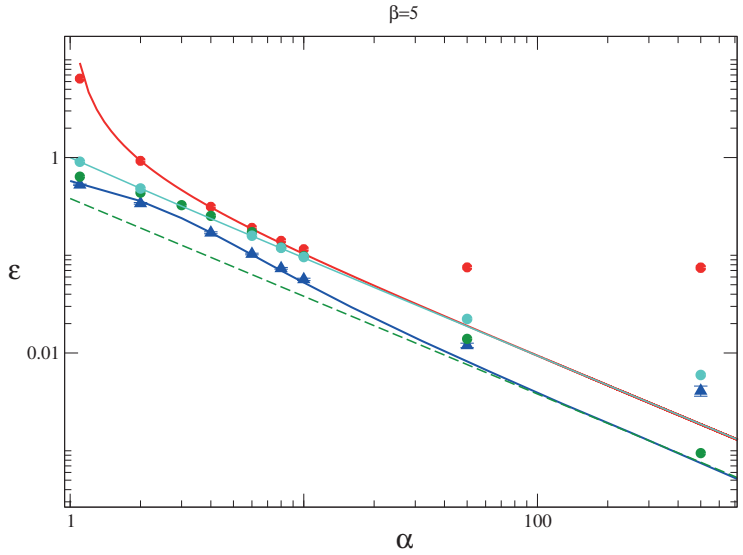

Figure 4.2: Mean squared error of the couplings inferred with different algorithms as a function of α. Red dots correspond to the mean field algorithm of Paper, green dots to maximum likelihood, blue dots to Naive EP and light blue dots to linear optimal. We consider a system of $N = 200$ spins with $\beta = 5$. Results are averaged over 25 istances of the network. Continuous line refer to the average error from the replica calculation, red for the mean field algorithm (see Paper) and blue for the Bayes estimator. The green dotted line shows the asymptotic error for the maximum likelihood algorithm.

curve of the three algorithms, which agrees very well with simulations.

By comparison with the asymptotic error of the maximum likelihood estimator, which has the property of asymptotic optimality, we assessed the performance of the considered methods. The error of linear estimators, such as the linear mean-field one, is asymptotically close to optimal one for weak couplings, whereas it deviates from optimality for stronger couplings.

If the prior corresponds to the true distribution of the parameters, the Bayes optimal estimator provides an asymptotically optimal estimator; the intractable integrals required to compute posterior averages can be approximated using the cavity method of statistical physics, and we solved the resulting set of equations by an algorithm of the Expectation Propagation type: the true posterior distribution of the couplings is approximated by a Gaussian distribution, whose mean and covariance are updated iteratively, in such a way that the approximated distribution is as close as possible to the true one (in the sense of KL-divergence). The fixed point equations for the moments of

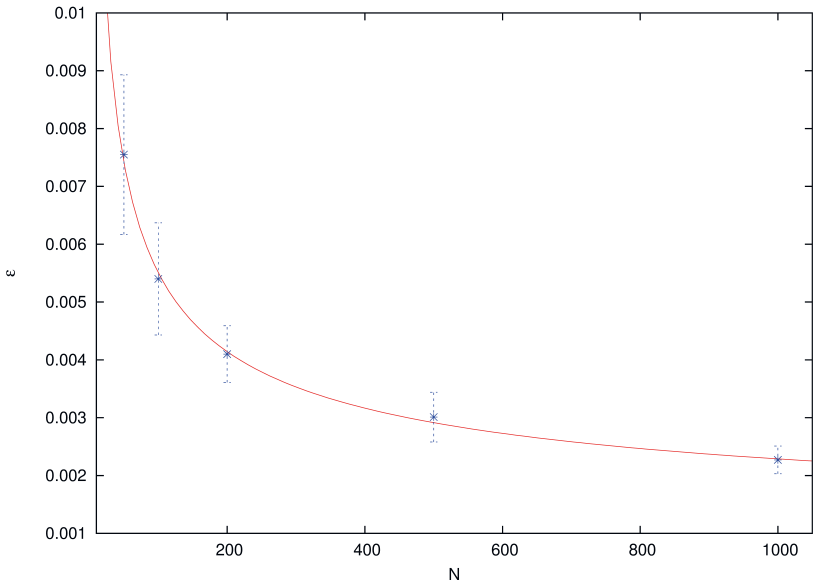

Figure 4.3: Mean squared error of the couplings inferred with the Naive EP algorithm, plotted as a function of N for fixed $\alpha = 500$ and $\beta = 5$. We fit a rescaled power law to the data to find an asymptotic value $\varepsilon_{N\to\infty} = 0.0007 \pm 0.0004$ and a decay exponent 0.48 ± 0.007.

the posterior distribution are equivalent to the ones obtained from cavity arguments. An interesting question, that we leave to future research, is whether our Bayesian estimator implemented via the Expectation Propagation algorithm becomes exact in the limit $N \to \infty$ (i.e., whether our approximation to the true posterior averages become exact in the thermodynamic limit).

Moreover, as a future direction, it would be interesting to extend our results to other types of networks. Particularly relevant for practical applications are sparse networks. Prior knowledge on the couplings could be introduced via a spike and slab distribution, widely used in machine learning for sparse linear models (see, e.g., [MB88, GM93, BBB$^+$03, BBB$^+$03]). Each weight J_{ij} of the prior would be set to zero with probability $1 - \pi$ and drawn from a Gaussian distribution with probability π . This would allow us to use the Expectation Propagation algorithm developed in this section to infer the couplings in sparse networks.

For what concerns the analytical analysis, the basic ideas underlying our replica formalism can be applied to other systems. The idea of treating one central spin as the output of a perceptron whose inputs are correlated through a cavity matrix C, will inspire the analysis presented in the next chapter.

Appendix

4.A The replica method: from spin glasses to neural networks

The replica trick is a long-established method in the analysis of disordered systems; dating back at least to Hardy [HLP34] as an identity for computing the average of a logarithm (see also the work of Kac [Kac68]), it was reintroduced by Edwards [Edw70] for a model of rubber elasticity and became well known with its application to spin glasses [EA75, SK75]. After the seminal work of Gardner [Gar87, Gar88], it has been widely used to study learning in neural networks in a statistical mechanics framework [WRB93, OK96, EVdB01].

In the following sections, I will first introduce the replica method in the context of spin glasses; then, I will show how this statistical mechanics formalism is applied to the problem of learning in neural networks.

4.A.1 Spin glasses and the replica trick

Spin glasses are the simplest models for glassy systems [Par06]. They have been widely studied in the last 40 years not only to derive some of the main properties of glassy systems, but also because they provided a framework to study properties of other physical systems, as fragile glasses, colloids and granular materials; moreover, and many ideas developed in the field were later applied to combinatorial optimization problems and learning in neural networks.

The Hamiltonian of a spin glass with pairwise interactions is:

$$H(\sigma) = -\sum_{i,j=1,\dots,N} J_{ij}\sigma_i\sigma_j - \sum_{i=1,\dots,N} h_i^{ex}\sigma_i \tag{4.52}$$

where σ_i are Ising variables (i.e., $\sigma_i = \pm 1$) located on a lattice vertices, the couplings Js are random variables located on the edges of the lattice and h_i^{ex} are local external fields.

Many models of spin glasses have been studied, according to the distribution of the couplings and the topology of the lattice. We will consider the

Sherrington-Kirkpatrick (SK) model c, introduced in 1975 as an exactly solvable model of a spin glass; all couplings are random variables with a Gaussian or bimodal distribution with variance $1/N$, and the network is fully connected.

The disorder induced by the randomness of the couplings is assumed to be quenched, which means that the changes in the Js happen on a time scale infinitely larger than the typical time scale of spin fluctuations. If the system observables depended on J, it would follow that the physical properties of spin-glasses are different for each different realization of the quanched disorder. In contrast, it turns out [Cav09] that extensive quantities, such as the free energy, have the property of self-averageness: in the thermodynamic limit (infinite volume limit) they assume the same value for each realization of the couplings. This means that analytically we can average over J, and the obtained result is in agreement with the physical value of the observable.

Let us now focus on the SK model, whose Hamiltonian is

$$H(\sigma) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i^{ex}\sigma_i \tag{4.53}$$

where the couplings J_{ij} are independent Gaussian random variables with zero mean and variance J_0/N, and h_i^{ex} is the external field. Denoting by f_J and Z_J the free energy and the partition function of a sample with a set J of couplings:

$$f_J = -\frac{1}{\beta N}\log Z_J = -\frac{1}{\beta N}\log \mathrm{Tr}_{\{\sigma\}} e^{-\beta H(\sigma)}, \tag{4.54}$$

where β is the inverse temperature, we are interested in computing the average value over the disordered distribution of the free energy,

$$f = \int dJ\, P(J) f_J = -\frac{1}{\beta N}\overline{\log Z_J}, \tag{4.55}$$

where the overbar $\overline{(...)}$ denotes the average over the disorder distribution, which is called *quenched average*. The computation of such average is technically difficult, since it requires averaging the logarithm of the partition function. A much easier approach would be to consider the *annealed average* of the free energy, that is the logarithm of the average of the partition function. However, while F is an extensive quantity, this is not the case for Z (which is exponential in the system size), and therefore Z is not in general self-averaging. To overcome the difficulty, Edwards and Anderson proposed to apply the replica trick, which makes use of the relation $log(x) = \lim_{n\to 0} \frac{x^n-1}{n}$ to transform the logarithm in a power law:

$$-\beta N f = \lim_{n\to 0} \frac{\overline{Z^n} - 1}{n}. \tag{4.56}$$

One first assumes that n is integer, and can see Z^n as the partition function of n replicas of the same system, that share the same realization of the couplings but are non-interacting. Then one must perform the limit $n \to 0$, and later $N \to \infty$, to get the self-averaging value of the free energy (in practice, it turns out that reverting the order of the two limits is not a source of trouble). Without going into the details of the calculation (an example will be given in Paper 4) we point out that the integral in

$$\overline{Z^n} = \int dJ\, P(J) \mathrm{Tr}_{\{\sigma^a\}} \exp\left[\beta \sum_{i<j} J_{ij} \sum_a \sigma_i^a \sigma_j^a + \beta \sum_i h_i^{ex} \sum_a \sigma_i^a\right], \quad (4.57)$$

where a is the replica index, can be easily performed; since the Js are coupled to a quadratic term in the spins, one can use the inverse Gaussian integral (Hubbard-Stratonivich transformation) to uncouple the spins σ_i^a in the sites and sum over all possible spin configurations. This procedure naturally introduces the spin overlap

$$q^{ab} = \frac{1}{N} \sum_i \sigma_i^a \sigma_i^b, \quad a < b, \quad (4.58)$$

which represents the typical overlap between two configurations in a given state belonging to two different replicas.

Sherrington and Kirkpatrick [SK75] considered a *replica symmetric* (RS) ansatz for the order parameter: the overlap is the same no matter what two replicas are chosen,

$$q^{ab} = q(1 - \delta^{ab}). \quad (4.59)$$

However, their result turned out to have some unphysical feature, such as a negative entropy at low temperatures. It was first thought to be a problem related to exchanging the $n \to 0$ limit with the large volume limit $N \to \infty$ when computing the free energy, but it later became clear [DAT78, BM80] that the problem resides in the symmetry of the replica ansatz.

The replica symmetry breaking (RSB) calculation was presented in a series of papers by Parisi [Par80a, Par80b, Par80b]; its solution was physically consistent and confirmed both by numerical simulations and by other analytical methods. Yet, it took over 20 years for the reluts predicted by the RSB calculation to be rigorously proven by Talagrand [Tal06], using the interpolation method of Guerra [Gue03].

Our work will focus on systems for which the replica symmetric ansatz is correct. Hence, we refer the reader to the literature mentioned above and to [MPV87] for a description of the replica symmetry breaking method; here, we just mention a few key concepts.

The replica symmetry breaking (RSB) procedure can be described as an iterative process that sequentially reparameterizes the n x n matrix with elements q^{ab}. The starting point (0-th step) is the ansatz (4.59), where the matrix has zero entries on the diagonal and values $q^{ab} = q_0$ for all non diagonal values. At the 1-st step, the n x n matrix is diveded in n/m_1 blocks of size m_1x m_1: the off-diagonal terms of the diagonal blocks take value q_1, the other terms remain unchanged. The correct solution was found by considering an infinite number of steps of RSB. The analysis of the probability distribution of the order parameters yielded a geometrical characterization of the space of solutions, that turned out to be an ultrametric space, where $q^{ac} \leq \min(q^{ab}q^{bc})$ [MPS$^+$84, MV85]. This is a signature of the complex free–energy landscape of the spin glass phase, where an infinitely large number of minima are separated by barriers that grow indefinitely as the system size increases [Par83]. The number of minima (metastable states) in exponentially large in the size of the system N [BM80], and so is the time spent by the system in every single valley: in the large N limit ergodicity is broken [MY82].

4.A.2 Statistical mechanics of learning: general setup

In this section, we introduce the statistical mechanics framework to analyze the theoretical performance of learning algorithms in the so called teacher-student scenario. For further reading, see [WRB93, OK96, EVdB01].

Let us begin by defining a neural network as a set of nodes, or neurons, that can take values ± 1 and influence each other's state through directed connections W_{ij}. Among the various architectures that have been studied, we will focus on layered networks; we refer to the the first layer as the input, and to the last layer as the output. For simplicity, we will further assume that the network has only one layer of N input nodes with values $\boldsymbol{S} = \{S_i\}$ and one node in the in the output layer, whose state is σ. The state of the neuron σ is set to a function of the weighted sum of the inputs, where the weights are the connections W_j (here $\boldsymbol{W} = \{W_i\}$ is an N dimensional vector):

$$\sigma(\boldsymbol{W}; \boldsymbol{S}) = g\left(\sum_j W_j S_j\right), \tag{4.60}$$

where g is a generic non-linear function. The term learning refers to the process of setting the weights to the values that make the network perform a desired task, that is a target input-output mapping which will be denoted as the *rule*. We will focus on *supervised* learning, where the weights are adjusted as to approximate as closely as possible a target function $\sigma_0(\boldsymbol{S})$. This is achieved by providing the network with a training set, that is a set of M input/output

pairs $\{\boldsymbol{S}^{(k)}, \sigma_0^{(k)}\}_{k=1}^{M}$ generated by some unknown mapping, and by requiring that the network adapts its weight to map each pair well[2]. We assume that the inputs are are generated independently at random from the input space according to some probability distribution $P(\boldsymbol{S})$. The target mapping can be represented by another network with weights $\boldsymbol{W}^*$, or *teacher* network, that knows the correct mapping and generated the examples. The learning network $\boldsymbol{W}$ is called the *student* and the prescription $\{\boldsymbol{S}^{(k)}, \sigma_0^{(k)}\} \to \boldsymbol{W}$ that, given the training set, specifies the student coupling vector is referred to as the *learning rule*. In particular, a rule for which a network in the student space exists that realizes the target function $\sigma_0(\boldsymbol{S})$ is called learnable. Otherwise the rule is called unlearnable.

In order to measure the deviation of the network output $\sigma(\boldsymbol{W}; \boldsymbol{S})$ from the target output $\sigma_0(\boldsymbol{S})$, we introduce an error function $\mathcal{E}(\mathbf{W}; \mathbf{S})$ which is zero if teacher and student agree on the output to $\mathbf{S}$ and larger than zero otherwise.

Based on the error function, one can define an extensive energy, which scales with the number of examples; if such energy is defined not to depend explicitly on the unknown rule, it can be used in a learning algorithm. A widely used choice is the training energy

$$E(\mathbf{W}) = \sum_{k=1}^{M} \mathcal{E}(\mathbf{W}; \mathbf{S}^{(k)}), \tag{4.61}$$

and training is usually achieved by minimizing such training energy, for example via gradient descent.

After the student network has learned a rule from a limited set of examples, it can make predictions on novel inputs. The ability of a network to generalize from a limited number of examples to the whole space of inputs is measured by the generalization function:

$$\varepsilon(\mathbf{W}) = \int d\boldsymbol{S}\, P(\boldsymbol{S}) \mathcal{E}(\mathbf{W}; \mathbf{S}). \tag{4.62}$$

Let us introduce one learning scenario that is particularly well suited for a theoretical analysis and that we will consider in chapters 4 and 5: the case of Gibbs learning at non-zero temperatures. In this case training is achieved by minimizing a generic training energy of the form (4.61), according to a stochastic dynamics governed by the Langevin relaxation equation

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_W E(\mathbf{W}) + \boldsymbol{\eta}(t), \tag{4.63}$$

[2]We will *not* consider the case where the data available for training are corrupted with noise.

where η is a white noise with variance

$$\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t'). \tag{4.64}$$

At zero temperature the noise drops out leaving us with a simple gradient descent equation. In learning algorithms, the noise can be useful in escaping local minima of the energy; the temperature is slowly decreased so that the system settles near to the global energy minimum at $T \approx 0$. The dynamics (4.63) generates at long times the Gibbs probability distribution on the parameter space for a canonical ensemble of networks

$$\rho(\mathbf{W}) = \frac{1}{Z}\exp[-\nu E(\mathbf{W})] \tag{4.65}$$

where $\nu = 1/T$ quantifies the noise of the training procedure, and the normalization integral

$$Z = \int d\mathbf{W}\exp[-\beta E(\mathbf{W})] \tag{4.66}$$

measures the weighted accessible volume in the configuration space. In the limit $\nu \to \infty$ the system settles at the global energy minimum [3]. The typical behaviour of a network can be now computed via thermal averages, denoted by $\langle \dots \rangle$, with respect to the distribution (4.65). Note however that the above quantities still depend on the random choice of a specific training set $\{\mathbf{S}^{(k)}\}_{k=1}^{M}$. Moreover, we do not want to consider a specific realization of the teacher network, but we assume that the teacher network is drawn at random from a teacher rule space. Both the teacher network and the data sets independently generated from it are randomly chosen and kept fixed during the learning procedure, and they represent - in the language of the statistical physics of spin glasses -a quenched disorder. It turns out however that the error (4.62) is self-averaging in the limit of $N \to \infty$, which means that almost any realization of the teacher network and training set will give the same result. We will denote

[3] This distribution, arising naturally for stochastic algorithms, was also introduced by Levin Tishby Solla [LTS90] from a statistical estimation theory perspective, where the training process in feedforward neural networks is seen as a parameter estimation problem. The solution can be found by setting the parameters to the value that maximizes the likelihood of the training set of M independent examples. Imposing that the maximization of the likelihood be equivalent to the minimization of an additive error of the form (4.61) for every set of independent training examples, the authors arrive at the Gibbs canonical distribution on the ensemble of all networks with the same parameter space (i.e., networks with the given architecture). The distribution depends on a free positive parameter, which determines the level of acceptable training error as well as the tevel of stochasticity in the training algorithm, and can be interpreted as an inverse temperature.

quenched averages by overlines :

$$\overline{(\dots)} = \int \left(\prod_k^M d\mathbf{S}^k \right) d\mathbf{W}^* \prod_k^M p(\mathbf{S}^k|\mathbf{W}^*) p(\mathbf{W}^*)(\dots). \tag{4.67}$$

The average generalization error $\overline{\langle \varepsilon(\mathbf{W}) \rangle}$ will then depend on the noise parameter ν in the thermal average and on the number of examples M in the quenched average, which we will assume to be proportional to the number of degree of freedom (i.e. independent synaptic weights): $M = \alpha N$, with α finite. Given a distribution of the inputs and an energy function, one can use the tools of statistical mechanics to calculate the quenched averages and derive the average generalization error from derivatives of the free energies, in the thermodynamic limit of $N \to \infty$. The calculation of the quenched average of the free energy per coupling,

$$F = -N^{-1} \nu^{-1} \overline{\log Z}, \tag{4.68}$$

can be carried out using the replica method. The quantity

$$\lim_{n \to 0} \frac{1}{n} \ln \overline{Z^n}$$

has to be evaluated for integer n and then analytically continued to $n = 0$. The replicated partition function yields

$$\overline{Z^n} = \int \left(\prod_{a=1}^n d\mathbf{W}^a \right) e^{-N\alpha \mathcal{G}_r[\{\mathbf{W}^a\}]} \tag{4.69}$$

where the replicated Hamiltonian is

$$\mathcal{G}_r[\{\mathbf{W}^a\}] = -\ln \int d\mathbf{S}\, d\mathbf{W}^* p(\mathbf{S}|\mathbf{W}^*) p(\mathbf{W}^*) \exp[-\beta \sum_{a=1}^n \mathcal{E}(\mathbf{W}^\mathbf{a}; \mathbf{S})]. \tag{4.70}$$

The average generalization error can then be computed as follows:

$$\begin{aligned} \overline{\langle \varepsilon(\mathbf{W}) \rangle} &= \lim_{n \to 0} \overline{Z^{n-1} \int d\mathbf{W}\, \varepsilon(\mathbf{W}) \exp[-\beta E(\mathbf{W})]} \\ &= \lim_{n \to 0} \int \left(\prod_{a=1}^n d\mathbf{W}^a \right) \varepsilon(\mathbf{W}^1)\, e^{-N\alpha \mathcal{G}_r[\{\mathbf{W}^a\}]}. \end{aligned} \tag{4.71}$$

The integration over the inputs will couple the weights of different replicas of the system, which makes it natural to introduce order parameters - representing the overlap of the weights of two copies of the student networks and the

overlap between the teacher and the student network - that will convey the dependence of $\mathcal{G}_r$ on the weights. The values of these order parameters are the ones that extremize $\mathcal{G}_r$; computing the saddle point equations for the parameters requires making an ansatz about the symmetry of the parameters at the saddle point. That simplest ansatz is the replica symmetric ansatz, whose validity can be assessed by studying the local stability of the replica symmetric saddle point. In this thesis, we will consider convex cost functions of the weight vector, which ensures the replica symmetric ansatz to be correct [EVdB01].

4.B Maximum likelihood estimator

Let us consider the Markovian dynamics for the Ising model that we introduced in Paper 1, which is described by the transition probability

$$p(\boldsymbol{\sigma}(t+1)|\boldsymbol{\sigma}(t)) = \prod_i^N \frac{\exp[\beta\sigma_i(t+1)h_i(t)]}{2\cosh\beta h_i(t)}, \tag{4.72}$$

where we defined the field $h_i(t) = \sum_j J_{ij}\sigma_j(t) + H_i^{\text{ext}}(t)$. The log-likelihood of the system parameters is

$$\mathcal{L}(\boldsymbol{J}, \boldsymbol{H}^{\text{ext}}) = \frac{1}{T}\sum_t\sum_i[\beta\sigma_i(t+1)h_i(t) - \log 2\cosh\beta h_i(t)]. \tag{4.73}$$

To find the maximum likelihood parameters, one starts from an initial sets of couplings and external fields, and then adjust them iteratively by gradient ascent; the derivatives are given by

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial H_i^{\text{ext}}(t)} &= \langle\sigma_i(t)\rangle_r - \langle\tanh h_i(t)\rangle_r,\\ \frac{\partial\mathcal{L}}{\partial J_{ij}} &= \frac{1}{T}\sum_t\langle\sigma_i(t)\sigma_j(t)\rangle_r - \langle\tanh h_i(t)\sigma_j(t)\rangle_r,\end{aligned} \tag{4.74}$$

where we assumed that N_r realizations of the trajectories can be observed, and the brackets $\langle\ldots\rangle_r$ represent empirical averages over different realizations. The derivatives (4.74) can be evaluated in N^2TN_r computational steps, which makes the computation much faster than the Likelihood of the equilibrium Ising model, where the normalizer of the Boltzmann distribution scales exponentially with the system size.

4.C Mean field estimators for the stationary state

Let us summarize the derivation of the mean field relation between the coupling matrix and the correlation matrix found in [RH11b] for a kinetic Ising model with parallel dynamics.

We start from the definition of one-step-delayed and equal time correlation matrices for the spin fluctuation $\delta\sigma_i(t) = \sigma_i(t) - m_i(t)$, that can be computed from data:

$$D_{ij} = \langle \delta\sigma_i(t+1)\delta\sigma_j(t) \rangle \tag{4.75}$$
$$C_{ij} = \langle \delta\sigma_i(t)\delta\sigma_j(t) \rangle \tag{4.76}$$

where $\langle ... \rangle$ are empirical averages; in the stationary case, averaging over time and repeats would be equivalent, so in this paragraph for any function of time $f(t)$ observed over a trajectory of length T, we define

$$\langle f(t) \rangle = \frac{1}{T} \sum_t f(t).$$

By setting the gradient of the likelihood (4.74) to zero, one gets

$$\langle \sigma_i(t+1)\sigma_j(t) \rangle = \langle \tanh[h_i(t)]\sigma_j(t) \rangle. \tag{4.77}$$

We now expand the effective local field $h_i(t)$ around its mean field solution. Formally, we write $s_i = m_i + \delta s_i$, and use the naive mean field equation $m_i = \tanh[\sum_j J_{ij} m_j + H_i]$ for the magnetization. From (4.77), expanding $\tanh[h_i(t)]$ in powers of δs_i we get to the leading order

$$\langle \delta\sigma_i(t+1)\delta\sigma_j(t) \rangle = (1 - m_j^2) \sum_k J_{ik}^{\mathrm{nMF}} \langle \delta\sigma_k(t)\delta\sigma_j(t) \rangle, \tag{4.78}$$

which can be written as

$$\boldsymbol{J}^{\mathrm{nMF}} = \boldsymbol{A}^{\mathrm{nMF}} \boldsymbol{D} \boldsymbol{C}^{-1}, \tag{4.79}$$

where

$$A_{ij}^{nMF} = \delta_{ij}(1 - m_i^2). \tag{4.80}$$

The TAP inversion formula is derived analogously [RH11b] and also results in the linear relation

$$\boldsymbol{J}^{\mathrm{TAP}} = \boldsymbol{A}^{\mathrm{TAP}} \boldsymbol{D} \boldsymbol{C}^{-1}, \tag{4.81}$$

where

$$J_{ij}^{TAP} = A_{ij}^{nMF}(1 - F_i), \tag{4.82}$$

and F_i is the smallest root of the following cubic equation:

$$F_i(1 - F_i)^2 = (1 - m_i^2) \sum_j (J^{nMF})_{ij}^2 (1 - m_j^2). \tag{4.83}$$

4.D Mean field estimators for the transient dynamics

In the case of out-of-equilibrium dynamics we refer to the three mean field theories of section 3.B; analogous relations to (4.79, 4.82) can be found for the the one-time-delayed and equal time correlation matrices, defined as

$$\begin{aligned} D_{ij}(t) &= \langle \delta s_i(t+1) \delta s_j(t) \rangle \\ C_{ij}(t) &= \langle \delta s_i(t) \delta s_j(t) \rangle, \end{aligned} \tag{4.84}$$

where now the matrices depend on time. Starting from the dynamical mean field equation of (3.38), by using the same expansion as in (4.C), one finds he following relation:

$$\boldsymbol{D}(t) = \boldsymbol{A}(t)\boldsymbol{J}(t)\boldsymbol{C}(t) \tag{4.85}$$

where $A_{ij} = \delta_{ij} a_i$ is a diagonal matrix with elements

$$a_i(t) = \left(1 - m_i^2(t)\right) \left[1 - \left(1 - m_i^2(t)\right) \sum_j J_{ij}^2 \left(1 - m_j^2(t-1)\right) \right]. \tag{4.86}$$

For the mean field theory (3.41), which is exact for asymmetric networks, the key observation is that when couplings scale as $1/\sqrt{N}$, also each matrix element will be of the order $1/\sqrt{N}$. We define the field fluctuation $\delta g_i(t) = \sum_j J_{ij} \delta s_j(t-1)$, and note that the joint distribution of $\delta g_i(t)$ and $\delta g_j(t)$ has small covariance $\epsilon = \langle \delta g_i(t) \delta g_j(t) \rangle$. By an expansion in small ϵ, one retrieve the relation (4.88), where now $A_{ij} = \delta_{ij} a_i$,

$$a_i(t) = \int \mathcal{D}x \left[1 - \tanh^2 \left(g_i(t) + H_i(t) + x\sqrt{\Delta_i(t)} \right) \right]. \tag{4.87}$$

In [MS14], the authors derive recursive equations that allow to compute correlations between spins at different times, starting from cavity arguments. The

equal time and one-time-delayed correlation matrices are related through the following relation:

$$(1-\delta)\boldsymbol{D}(t) = \boldsymbol{A}(t)\boldsymbol{J}(t)\boldsymbol{C}(t), \tag{4.88}$$

where δ is the unit matrix, so that $(1-\delta)\boldsymbol{D}(t)$ contains only non-diagonal terms of the covariance matrix $\boldsymbol{D}$, i.e. D_{ij} $i \neq j$; its diagonal elements of the form $D_{ii}(t)$ have to be computed separately according to (3.45). Please note that in the present chapter we are using the notation $D_{ij}(t) = \langle \delta s_i(t+1)\delta s_j(t)\rangle$, to draw a parallel between techniques used for the stationary case and the ones valid for the transient dynamics; in section 3.B we were using $C_{ij}(t+1,t) = \langle \delta s_i(t+1)\delta s_j(t)\rangle$. The elements of the diagonal matrix $\boldsymbol{A}$ in (4.88) are

$$a_i(t) = \int \mathcal{D}x \left[1 - \tanh^2\left(g_i(t) + H_i(t) + x\sqrt{V_{ii}(t,t)}\right)\right], \tag{4.89}$$

where the definitions of $g_i(t)$ and $V_{ii}(t,t)$ are respectively (3.39) and (3.43).

4.E Expectation Propagation algorithm: generating function of the moments

The moment generating function for the distribution (4.28) is

$$\begin{aligned} Z_t(\psi) &= \int dW q^{\setminus t}(W) f_t(W) e^{\sum_j W_j\psi_j} \\ &= \left(2\pi\lambda^{\setminus t}\right)^{-N/2} \int dW e^{-\frac{1}{2\lambda^{\setminus t}}\sum_j (W_j - \mu_j^{\setminus t})^2} \\ &\quad \frac{e^{\beta\sigma_i(t)\frac{1}{\sqrt{N}}\sum_j W_j\sigma_j(t-1)}}{2\cosh\frac{\beta}{\sqrt{N}}\sum_j W_j\sigma_j(t-1)} e^{\sum_j W_j\psi_j}. \end{aligned} \tag{4.90}$$

Enforcing the definition of h_t by delta function and introducing the integral representation of the delta, one obtains:

$$\begin{aligned} Z_t(\psi) = \left(2\pi\lambda^{\setminus t}\right)^{-N/2} &\int dW\, dh\, d\hat{h}\, e^{-\frac{1}{2\lambda^{\setminus t}}\sum_j (W_j - \mu_j^{\setminus t})^2} \frac{e^{\beta\sigma_i(t)h}}{2\cosh(\beta h)} \\ &\exp\left\{ i\hat{h}\left[h - \frac{1}{\sqrt{N}}\sum_j W_j^{(i)}\sigma_j(t-1)\right] + \sum_j W_j\psi_j \right\}. \end{aligned} \tag{4.91}$$

The integration over dW yields:

$$Z_t(\psi) = \int \mathcal{D}\phi\, dh\, d\hat{h}\, \frac{e^{\beta\sigma_i(t)h}}{2\cosh(\beta h)}\, \exp\left\{ i\hat{h}\left[h - \frac{\lambda^{\setminus t}}{\sqrt{N}} \sum_j \sigma_j(t-1)\psi_j \right.\right. \\ \left.\left. -\frac{1}{\sqrt{N}} \sum_j \sigma_j(t-1)\mu_j^{\setminus t} - \phi \right] + \frac{\lambda^{\setminus t}}{2} \sum_j \psi_j^2 + \sum_j \psi_j \mu_j^{\setminus t} \right\}, \tag{4.92}$$

where $\mathcal{D}x = (dx/\sqrt{2\pi\lambda^{\setminus t}})e^{-\phi^2/2\lambda^{\setminus t}}$ is the probability density for a Gaussian variables with zero mean and variance $\lambda^{\setminus t}$. One recovers (4.33) from $\frac{\partial \log Z(\psi)}{\partial \psi_j}$ and (4.35) from $\frac{\partial^2 \log Z(\psi)}{\partial \psi_j^2}$ in the limit $\psi 0$.

4.F Fixed point of the Expectation Propagation algorithm

From (4.33), (4.36) and (4.37) we find

$$\mu_j \left(\frac{1}{\lambda} - \sum_{t=1}^{T} \frac{1}{\tilde{\lambda}} \right) = \sum_{t=1}^{T} \frac{\beta}{\sqrt{N}} \sigma_j(t-1) \left[\sigma(t) - \frac{\langle \tanh(\beta h^{\setminus t}) f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right]. \tag{4.93}$$

Using (4.38), this yields equation (4.18), while (4.24) is recovered from (4.33) and (4.104). From (4.36) and (4.38) we observe that

$$\lambda^{\setminus t} = \left(1 + \sum_{\tau \neq t} \frac{1}{\tilde{\lambda}(\tau)} \right)^{-1} \approx \lambda \quad \text{for large } t, \tag{4.94}$$

which is equivalent to (4.23). Hence, from (4.35), we get

$$\begin{aligned} \frac{1}{\tilde{\lambda}(t)} =& \frac{\tilde{\lambda}^2(t)}{\tilde{\lambda}(t)\lambda} \frac{\beta^2}{N} \left\{ 2 \frac{\langle \tanh(\beta h^{\setminus t})[\tanh(\beta h^{\setminus t}) - \sigma(t)] f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right. \\ & \left. + \left[\sigma(t) - \frac{\langle \tanh(\beta h^{\setminus t}) f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right]^2 \right\} \\ \approx & \frac{\beta^2}{N} \left\{ 2 \frac{\langle \tanh(\beta h^{\setminus t})[\tanh(\beta h^{\setminus t}) - \sigma(t)] f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right. \\ & \left. + \left[\sigma(t) - \frac{\langle \tanh(\beta h^{\setminus t}) f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right]^2 \right\}, \end{aligned} \tag{4.95}$$

where in the last equality we used (4.94). Inserting the above equation in (4.38), the following expression for the posterior variance is found:

$$\lambda \approx \left\{ 1 + \frac{\beta^2}{N} \sum_{t=1}^{T} \left[2 \frac{\langle \tanh(\beta h^{\setminus t})[\sigma(t) - \tanh(\beta h^{\setminus t})] f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} + \left(\sigma(t) - \frac{\langle \tanh(\beta h^{\setminus t}) f_t(h_t) \rangle_{\setminus t}}{\langle f_t(h_t) \rangle_{\setminus t}} \right)^2 \right] \right\}^{-1}. \tag{4.96}$$

This equals the combination of equation (4.23) and (4.19), when the latter is expressed through the cavity fields.

4.G Complete Expectation Propagation

We now approximate the posterior (4.14) by a gaussian distribution of the form

$$q(\boldsymbol{W}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}) \tag{4.97}$$

where C is the full $N \times N$ covariance matrix. We assume that $q(\boldsymbol{W})$ can be also written as a product of factors,

$$q(\boldsymbol{W}) = \frac{1}{\tilde{Z}} \prod_{t=0}^{T} \tilde{f}_t(\boldsymbol{W}), \tag{4.98}$$

where $\tilde{f}_0(\boldsymbol{W}) = f_0(\boldsymbol{W})$ and for $t = 1, \ldots, T$

$$\tilde{f}_t(\boldsymbol{W}) = \left(2\pi |\tilde{\boldsymbol{C}}^{\boldsymbol{t}}| \right)^{-N/2} \exp \left[-\frac{1}{2} \sum_{ij} (W_i - \tilde{\mu}_i^t)(\tilde{C}^t)_{ij}^{-1} (W_j - \tilde{\mu}_j^t) \right]. \tag{4.99}$$

As before, we introduce the 'cavity' distributions:

$$q^{\setminus t}(\boldsymbol{W}) = \frac{q(\boldsymbol{W})}{\tilde{f}_t(\boldsymbol{W})} = \mathcal{N}(\boldsymbol{\mu}^{\setminus t}, \boldsymbol{C}^{\setminus t}), \tag{4.100}$$

where

$$\boldsymbol{C}^{\setminus t} = (\boldsymbol{C}^{-1} - (\tilde{\boldsymbol{C}}^t)^{-1})^{-1}, \quad \boldsymbol{\mu}^{\setminus t} = \boldsymbol{C}^{\setminus t} \cdot (\boldsymbol{C}^{-1} \cdot \boldsymbol{\mu} - (\tilde{\boldsymbol{C}}^t)^{-1} \cdot \tilde{\boldsymbol{\mu}}^t). \tag{4.101}$$

We then update the approximate posterior by matching its first and second moments with the ones of the distribution

$$\frac{1}{Z_t} f_t(W) q^{\setminus t}(W).$$

Those moments can be calculated by derivatives of the generating functional

$$\begin{aligned} Z_t(\boldsymbol{\psi}) &= \int dW q^{\setminus t}(W) f_t(W) e^{\sum_j W_j \psi_j} \\ &= \left(2\pi |\boldsymbol{C}^{\setminus t}|\right)^{-N/2} \int dW e^{-\frac{1}{2}\sum_{ij}(W_i - \mu_i^{\setminus t})(C^{\setminus t})^{-1}_{ij}(W_j - \mu_j^{\setminus t})} \\ &\quad \frac{e^{\beta\sigma_i(t)\frac{1}{\sqrt{N}}\sum_j W_j \sigma_j(t-1)}}{2\cosh\frac{\beta}{\sqrt{N}}\sum_j W_j\sigma_j(t-1)} e^{\sum_j W_j\psi_j} \\ &= \left(2\pi|\boldsymbol{C}^{\setminus t}|\right)^{-N/2} \int \mathcal{D}\phi\, dg\, d\hat{g} \frac{e^{\beta\sigma_i(t)g}}{2\cosh\beta g} \\ &\quad e^{-i\hat{g}\left[g - \frac{1}{\sqrt{N}}\boldsymbol{\mu}^{\setminus t}\cdot\boldsymbol{\sigma}(t-1) - \frac{1}{\sqrt{N}}\boldsymbol{\psi}\cdot\boldsymbol{C}^{\setminus t}\cdot\boldsymbol{\sigma}(t-1) - \phi\right]} e^{\frac{1}{2}\boldsymbol{\psi}\cdot\boldsymbol{C}^{\setminus t}\boldsymbol{\psi} + \boldsymbol{\mu}^{\setminus t}\cdot\boldsymbol{C}^{\setminus t}} \end{aligned} \tag{4.102}$$

in the limit $\boldsymbol{\psi} \to 0$, where $\mathcal{D}\phi = \left(\frac{N}{2\pi\,\boldsymbol{\sigma}(t-1)\cdot\boldsymbol{C}^{\setminus t}\cdot\boldsymbol{\sigma}(t-1)}\right)^{N/2} e^{-\frac{\phi^2}{2}\frac{N}{\boldsymbol{\sigma}(t-1)\cdot\boldsymbol{C}^{\setminus t}\cdot\boldsymbol{\sigma}(t-1)}}$. The calculation of the first moment yields:

$$\mu_j = \mu_j^{\setminus t} + \frac{\beta}{\sqrt{N}}\sum_k C_{jk}^{\setminus t}\sigma_k(t-1)\left[\sigma(t) - \frac{\langle\tanh(\beta h^{\setminus t}) f_t(h_t)\rangle_{\setminus t}}{\langle f_t(h_t)\rangle_{\setminus t}}\right], \tag{4.103}$$

where the average is a over the gaussian field with variance

$$(\boldsymbol{\sigma}(t-1)\cdot\boldsymbol{C}^{\setminus t}\cdot\boldsymbol{\sigma}(t-1))/N$$

and mean

$$\gamma^{\setminus t} = \frac{1}{\sqrt{N}}\sum_j \sigma_j(t-1)\mu_j^{\setminus t}. \tag{4.104}$$

For the second moments we get

$$C_{ij} = C_{ij}^{\setminus t} + \frac{\beta^2}{N}\sum_{kl} C_{ik}^{\setminus t} C_{lj}^{\setminus t}\sigma_k(t-1)\sigma_l(t-1)\left[\frac{\langle\tanh^2(\beta h^{\setminus t}) f_t(h_t)\rangle_{\setminus t}}{\langle f_t(h_t)\rangle_{\setminus t}} - 1\right]. \tag{4.105}$$

As last step of the iteration, one evaluates and store the new factors $\tilde{f}_t(W)$ (4.99), whose first two moments satisfy the following equations:

$$\tilde{\boldsymbol{C}}^t = (\boldsymbol{C}^{-1} - (\boldsymbol{C}^{\setminus t})^{-1})^{-1}, \quad \tilde{\boldsymbol{\mu}}^t = \tilde{\boldsymbol{C}}^t\cdot(\boldsymbol{C}^{-1}\cdot\boldsymbol{\mu} - (\boldsymbol{C}^{\setminus t})^{-1}\cdot\boldsymbol{\mu}^{\setminus t}). \tag{4.106}$$

4.H Details of the replica calculation

It is convenient to split the computation of the free energy into two parts. The first one represents the weight of the coupling vectors W which are constrained

by the order parameters:

$$F_0 = -\lim_{n\to 1}\frac{\partial}{\partial n} N^{-1}\ln Z_{\rm c}^n, \tag{4.107}$$

with

$$\begin{aligned} Z_{\rm c}^n &= \int\prod_a d\boldsymbol{W^a} e^{-\frac{1}{2}\sum_a \boldsymbol{W^a}\cdot\boldsymbol{W^a}} \prod_{a<b}\int dq\;\delta\left(\sum_{ij} W_i^a C_{ij} W_j^b - Nq\right) \\ &= \int\prod_a d\boldsymbol{W^a} e^{-\frac{1}{2}\sum_a \boldsymbol{W^a}\cdot\boldsymbol{W^a}} \prod_{a<b}\int dq\frac{d\hat{q}}{2\pi} e^{-\frac{\hat{q}}{2}\left\{\sum_{ij} W_i^a C_{ij} W_j^b - Nq\right\}}. \end{aligned} \tag{4.108}$$

Note, there is no need to introduce extra conditions on diagonal overlaps as they are taken care of by the prior. As we did in the Paper, we can decouple the integrals over different spins by diagonalising $C = U\Lambda U^\top$ and transforming to new variables $U^\top W^a \to W^a$ which we give just the same name. Hence

$$Z_{\rm c}^n = \int\prod_a d\boldsymbol{W^a} e^{-\frac{1}{2}\sum_a\sum_i (W_i^a)^2} \prod_{a<b}\int dq\frac{d\hat{q}}{2\pi} e^{-\frac{\hat{q}}{2}\left\{\sum_i W_i^a \Lambda_i W_i^b - Nq\right\}}. \tag{4.109}$$

Once the sites are decoupled we consider the limit $N\to\infty$ and evaluate the integral with the saddle point method. We get the following result, where we write $Z_{\rm c}^n$ as a function of the parameters $q, \hat{q}$; the physical value of the free energy will be then obtained by extremization over $q, \hat{q}$:

$$\begin{aligned} &\frac{1}{N}\ln Z_{\rm c}^n(q,\hat{q}) = \frac{1}{N}\sum_i \ln\int\prod_a dW^a e^{-\frac{1}{2}\sum_a (W^a)^2} e^{-\frac{\hat{q}\Lambda_i}{2}\sum_{a\neq b}(W^a W^b - q)} \\ &= \frac{\hat{q}qn(n-1)}{2}\frac{1}{N}\sum_i \Lambda_i + \frac{1}{N}\sum_i \ln\int Dz\left(\frac{1}{\sqrt{2\pi(1-\hat{q}\Lambda_i)}} e^{-\frac{1}{2}\frac{\hat{q}\Lambda_i z^2}{1-\hat{q}\Lambda_i}}\right)^n \\ &= \frac{\hat{q}qn(n-1)}{2}\frac{1}{N}{\rm Tr}(C) - \frac{n-1}{2N}{\rm Tr}(\ln(I-\hat{q}C)) - \frac{1}{2N}{\rm Tr}(\ln[I+(n-1)\hat{q}C]). \end{aligned}$$

Using $\frac{1}{N}{\rm Tr}C = 1$ and (4.107) one finds:

$$F_0(q,\hat{q}) = -\frac{1}{2}\hat{q}(q-1) + \frac{1}{2N}{\rm Tr}(1-\hat{q}C) \tag{4.110}$$

The second term of the free energy involves averages over the Gaussian random fields (4.44). Its computation follows the steps of standard calculations for the perceptron learning problem [EVdB01, OK96, NY96] and the result is written in (4.47).

5 Learning Curves for the inverse Ising problem

5.1 Introduction

Having discussed the average performance of inference algorithms for the kinetic Ising model in the previous chapter, we will now examine the corresponding problem for the equilibrium case. We aim to compute the average error of learning the couplings from independent data generated by the equilibrium Ising model. As exact inference via the maximum likelihood method is computationally intractable for large systems, a vast amount of literature has been devoted to design approximate inference algorithms (for a review, see [NZB17b]).

As for numerical approaches, Monte Carlo methods can be exploited in a maximum likelihood algorithm [FS88, BDT$^+$07] or be used to directly sample from the posterior probability distribution of the parameters [Fer16].

Mean field analytical approximations to likelihood maximization have been derived by information geometric approaches [Tan00], cavity methods [OW01b], weak-coupling expansions [Ple82]; a related technique is based on a perturbative expansion of the entropy functional in terms of connected correlations [SM09]. These approaches are exact in the thermodynamic limit for densely and weakly interacting systems, but constitute a poor approximation when the couplings are strong.

An approximation that is valid also for networks with strong couplings was developed by [CM11, CM12]. It consists in constructing and selecting specific subsets of variables of increasing size, called clusters. The algorithm retains the clusters of variables contributing most to the cross-entropy and rejects the small contributions. It works well when the networks have many short loops.

In the opposite limit, i.e. for networks with no loops, the Bethe-Pearls ansatz [Pei36] of pair-wise factorised form for the spin distribution is exact; it can be used within a variational approximation to reconstruct the couplings; it is exact on trees but can be also used on networks that are locally tree-like [Bet35]. A related method is a message passing algorithm called susceptibility propagation, which combines belief propagation [Pea14] and linear response

theory [NB12a]. Its performance is investigated in [NB12a, MVK10].

Based on very different approaches, two consistent estimators for the couplings are given by minimum probability flow [SDBD09] and pseudolikelihood maximisation. The latter method, inspired by logistic regression, was developed in the statistics community [Bes74] and recently became very popular within the physics community [AE12, Bes74, ELL$^+$13, MDP14]. The log-likelihood function to be maximised is replaced by a tractable sum of local log-likelihood functions (i.e., distributions of single random variables conditioned on the others). With respect to exact maximum likelihood, the computational complexity is reduced from exponential to polynomial in the system size (and in the sample size). Moreover, it outperforms most of the methods cited above at low temperatures [AE12, NZB17b].

In this chapter, we provide a setting for a theoretical comparison of the performance of some of these algorithms. We compute the average error of learning the couplings in the teacher-student scenario, where the teacher network will be kept fixed during the calculation. As we saw in the last chapter, the computation requires performing 'thermal' averages over student couplings and quenched averages over the spin configurations. Such configurations are distributed according to the Boltzmann measure, and the intractability of the partition function constitutes the main technical difficulty. A first approach to this problem was published in [KKC98], a work that analyses the performance of various online algorithms for learning the parameters in a spin glass from data about its metastable states. The authors, inspired by the work of Palmer and Pond [PP79], considered an approximation for the distribution of local fields that factors in the sites. They showed that it is possible to learn the Hamiltonian from a small set ($\mathcal{O}(N)$) of metastable states; however, the reconstruction error does not match well with the one from simulations due to the crude approximation on the field distribution.

In Paper 3, using ideas from the cavity methods of statistical physics [MPV87], we develop a formalism that takes into account the correlations between local fields; in this way, we are able predict the error with great accuracy for the first time. An alternative approach, based on a replica calculation, can be found in [Ber16]. Our method allows us to simply study the performance of algorithms based on the minimisation of a local cost function, such as maximum pseudolikelihood and a mean-field approximation to maximum likelihood.[1]

We also derive the form of the optimal local cost function that achieves minimal error. The formalism to address this problem dates to the study of one-layer perceptrons, when Kinouchi and Caticha (1992) presented a modi-

[1] Another recently proposed algorithm based on convex optimization is interaction screening [VMLC16]; its performance has been analysed in [Ber16].

fied version of the Hebb algorithm for the one-layer perceptron that minimises the generalisation error. We follow the more recent approach of [AG16] (introduced for the problem of optimal regression and followed in [Ber16] for the inverse Ising problem) and perform a functional minimisation of the error with respect to the cost function.

Our results will depend on parameters related to the statistics of the network that generated the data; in the last part of the chapter, we will show how such parameters can be estimated from the data.

The explicit equations for maximum likelihood and maximum pseudo-likeliood are given in Appendix 5.A.

5.2 Paper 3.

Author's contribution: I performed the analytical and numerical calculations, prepared the figures and contributed to writing the paper.

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

A statistical physics approach to learning curves for the inverse Ising problem

Ludovica Bachschmid-Romano and Manfred Opper

Department of Artificial Intelligence, Technische Universität Berlin, Marchstraße 23, Berlin 10587, Germany
E-mail: ludovica.bachschmidromano@tu-berlin.de and manfred.opper@tu-berlin.de

Received 24 January 2017
Accepted for publication 6 May 2017
Published 23 June 2017

Online at stacks.iop.org/JSTAT/2017/063406
https://doi.org/10.1088/1742-5468/aa727d

Abstract. Using methods of statistical physics, we analyse the error of learning couplings in large Ising models from independent data (the inverse Ising problem). We concentrate on learning based on local cost functions, such as the pseudo-likelihood method for which the couplings are inferred independently for each spin. Assuming that the data are generated from a true Ising model, we compute the reconstruction error of the couplings using a combination of the replica method with the cavity approach for densely connected systems. We show that an explicit estimator based on a quadratic cost function achieves minimal reconstruction error, but requires the length of the true coupling vector as prior knowledge. A simple mean field estimator of the couplings which does not need such knowledge is asymptotically optimal, i.e. when the number of observations is much larger than the number of spins. Comparison of the theory with numerical simulations shows excellent agreement for data generated from two models with random couplings in the high temperature region: a model with independent couplings (Sherrington–Kirkpatrick model), and a model where the matrix of couplings has a Wishart distribution.

Keywords: analysis of algorithms, learning theory, statistical inference

1742-5468/17/063406+28$33.00

Contents

1. Introduction 2
2. Estimators for the inverse Ising model 4
3. Local learning 4
4. Teacher–student scenario and statistical physics analysis 5
5. Cavity approach I: quenched averages 7
6. Replica result 8
7. Quadratic cost functions 9
8. The optimal local cost function 11
9. Cavity approach II: TAP equations and approximate mean field ML estimator 12
10. Reconstruction error for MF-ML estimator 14
11. Asymptotics 15
12. Numerical results 17
13. Discussion and outlook 18

Acknowledgments 21

Appendix A. Details of the replica calculation 21

Appendix B. Saddle point equations for the order parameters 22

Appendix C. Relation between order parameters 23

Appendix D. Replica result for quadratic cost functions 24

Appendix E. Asymptotics from the replica approach 25

Appendix F. Asymptotic error for pseudo-likelihood estimator 25

Appendix G. Error dependence on the system size 26

References 27

1. Introduction

In recent years, there has been an increasing interest in applying classical Ising models to data modelling. Applications range from modelling the dependencies of spikes recorded from ensembles of neurons [1, 2] to protein structure determination [3] or gene expression analysis [4]. An important issue for such applications is the so-called

https://doi.org/10.1088/1742-5468/aa727d

J. Stat. Mech. (2017) 063406

inverse Ising problem, i.e. the statistical problem of fitting model parameters, external fields and couplings, to a set of data. Unfortunately, the exact computation of statistically efficient estimators such as the *maximum likelihood* estimator is computationally intractable for large systems. Hence, to overcome this problem researchers have suggested two possible solutions: the first one tries to approximate maximum likelihood estimators by computationally efficient procedures such as Monte Carlo sampling [5] or mean field types of analytical computations, see e.g. [6–9]. A second line of research abandons the idea of maximising the likelihood function and replaces it by other cost functions which are easier to optimise. The most prominent example is the so-called pseudo-likelihood method [10–14]. In general it is not clear which of the two methods leads to better reconstruction of an Ising model. The quality of such estimators, e.g. measured by the mean squared reconstruction error of network parameters, will depend on the problem at hand.

As an alternative to analysing specific instances of problems, one may study the typical prediction performance of algorithms assuming that the *true* Ising parameters are drawn at random from a given ensemble distribution. For such random problem cases, one can apply powerful methods of statistical physics to compute (scaled) reconstruction errors *exactly* in the limit where the number of spins grows to infinity and the number of data is increased proportionally to the number of spins. Such an approach has been applied extensively to statistical learning in large neural networks in the past [15–17] and also to learning in an Ising spin glass with binary teacher couplings [18], where learning is performed in an online fashon. In a previous paper [19] we have applied this method to the learning from dynamical data which are modelled by a kinetic Ising model with random independent couplings. This problem is theoretically simpler compared to the static, 'equilibrium' Ising case discussed in the present paper. This is because the spin statistics of the dynamical model is fairly simple in the 'thermodynamic' limit of a large network and gives rise to Gaussian distributed fields.

We will show in the following that a related approach is possible to data drawn independently from an equilibrium Ising model when we assume that couplings are learnt independently for each spin using local cost functions. Although the spin statistics is more complicated, computations are possible, when the so-called 'cavity' method [20] is applicable to the true teacher Ising model.

The paper is organised as follows: section 2 explains the inverse Ising problem and maximum likelihood estimation. Section 3 introduces simpler estimators which are derived from local cost functions. In section 4, we review the statistical physics approach for analysing learning performances within the so-called teacher student scenario. In section 5 we explain the cavity method for performing quenched averages over spin configurations. Section 6 presents explicit results of our method applied to the inverse Ising model with independent Gaussian couplings (SK-model). In section 7 we study the learning performance of algorithms based on local quadratic cost functions and we compute the optimal local quadratic cost function. In section 8 we show that an optimal quadratic function provides the best local estimator for the couplings. Section 9 introduces further applications of the cavity method which allow us to simplify order parameters corresponding to the true teacher couplings. As an example, we compute the reconstruction error for an Ising model with Wishart distributed, i.e. weakly dependent couplings. The method is also applied to re-derive a simple mean field approximation to the maximum likelihood estimator. Section 10 explains how the

J. Stat. Mech. (2017) 063406

mean field estimator can be obtained from a local cost function and presents results for the reconstruction errors. Section 11 discusses the asymptotics of the reconstruction errors for large number of data and relates these results to expressions known from classical statistics. Section 12 contains comparisons of our results with those of simulations of the estimators and section 13 presents a summary and an outlook.

2. Estimators for the inverse Ising model

Let us consider a system of N binary spin variables $\boldsymbol{\sigma} = (\sigma_0, \ldots, \sigma_{N-1})$ connected by pairwise interactions J_{ij} and subject to external local fields H_i. The probability distribution of the spin set is given by the Boltzmann equilibrium distribution

$$P(\boldsymbol{\sigma}|\boldsymbol{J}, \boldsymbol{H}) = Z_{\text{Ising}}^{-1} \exp\left[\beta \sum_{i<j} J_{ij}\sigma_i\sigma_j + \beta \sum_i H_i\sigma_i\right], \tag{1}$$

where Z_{Ising} is the partition function and β is the inverse temperature. Given a set of M independent observations $\{\boldsymbol{\sigma}^k\}_{k=1}^M$ drawn independently from (1), the inverse Ising problem consists of estimating the model parameters $\boldsymbol{H}$ and $\boldsymbol{J}$ from the data. A standard approach for parameter estimation is the maximum likelihood (ML) method, which has the properties of consistency and asymptotic efficiency [21]. Maximum likelihood can be formulated as the minimisation of the following cost function (negative log-likelihood)

$$E_{\text{ML}}(\boldsymbol{J}, \boldsymbol{H}) = -\sum_{k=1}^{M} \ln P(\boldsymbol{\sigma}^k|\boldsymbol{J}, \boldsymbol{H}) \tag{2}$$

with respect to the matrix of couplings $\boldsymbol{J}$ and the field vector $\boldsymbol{H}$. As is well known, the minimisation of (2) is equivalent to a simple set of conditions for the first and second moments of the ensemble (1) of spins: the parameters estimated by ML lead to the matching of the empirical (data averaged) magnetisations to the magnetisation given by the model (1). Likewise we have the matching of all empirical pair correlations of spins with their model counterparts. Despite the simplicity of this rule, the practical minimisation of (2) requires the computation of these spin moments for a given set of couplings and fields which is equivalent to averaging over 2^N spin configurations, which is intractable for larger N. An approximation of such averages by Monte Carlo sampling is possible but requires sufficient time for equilibration. Alternatively, different approximation techniques have been developed to provide a good estimate of the parameters at a smaller computational cost, see e.g. [8, 9, 11, 12, 22–25].

3. Local learning

If we neglect the symmetry of coupling matrix, i.e. the equality $J_{ij} = J_{ji}$, we can develop estimators which learn the 'ingoing' coupling vectors J_{ij} for $j = 0, \ldots, i-1, i+1, \ldots, N-1$ for each spin σ_i independently. It turns out that the

J. Stat. Mech. (2017) 063406

corresponding (local) algorithms can often be performed in a much more efficient way compared to the ML method.

In the following we will concentrate on the estimation of the couplings only and set the external fields H_i to zero. We will specialise on the couplings for spin σ_0 and assuming that the typical couplings J_{ij} are variables with magnitude scaling like $1/\sqrt{N}$ for large N. We define a vector of rescaled couplings (weights) as

$$\boldsymbol{W} = (W_1, \ldots, W_{N-1}) \doteq \sqrt{N}(J_{01}, \ldots, J_{0N-1}). \tag{3}$$

We will assume that an estimator for $\boldsymbol{W}$ is defined by the minimisation of a cost function

$$E(\boldsymbol{W}) = \sum_{k=1}^{M} \mathcal{E}(\boldsymbol{W}; \boldsymbol{\sigma}^k) \tag{4}$$

which is additive in the observed data. An important and widely used case is the pseudo-likelihood approach, where the cost function

$$\begin{aligned} \mathcal{E}(\boldsymbol{W}; \boldsymbol{\sigma}) &= -\ln P(\sigma_0 | \boldsymbol{\sigma}_{\setminus 0}, \boldsymbol{W}) \\ &= -\beta\sigma_0 \sum_{j\neq 0} \frac{W_j \sigma_j}{\sqrt{N}} + \ln\left(2\cosh\beta \sum_{j\neq 0} \frac{W_j\sigma_j}{\sqrt{N}}\right) \end{aligned} \tag{5}$$

is given by the negative log-probability of spin σ_0 conditioned on all other spins $\boldsymbol{\sigma}_{\setminus 0}$. In contrast to the ML approach, the gradient of this function can be computed in an efficient way.

4. Teacher–student scenario and statistical physics analysis

We assume in the following that data are generated independently at random from a 'teacher' network with coupling matrix J^*_{ij}. A local learning algorithm based on the minimisation of (4) produces 'student' network couplings $\boldsymbol{W}$ as estimators for the teacher network couplings $\boldsymbol{W}^* = \sqrt{N}(J^*_{01}, \ldots, J^*_{0N-1})$. To measure the quality of a given local learning algorithm, we will compute the average square reconstruction error given by

$$\varepsilon = N^{-1}\overline{(\boldsymbol{W}^* - \boldsymbol{W})^2} = Y - 2\rho + Q, \tag{6}$$

where we define order parameters

$$Y = N^{-1}\overline{(\boldsymbol{W}^*)^2} \qquad Q = N^{-1}\overline{(\boldsymbol{W})^2} \qquad \rho = N^{-1}\overline{\boldsymbol{W}^* \cdot \boldsymbol{W}}, \tag{7}$$

representing, respectively, the squared lengths of the teacher and student coupling vectors and the overlap between teacher and a student coupling vectors. Here the overline defines an expectation over the ensemble of $M = \alpha N$ training data drawn at random from an Ising model with teacher couplings $\boldsymbol{J}^*$, i.e.

$$\overline{(\ldots)} = \sum_{\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^M} \prod_{k=1}^{M} P(\boldsymbol{\sigma}^k | \boldsymbol{J}^*)(\ldots). \tag{8}$$

J. Stat. Mech. (2017) 063406

Since there is often no explicit analytical solution to the minimisers $\boldsymbol{W}$ of (4), we will resort to a statistical physics approach which has been successfully applied to the analysis of a great variety of problems related to learning in neural networks [15–17]. In this approach one defines a statistical ensemble of student weights by a Gibbs distribution [26]

$$p(\boldsymbol{W}) = \frac{1}{Z}\exp[-\nu E(\boldsymbol{W})], \tag{9}$$

with the partition function

$$Z = \int \mathrm{d}\boldsymbol{W}\exp[-\nu E(\boldsymbol{W})], \tag{10}$$

where $1/\nu$ represents an effective temperature which controls the fluctuations of the 'training energy' $E(\boldsymbol{W})$. Using techniques from statistical physics of disordered systems one computes order parameters at nonzero temperature and performs the limit $\nu \to \infty$ at the end of the calculation. The 'thermal average' $\langle \boldsymbol{W} \rangle$ with respect to the distribution (9) converges to the minimiser of the cost function $E(\boldsymbol{W})$. Order parameters can be extracted from the quenched average of the free energy F corresponding to (8) using the replica method:

$$F = -N^{-1}\nu^{-1}\overline{\ln Z} = -\lim_{n\to 0} N^{-1}\nu^{-1}\frac{\partial}{\partial n}\ln \overline{Z^n}, \tag{11}$$

where the average replicated partition function for integer n is given by

$$\overline{Z^n} = \int \prod_{a=1}^{n} \mathrm{d}\boldsymbol{W}^a \left\{ \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}|\boldsymbol{J}^*)\exp[-\nu \sum_{a=1}^{n} \mathcal{E}(\boldsymbol{W}^a;\boldsymbol{\sigma})] \right\}^{\alpha N}. \tag{12}$$

To allow for an analytical treatment, we assume that the local cost function $\mathcal{E}(\boldsymbol{W}^a;\boldsymbol{\sigma})$ depends on the spins and couplings only via σ_0 and the local field $h \doteq \frac{1}{\sqrt{N}}\sum_{j\neq 0} W_j\sigma_j$ in the following way:

$$\mathcal{E}(\boldsymbol{W};\boldsymbol{\sigma}) = \Phi(\sigma_0 h). \tag{13}$$

Obviously, the pseudo-likelihood cost function (5) belongs to this class of functions. The goal of the following section is to perform the expectation (12). The resulting expression depends on a set of order parameters and can for integer n be evaluated by standard saddle-point methods in the limit $N \to \infty$. Performing an analytical continuation for $n \to 0$ yields both the free energy and the self-averaging values of these order parameters. While in most previous applications [15–17] of this programme to learning in neural networks, the quenched average over data in (12) is straightforward, the required average over Ising spin configurations drawn from the distribution (1) cannot be performed (for arbitrary N) in closed form. One might attempt a solution to this problem by introducing a second set of replicas which would deal with the partition function Z_{Ising}^{-1} in the denominator of (1). We expect that such an approach can be carried out for random teacher couplings but may lead to complicated expressions which have to be carefully evaluated for $N \to \infty$. In the next Section we will use a simpler approach using ideas of the cavity method [20] which allows, under certain assumptions on the teacher coupling matrix $\boldsymbol{J}^*$, the explicit computation of the quenched average for $N \to \infty$.

J. Stat. Mech. (2017) 063406

5. Cavity approach I: quenched averages

In order to perform the quenched averages in (12), we will combine the replica approach with ideas of the so-called cavity method. In doing so we write the Gibbs distribution (1) corresponding to the teacher couplings in the form

$$P(\boldsymbol{\sigma}|\boldsymbol{J}^*) \propto \exp\left[\beta\sigma_0 \sum_{j\neq 0} J^*_{0j}\sigma_j\right] P_{\rm cav}(\boldsymbol{\sigma}\backslash\sigma_0), \tag{14}$$

where $P_{\rm cav}$ denotes the distribution of the remaining spins in a system where the spin σ_0 was *removed*, creating a cavity at this site, which gives the method its name. The replicated partition function depends only on the fields $h_a \doteq \frac{1}{\sqrt{N}}\sum_{j\neq 0} W^a_j\sigma_j$ where $a \in \{*, 1, \ldots, n\}$. The cavity assumption for the statistics of such fields in densely connected systems can be summarised as follows: in performing expectations over $P_{\rm cav}$, we can assume that dependencies between spins are so weak that random variables h_a become jointly Gaussian distributed in the limit $N \to \infty$. Hence, the joint distribution of spin σ_0 and the fields can be expressed as

$$P(\sigma_0, h_*, h_1, \ldots h_n) = \frac{1}{Z_0}\mathrm{e}^{\beta\sigma_0 h_*} p_{\rm cav}(h_*, h_1, \ldots h_n) \tag{15}$$

with the normalisation

$$Z_0 = 2\int \cosh(\beta h_*) p_{\rm cav}(h_*)\mathrm{d}h_*. \tag{16}$$

Assuming that in absence of external fields we have vanishing magnetisations (paramagnetic phase), the distribution $p_{\rm cav}(h_*, h_1, \ldots h_n)$ is a multivariate Gaussian density with zero mean and covariance

$$\langle h_a h_b\rangle = \frac{1}{N}\sum_{i,j\neq 0} W^a_i C^{\backslash 0}_{ij} W^b_j. \tag{17}$$

The matrix $C^{\backslash 0}$ is the correlation matrix of the reduced spin system (without σ_0), which does not depend on the couplings $\boldsymbol{W}^*$. We have $C^{\backslash 0}_{ii} = 1$ and assume that typically $C^{\backslash 0}_{ij} = O(\frac{1}{\sqrt{N}})$ for $i \neq j$ and large N. However, this scaling does not mean that we can neglect the non-diagonal matrix elements. We will later see that they give *nontrivial contributions* to the final reconstruction error. Within this framework, the quenched average in (12) is rewritten in terms of integrals over the random variables h_a as follows:

$$\begin{aligned}&\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}|\boldsymbol{J}^*)\exp[-\nu\sum_{a=1}^{n}\mathcal{E}(\boldsymbol{W}^{\boldsymbol{a}};\boldsymbol{\sigma})]\\ &= \sum_{\sigma_0}\int \mathrm{d}h_* \prod_{a=1}^{n}\mathrm{d}h_a \frac{1}{Z_0}\exp\left[\beta\sigma_0 h_*\right]\exp\left[-\nu\sum_{a=1}^{n}\Phi(\sigma_0 h_a)\right] p_{\rm cav}(h_*, h_1, \ldots h_n).\end{aligned} \tag{18}$$

This result can be expressed by the covariances (17) which in the limit $N \to \infty$ will become *self averaging order parameters* which will be computed by the replica method (appendix A). Under the assumption of replica symmetry (which is expected to be

J. Stat. Mech. (2017) 063406

correct for convex cost functions, which holds e.g. in the case of pseudo-likelihood), these new order parameters and their physical meaning are denoted as:

$$
\begin{aligned}
V &\doteq \frac{1}{N}\sum_{i,j\neq 0} W_i^* C_{ij}^{\backslash 0} W_j^* \\
R &\doteq \frac{1}{N}\sum_{i,j\neq 0} W_i^* C_{ij}^{\backslash 0} \langle W_j\rangle_w = \frac{1}{N}\sum_{i,j\neq 0} W_i^* C_{ij}^{\backslash 0} W_j^a \qquad a\neq *, \\
q_0 &\doteq \frac{1}{N}\sum_{i,j\neq 0} \langle W_i W_j\rangle_w C_{ij}^{\backslash 0} = \frac{1}{N}\sum_{i,j\neq 0} W_i^a C_{ij}^{\backslash 0} W_j^a \qquad a\neq *, \\
q &\doteq \frac{1}{N}\sum_{i,j\neq 0} \langle W_i\rangle_w C_{ij}^{\backslash 0} \langle W_j\rangle_w = \frac{1}{N}\sum_{i,j\neq 0} W_i^a C_{ij}^{\backslash 0} W_j^b \quad a\neq b\neq *,
\end{aligned}
\tag{19}
$$

where the brackets $\langle\ldots\rangle_w$ denote averages with respect to the distribution of couplings (9).

6. Replica result

Using a replica symmetric ansatz, the computations follow the approach summarised in appendix A. In the zero temperature limit $\nu\to\infty$ the fluctuations of student couplings vanish and we obtain the convergence of the order parameters $q_0\to q$ with the limiting 'susceptibility'

$$
x \doteq \lim_{\nu\to\infty}(q_0-q)\nu = \lim_{\nu\to\infty}\frac{\nu}{N}\sum_{i,j\neq 0}\left(\langle W_i W_j\rangle_w - \langle W_i\rangle_w\langle W_j\rangle_w\right) C_{ij}^{\backslash 0}
$$

remaining finite and nonzero. As a main result, we find that the auxiliary order parameters (19) are obtained by extremizing the limiting free energy function

$$
F = -\operatorname*{extr}_{q,R,x}\left\{\frac{1}{2}\frac{q-R^2/V}{x} + \alpha\int \mathrm{d}v\, G_{\beta R,q}(v)\ \max_y\left[-\frac{(y-v)^2}{2x}-\Phi(y)\right]\right\},
\tag{20}
$$

where $G_{\mu,\omega}(v)$ denotes a scalar Gaussian density with mean μ and standard deviation ω. Remarkably, this free energy does (for any fixed cost function Φ) only depend on the teacher couplings $\boldsymbol{J}^*$ via the order parameter V, defined in equation (19). To compute the prediction error, however, we need the 'original' order parameters (7). These can be expressed by the auxiliary ones q, R and x. This relation can be derived from the free energy (appendix C) in a standard way by adding corresponding external fields to the 'Hamiltonian' in the Gibbs free energy (9). This relation brings back further statistics related to the teacher couplings $\boldsymbol{J}^*$ via

$$
\begin{aligned}
\rho &= \frac{RY}{V}, \\
Q &= \left(q-\frac{R^2}{V}\right)\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}} + \frac{R^2 Y}{V^2},
\end{aligned}
\tag{21}
$$

with the corresponding reconstruction error

J. Stat. Mech. (2017) 063406

$$\varepsilon = (q - \frac{R^2}{V})\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}} + Y(1 - \frac{R}{V})^2. \tag{22}$$

In deriving these results, we have also assumed that for $N \to \infty$, $\frac{1}{N}\mathrm{Tr}\overline{(\boldsymbol{C}^{\setminus 0})^{-1}} \to \frac{1}{N}\mathrm{Tr}\overline{(\boldsymbol{C})^{-1}}$. Note that the prediction error is larger than the one we would get if we had neglected the off-diagonal elements of the correlation matrix $\boldsymbol{C}^{\setminus 0}$. The error (22) depends on the teacher couplings $\boldsymbol{J}^*$ through the parameter Y and the parameter V (the cavity variance of the teacher field) and through the trace of the inverse correlation matrix $\boldsymbol{C}$ corresponding to the teacher's spin distribution. We will show later that the latter quantity can be expressed by the former using a second application of the cavity method. In the next section, we will see that the parameter V can be estimated from the data.

We will illustrate the result (22) for the case of random teacher couplings J^*_{ij} drawn independently for $i < j$ from a Gaussian density of variance 1. This corresponds to the celebrated *Sherrington–Kirkpatrick* (SK) model [27]. For $\beta < 1$, i.e. outside of the spin-glass phase, our simple form of the cavity arguments are known to be correct [20] and one finds the values

$$\begin{aligned} &V = Y = 1, \\ &\lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\overline{(\boldsymbol{C})^{-1}} = 1 + \beta^2, \end{aligned} \tag{23}$$

for zero magnetisations $m_i = 0$ in the literature [28]. A comparison of the theory (22) with numerical simulations is shown in section 12.

7. Quadratic cost functions

Among the simplest functions satisfying the property (13), we consider quadratic cost functions of the form

$$E_\eta(\boldsymbol{W}) = \frac{1}{2}\sum_{i\neq 0, j\neq 0} W_i \hat{C}_{ij} W_j - \eta\sqrt{N}\sum_{j\neq 0}\hat{C}_{0j}W_j, \tag{24}$$

where the empirical correlation matrix is defined as

$$\hat{C}_{ij} \doteq \frac{1}{M}\sum_{k=1}^{M}\sigma_i^k\sigma_j^k. \tag{25}$$

These allow for an explicit computation of the estimator in terms of a matrix inversion. The estimator minimizing (24) is given by

$$W_i^\eta = \eta\sqrt{N}\sum_{j\neq 0}(\hat{C}_{-0}^{-1})_{ji}\hat{C}_{0j} \qquad i \neq 0, \tag{26}$$

where the matrix $\hat{C}_{-0}$ is the submatrix of $\hat{C}$ where the 0th column and 0th row are deleted (not to be confused with the cavity matrix $C^{\setminus 0}$) and η is a free parameter. The estimation error can be computed from the free energy (20) by setting

J. Stat. Mech. (2017) 063406

$$\Phi(h) = \frac{h^2}{2} - \eta h, \tag{27}$$

and gives (see appendix D)

$$\varepsilon = \left(\frac{\beta\eta}{1+\beta^2 V} - 1\right)^2 Y + \frac{\eta^2}{(\alpha-1)(1+\beta^2 V)} \frac{1}{N} Tr\overline{\boldsymbol{C}^{-1}}. \tag{28}$$

The optimal choice for the quadratic cost function (24) is found by fixing the parameter η to the value that minimizes the error (28), namely

$$\eta_{\text{opt}} = \frac{(\alpha-1)(1+\beta^2 V)\beta Y}{(\alpha-1)\beta^2 Y + (1+\beta^2 V)\frac{1}{N}\text{Tr}\overline{\boldsymbol{C}^{-1}}}, \tag{29}$$

with the corresponding minimal error

$$\varepsilon_{\text{opt}} = \frac{(1+\beta^2 V) Y \frac{1}{N}\text{Tr}\overline{\boldsymbol{C}^{-1}}}{(\alpha-1)\beta^2 Y + (1+\beta^2 V)\frac{1}{N} Tr\overline{\boldsymbol{C}^{-1}}}. \tag{30}$$

In general, the computation of the optimal parameter η_{opt} requires the knowledge of the three parameters Y, V and $\frac{1}{N}\text{Tr}\overline{\boldsymbol{C}^{-1}}$ which characterise the statistical ensemble to which he unknown teacher matrix $\boldsymbol{J}^*$ belongs. However, (29) simplifies as $\alpha \to \infty$ and we get

$$\lim_{\alpha\to\infty} \eta_{\text{opt}} = \frac{1+\beta^2 V}{\beta}. \tag{31}$$

We will now show that the remaining parameter V can be estimated from the observed data. We use the fact that at its minimum, the cost function (24) equals

$$E_\eta(\boldsymbol{W}^\eta) = -\frac{N}{2}\eta^2 \Delta, \tag{32}$$

where we have used (26) and defined

$$\Delta = \sum_{i\neq 0, j\neq 0} \hat{C}_{0i} (\hat{C}_{-0}^{-1})_{ij} \hat{C}_{0j}, \tag{33}$$

which only depends on the spin data. On the other hand in the situations where our statistical physics formalism applies, the minimal training energy (32) will be self-averaging in the thermodynamic limit $N \to \infty$ and can be computed as the *zero temperature limit of the free energy*, i.e. the free energy function (20) evaluated at the stationary values of the order parameters. The calculation in (appendix D) yields

$$\Delta = \frac{1+\alpha\beta^2 V}{\alpha(1+\beta^2 V)}. \tag{34}$$

This shows, that the unknown parameter V and the asymptotically optimal parameter η can be directly estimated from the observed spin correlations.

In the next section, we will show that the optimal *quadratic* cost function yields in fact the *total* optimum of the reconstruction error with respect to free variations of the cost function Φ.

J. Stat. Mech. (2017) 063406

8. The optimal local cost function

In this section, we will derive the form of the optimal local cost function Φ within the cavity/replica approach and show that it is quadratic. Hence, the results of the previous section can be applied, where the optimal quadratic cost function was already computed. We will give a derivation of this fact for the case of finite inverse 'temperature' ν, assuming that the argument can be continued to $\nu \to \infty$.

The optimisation of cost functions for learning problems within the replica approach goes back to the work of Kinouchi and Caticha [29]. We will follow the framework of [30] (see also [31]). Our goal is to minimise an error measure for a learning problem which is of the form $\varepsilon(R, q, q_0)$ such as (22). It depends on order parameters which are computed by setting the derivatives of a free energy function $F_\Phi(R, q, q_0)$ (such as A.10) equal to zero. The main idea is to take these conditions into account within a Lagrange function

$$\varepsilon(R, q, q_0) + \sum_{S \in R, q, q_0} \lambda_S \frac{\partial}{\partial S} F_\Phi(R, q, q_0), \tag{35}$$

where the λ_S are the corresponding Lagrange multipliers. The optimal function Φ is obtained from the variation

$$\frac{\delta}{\delta \Phi} \sum_{S \in R, q, q_0} \lambda_S \frac{\partial}{\partial S} F_\Phi(R, q, q_0) = 0. \tag{36}$$

For our problem, we can write (see (A.2) and (A.10))

$$F_\Phi(R, q, q_0) = F_0(R, q, q_0) - \frac{\alpha}{\nu} \int G_{\beta R, q}(v) \ln \Psi_{q_0 - q}(v) \mathrm{d}v, \tag{37}$$

where $F_0(R, q, q_0)$ is independent of Φ and $G_{\mu,\omega}(v)$ denotes a scalar Gaussian density with mean μ and standard deviation ω. The free energy depends on Φ through the function

$$\Psi_{q_0 - q}(v) \doteq \int G_{v, q_0 - q}(y)\, \mathrm{e}^{-\nu \Phi(y)} \mathrm{d}y. \tag{38}$$

We will first derive a condition on the form of the optimal function Ψ from the variation

$$\frac{\delta}{\delta \Psi} \sum_{S \in R, q, q_0} \lambda_S \frac{\partial}{\partial S} \int G_{\beta R, q}(v) \Psi_{q_0 - q}(v) \mathrm{d}v = 0. \tag{39}$$

From this, we will recover the form of the optimal Φ. To obtain the derivatives with respect to the order parameters we use the following rules for expectations over Gaussian measures, which can be easily derived using integration by parts

$$\frac{\partial}{\partial \mu} \int G_{\mu,\omega}(v) f(v) \mathrm{d}v = \int G_{\mu,\omega}(v) \partial_v f(v) \mathrm{d}v, \tag{40}$$

$$\frac{\partial}{\partial \omega} \int G_{\mu,\omega}(v) f(v) \mathrm{d}v = \frac{1}{2} \frac{\partial^2}{\partial \mu^2} \int G_{\mu,\omega}(v) f(v) \tag{41}$$

J. Stat. Mech. (2017) 063406

$$= \frac{1}{2}\int G_{\mu,\omega}(v)\partial_v^2 f(v)\mathrm{d}v. \tag{42}$$

Hence, the derivatives required for (39) are

$$\frac{\mathrm{d}}{\mathrm{d}R}\int G_{\beta R,q}(v)\ln\Psi_{q_0-q}(v)\mathrm{d}v = \beta\int G_{\beta R,q}(v)\ \partial_v \ln\Psi_{q_0-q}(v)\mathrm{d}v, \tag{43}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}q_0}\int G_{\beta R,q}(v)\ \ln\Psi_{q_0-q}(v) &= \frac{1}{2}\int G_{\beta R,q}(v)\ \frac{\partial_v^2\Psi_{q_0-q}(v)}{\Psi_{q_0-q}(v)}\mathrm{d}v \\ &= \frac{1}{2}\int G_{\beta R,q}(v)\ \left\{\partial_v^2\ln\Psi_{q_0-q}(v) + (\partial_v\ln\Psi_{q_0-q}(v))^2\right\}\mathrm{d}v,\end{aligned} \tag{44}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}q}\int G_{\beta R,q}(v)\ \ln\Psi_{q_0-q}(v)\mathrm{d}v = \frac{\beta^2}{2}\int G_{\beta R,q}(v)\ \partial_v^2\ \ln\Psi_{q_0-q}(v)\mathrm{d}v \\ -\frac{\mathrm{d}}{\mathrm{d}q_0}\int G_{\beta R,q}(v)\ \ln\Psi_{q_0-q}(v)\mathrm{d}v.\end{aligned} \tag{45}$$

An application of standard variational calculus to a linear combination of these order parameter derivatives shows that

$$\partial_v\ln\Psi_{q_0-q}(v) = c_1 + c_2\partial_v\ln G_{\beta R,q}(v), \tag{46}$$

where $c_{1,2}$ are independent of v. Since the logarithm of the Gaussian density $\ln G_{\beta R,q}(v)$ is a quadratic function in v, we conclude that also $\ln\Psi_{q_0-q}(v)$ is a quadratic expression in the variable v, making $\Psi_{q_0-q}(v)$ a (non-normalised) Gaussian density.

To conclude our argument on the optimal form of Φ, we use relation (38). This shows that the Gaussian density $\Psi_{q_0-q}(v)$ is the convolution of a (non-normalised) Gibbs density $\mathrm{e}^{-\nu\Phi(y)}$ of a random variable y with the density $G_{v,q_0-q}(y) = G_{y,q_0-q}(v)$ of a Gaussian random variable v. As a convolution corresponds to the addition two random variables, we know that $v+y$ is also a Gaussian random variable. Since v is Gaussian, then $\mathrm{e}^{-\nu\Phi(y)}$ is also a Gaussian density and $\Phi(y)$ is quadratic in y. We have a already computed the best quadratic cost function in the previous Section, and we conclude that the estimator (26) with (29) is the best local estimator of the couplings.

9. Cavity approach II: TAP equations and approximate mean field ML estimator

So far we have ignored the symmetry of the coupling matrix by restricting ourselves to estimators derived from local cost functions. In this Section, we will discuss a well known approximation [32] of the (symmetric) maximum likelihood estimator which is based on mean field theory. We will re-derive this estimator using the more advanced (adaptive) TAP mean field theory, because its results for the spin correlation matrix will also be needed in the following. We will later compute its reconstruction error in section 10. Our starting point is a generalisation of the well known TAP mean field approach developed for the SK model. Using the cavity approach [32] one derives the following 'adaptive' TAP equations for the magnetisations

J. Stat. Mech. (2017) 063406

$$m_i = \tanh\left(\beta \sum_j J_{ij} m_j - \beta^2 V_i m_i + \beta H_i\right), \tag{47}$$

where

$$V_i = \left\langle \left\{ \sum_j J_{ij}(\sigma_j - \langle \sigma_j \rangle_{\setminus i}) \right\}^2 \right\rangle_{\setminus i} \tag{48}$$

is the variance of the cavity field at spin i. Using a linear response argument (i.e. by taking the derivative of m_i equation (47) with respect to H_j), one obtains the following cavity approximation to the susceptibility $\chi_{ij} = C_{ij} - m_i m_j$, i.e. the covariance matrix of the spins:

$$\boldsymbol{\chi}(\boldsymbol{J}) = (\boldsymbol{\Lambda} - \beta \boldsymbol{J})^{-1}, \tag{49}$$

where the diagonal matrix $\boldsymbol{\Lambda}$ has elements

$$\Lambda_{ii} = \beta^2 V_i + \frac{1}{\chi_{ii}} = \beta^2 V_i + \frac{1}{1 - m_i^2}. \tag{50}$$

From this result, we can draw the following conclusions:

(i) Writing the moment matching conditions for the maximum likelihood estimator as

$$C_{ij}(J) \doteq \langle \sigma_i \sigma_j \rangle = \hat{C}_{ij} \doteq \frac{1}{M} \sum_{k=1}^{M} \sigma_i^k \sigma_j^k \tag{51}$$

and specialising to the paramagnetic case $H_i = m_i = 0$, we have $\boldsymbol{C}(\boldsymbol{J}) = \boldsymbol{\chi}(\boldsymbol{J})$. Hence, the cavity approximation (49) yields the mean field (MF) estimator given by [6]

$$J_{ij}^{\mathrm{MF}} = -\frac{1}{\beta} \left(\hat{\boldsymbol{C}}^{-1}(\boldsymbol{J})\right)_{ij} \quad \text{for } i \neq j. \tag{52}$$

At first glance, this simple and explicit form of a (symmetric) coupling estimator does not seem to fit into the framework developed in this paper. Surprisingly, we will derive a local cost function in the next section which allows for the computation of the reconstruction error using the statistical physics approach.

(ii) Inverting (49) and using (50) for $m_i = 0$, we get an expression for the trace of the inverse spin correlation matrix in terms of the variances of the cavity fields at all spins which is given by

$$\frac{1}{N} \mathrm{Tr} \overline{\boldsymbol{C}^{-1}} = \frac{\beta^2}{N} \sum_i V_i + 1. \tag{53}$$

J. Stat. Mech. (2017) 063406

If we assume that the teacher couplings $\boldsymbol{J}^*$ can be viewed as generated from a random matrix ensemble for which the V_i become self-averaging, i.e. $V_i \equiv V$ as $N \to \infty$ we finally obtain the simple result

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{Tr} \overline{\boldsymbol{C}^{-1}} = \beta^2 V + 1. \tag{54}$$

With this result, we can eliminate another unknown parameter of the teacher's ensemble of couplings, as we have shown that V can be estimated from the observed spin data, see (33) and (34).

Equation (54) agrees with the special result (23) for the SK model, since the 'Onsager correction' in the TAP equations for the SK model gives $V = 1$. As an application of the general result (54), we present numerical results for the reconstruction error for the Wishart ensemble in section 12, where the couplings are given by

$$J^*_{ij} = \frac{1}{N} \sum_{\mu=1}^{\gamma N} \xi_i^\mu \xi_j^\mu \tag{55}$$

and the ξ_j^μ are independent zero mean Gaussian random variables with unit variance. The thermodynamics of this model agrees with that of the celebrated Hopfield model of a neural network (where $\xi_i^\mu = \pm 1$) [33], in the phase where there is no macroscopic overlap between the spin configurations and a stored pattern. Hence, we can read off the cavity variance from the TAP mean field equations obtained by [26], setting $m_i = 0$. One finds

$$V = \frac{\gamma}{1-\beta}. \tag{56}$$

For other random matrix ensembles which are invariant against orthogonal transformations it is possible to obtain a general expression for the cavity variance in terms of the so-called R-transform of the matrix ensemble (for details, see [34, 35]) and can be expressed by the limiting eigenvalue spectrum of the matrices.

10. Reconstruction error for MF-ML estimator

We will now turn to the computation of the reconstruction error for the MF-ML estimator (52). At first glance, this estimator does not seem to be related to a local cost function in the style of (4). But surprisingly, it is not hard to construct such a function. If we specialise again to the estimation of the coupling vector $\boldsymbol{W}$ corresponding to spin σ_0, we can simplify the estimator (52) using the matrix inversion lemma [36] in the form

$$W_i^{\mathrm{MF}} = \sqrt{N} J_{0i}^{\mathrm{MF}} = -\frac{\sqrt{N}}{\beta} (\hat{C}^{-1})_{0i} = \frac{\sqrt{N}}{\beta} \phi_0 \sum_{j\neq 0} (\hat{C}_{-0}^{-1})_{ji} \hat{C}_{0j} \qquad i \neq 0, \tag{57}$$

J. Stat. Mech. (2017) 063406

where

$$\phi_0 = \frac{1}{1 - \sum_{i,j\neq 0} \hat{C}_{0i}(\hat{C}_{-0}^{-1})_{ij}\hat{C}_{0j}} = \frac{1}{1-\Delta}, \tag{58}$$

where Δ was introduced in (33). Assuming as before, that Δ is self-averaging for $N \to \infty$, the mean field estimator is of the form (26) and is associated to a cost function of the form (24). Hence, the results of section 7 apply. In particular, from (34) and (58) we compute

$$\phi_0 = \frac{1}{1-\Delta} = \frac{\alpha(1+\beta^2 V)}{\alpha - 1}, \tag{59}$$

and the estimation error is given by (28) with the parameter $\eta_{\mathrm{MF}} = \phi_0/\beta$:

$$\varepsilon_{\mathrm{MF}} = \frac{Y}{(\alpha-1)^2} + \frac{\alpha^2}{\beta^2(\alpha-1)^3}(1+\beta^2 V)\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}}. \tag{60}$$

11. Asymptotics

We will now investigate the limiting scaling of the reconstruction error as the number of data M grows much larger than the number of parameters (per spin) N to be estimated. This means we consider the limit $\alpha \to \infty$. This is of special interest, because we can compare the results obtained by our replica/cavity approach with results derived independently by standard arguments of classical statistics. From (30) and (60) and appendix E we can see that as $\alpha \to \infty$, the scaling of the reconstruction errors for the pseudo-likelihood estimator, the optimal local estimator and the mean field estimator is

$$\varepsilon \simeq \frac{c}{\alpha}, \tag{61}$$

where

$$c_{\mathrm{PLM}} = \frac{1}{\beta^2}\frac{1}{\int \mathrm{d}v\, G_{\beta V,V}(v)\left(1-\tanh^2(\beta v)\right)}\frac{1}{N}\mathrm{Tr}\overline{C^{-1}}, \tag{62}$$

$$c_{\mathrm{OPT}} = \frac{1+\beta^2 V}{\beta^2}\frac{1}{N}\mathrm{Tr}\overline{\boldsymbol{C}^{-1}} = \frac{(1+\beta^2 V)^2}{\beta^2}, \tag{63}$$

$$c_{\mathrm{MF-ML}} = c_{\mathrm{OPT}}, \tag{64}$$

where in the second equality of the second line, we have used (54). Hence, asymptotically the simple mean field estimator and the optimal estimator converge to the true couplings at the same speed. Thus, one might conjecture that the mean field estimator is equivalent to the true maximum likelihood estimator in the thermodynamic limit, assuming that the cavity approach is correct.

J. Stat. Mech. (2017) 063406

The validity of the inequality $c_{\mathrm{PLM}} > c_{\mathrm{OPT}}$ given by (62) and (63) will depend on the temperature β and can be established at least for β small enough. For the SK model, this region covers the entire paramagnetic phase $\beta < 1$ where our simple cavity method is valid. However, the difference between the two is not very large for small β. In fact, expanding (62) in powers of β shows that error coefficients c for both estimators agree up to terms of order β^2. One may however argue that the comparison between the two estimators is not fair, because the pseudo-likelihood estimator does not yield symmetric couplings J_{ij} whereas the mean field one (and hence, asymptotically the optimal one) does. One might thus get a better estimate by a final symmetrisation. Unfortunately, with our present method, the effect of symmetrisation on the reconstruction error cannot be computed. We expect that methods of random matrix theory would be needed for this. Hence, we postpone a treatment of this problem to future publications. On the other hand, preliminary simulations show that the improvement of the pseudo likelihood estimator after symmetrisation is rather weak (at least for the systems with random couplings studied in this paper). This result is further supported by the fact that for small β, the pseudo likelihood estimator is already *almost symmetric*, a fact that can be easily shown, if we expand (5) for small β. The lowest order term yields an explicit result which is symmetric.

We want to compare the replica based asymptotics (62), (63) and (64) with exact asymptotic expressions for the errors of statistical estimators which are defined by the minimisation of smooth cost functions of the type (4), see e.g. [21] or [26] for an alternative derivation using replicas. The idea behind such asymptotic results is an expansion of the cost function in terms of the parameters $\boldsymbol{W}$ around the teacher parameters $\boldsymbol{W}^*$ (assuming convergence to the teacher in the infinite data limit). Setting $\delta \boldsymbol{W} \doteq \boldsymbol{W} - \boldsymbol{W}^*$ and using the law of large numbers and central limit arguments one can show the following equation for the data averaged correlations

$$\overline{\delta W_i \delta W_j} \simeq \frac{1}{N\alpha}[(\boldsymbol{U}^{-1}\boldsymbol{B}\boldsymbol{U}^{-1})_{ij}] \quad \text{for } \alpha \to \infty, \tag{65}$$

together with $\overline{\delta \boldsymbol{W}} \simeq 0$. The matrices are given by

$$\begin{aligned} U_{ij} &= \langle \partial_i \partial_j \mathcal{E}(\boldsymbol{W}^*; \boldsymbol{\sigma}) \rangle \,, \\ B_{ij} &= \langle \partial_i \mathcal{E}(\boldsymbol{W}^*; \boldsymbol{\sigma}) \partial_j \mathcal{E}(\boldsymbol{W}^*; \boldsymbol{\sigma}) \rangle \,. \end{aligned} \tag{66}$$

The partial derivatives are with respect to the components of $\boldsymbol{W}^*$ and the brackets are averages over spins using the distribution $P(\boldsymbol{\sigma}|\boldsymbol{J}^*)$. For the pseudo-likelihood case, (65) can be further simplified. In appendix F, we show that in this case $\boldsymbol{U} \equiv -\boldsymbol{B}$ and we finally obtain

$$\varepsilon \simeq N^{-1} \sum_i \overline{(\delta W_i)^2} = \frac{1}{N\alpha} \mathrm{Tr} B^{-1}, \tag{67}$$

with

$$B_{ij} = \beta^2 \left\langle \sigma_i \sigma_j \left[1 - \tanh^2 (\beta \frac{1}{\sqrt{N}} \sum_{j \neq 0} W^*_{0j} \sigma_j) \right] \right\rangle . \tag{68}$$

If we neglect the correlations between $\sigma_i\sigma_j$ and the field $\frac{1}{\sqrt{N}}\sum_{j\neq 0} W^*_{j0}\sigma_j$ for large N and note that $\langle\sigma_i\sigma_j\rangle = C_{ij}$, this result is in agreement with (62).

A similar calculation is possible for the OPT/MF-ML case. Here we get

$$\begin{aligned} U_{ij} &= \frac{\beta}{\phi_0 N} C_{ij}, \\ B_{ij} &= \frac{\beta^2}{\phi_0^2 N}\langle\sigma_i\sigma_j h_*^2\rangle + \frac{1}{N}C_{ij} - 2\frac{\beta}{\phi_0 N}\langle h_*\sigma_0\sigma_i\sigma_j\rangle. \end{aligned} \tag{69}$$

To obtain the asymptotics of the replica result (63) form these matrices, we assume that the dependencies between the random variables $\sigma_i\sigma_j$ on the one hand and respectively h_*^2 and $h_*\sigma_0$ on the other hand can be neglected for $N\to\infty$. Using the facts that $\langle h_*^2\rangle = \beta^2V^2 + V$, $\langle h_*\sigma_0\rangle = \beta V$ and $\lim_{\alpha\to\infty}\phi_0 = 1+\beta^2 V$ finally yields (63) .

12. Numerical results

In the previous sections we saw that the error of any algorithm that infers the network couplings by minimizing a cost function of the kind (13) satisfies (22), in the large N limit, when the cavity arguments apply. The order parameters are the ones extremizing the free energy (20). For pseudo-likelihood maximization, the set of equation (B.5) for the order parameters has to be solved numerically, whereas for the local optimal and MF-ML estimators we computed analytically the error in the form, respectively, (30) and (60). Note that the error (22) is expressed in terms of three parameters that depend on the distribution of the teacher couplings: Y, V and the trace of the inverse correlation matrix $\boldsymbol{C}$. As an example, we considered the the Gaussian ensemble of the SK model, with parameters given by $Y=1$ and relation (23) and the Wishart ensamble of (55) with parameters given by $Y=\gamma$ and relations (54) and (56). Figure 1 compares the predicted error with the mean squared error that we get from simulations, as a function of α. We show results for the pseudo-likelihood, local optimal and MF-ML algorithms applied to the SK and the Wishart model. We only report results for the high-temperature (paramagnetic) region, i.e. for $\beta<\beta_c$ where β_c defines the onset of spin-glass ordering. In this region, we expect that on the one hand, the cavity arguments are exact and the other hand, the convergence of the spin simulations to the thermal equilibrium is sufficiently fast. For the SK model, we have $\beta_c=1$ and for the Wishart model $\beta_c\simeq 1/(1+\sqrt{\gamma})$ for zero magnetization and small q [37], where q is the Edwards-Anderson order parameter. The data are generated by Monte Carlo sampling with a burn in time of 10^7 N spin updates and sampling every 10 N updates, and the couplings are recovered either by minimizing the pseudo-likelihood cost function (5) using a Newton method or from the empirical correlation matrices: see (26) and (29) for the local optimal algorithm and (52) for MF-ML. The plot shows that the replica calculation predicts rather well the results of the simulations for systems of $N=100$ spins. In addiction, it is clear that the optimal local algorithm outperforms the other two methods and, in the high-temperature regime considered here, the MF-ML algorithm performs better than pseudo-likelihood maximization. This performance difference is more relevant for increasing β and in the small α region, whereas it is almost negligible

J. Stat. Mech. (2017) 063406

for large α, in agreement with the asymptotic expansions. Finally, we compare the analytical results for the asymptotic behavior of the error computed in section 11 with the results from simulations. Assuming the scaling (61), we fitted the function $\varepsilon = c/\alpha$ to the mean squared error of the couplings inferred from simulations at large α. In table 1 we show that this 'experimental' value of c is consistent with c_{PLM} (62) and $c_{\mathrm{OPT}} = c_{\mathrm{MF-ML}}$ (63) and (64). We then plot the predicted value of c as a function of β in figure 2, where we can see that the difference between pseudo-likelihood maximization, the local optimal and MF-ML algorithms is almost indistinguishable and goes to zero for small β, as we would expect by noticing that the analytical formula for c_{PLM} (62) agrees with $c_{\mathrm{OPT}} = c_{\mathrm{MF-ML}}$ (63) and (64) up to second order in β. From the plot it is also clear that for larger β—i.e. smaller stochasticity of the spins—the error in predicting the couplings is smaller.

The three algorithms show different behaviors in the small α region. As the MF-ML algorithm relies on the inversion of the correlation matrix (52), that becomes singular at $\alpha = 1$, its error diverges at $\alpha = 1$, as can be seen from (60). On the contrary, the error of the optimal local algorithm shows no divergence, since $\eta_{opt} = 0$ and $\varepsilon = Y$ at $\alpha = 1$ (see (29) and (30)). From simulations we also observe that the error of the pseudo-likelihood estimator increases for decreasing α and for $\alpha < 2$ it reaches large values, with large variations across trials, while the extremization of the free energy (see B.5) fails in the region $\alpha < 2$. A way to overcome this divergence is to introduce a regularizing term in the objective function. We postpone the study of regularized estimators to future work. We present additional plots in appendix G, showing the error dependence on the system size.

13. Discussion and outlook

We have presented a statistical physics approach for calculating the reconstruction error of algorithms for learning the couplings of large Ising models. Our method assumes local cost functions for learning and is based on a combination of the replica trick and of cavity arguments for computing quenched averages over spin configurations which are drawn at random from a teacher network. A replica symmetric ansatz seems to be justified as long as the learning algorithms are based on convex cost functions. The cavity approach assumes a large densely connected network with couplings that are roughly of the same size leading to only weakly correlated spins. These assumptions are correct in the thermodynamic limit for certain statistical ensembles of network couplings but may also give good approximations for realistic networks. While our method is so far restricted to problems which are realisable by pairwise spin-interactions, it could nevertheless be of practical interest in providing approximate statistics for hypotheses testing against more complicated network models (having e.g. 3-spin interactions).

Our results show that the learning problem is, at least within our framework, surprisingly simple: an explicit estimator based on a quadratic cost function achieves minimal error and outperforms the more complicated pseudo-likelihood estimator. This optimal estimator only requires prior knowledge of the length of the true coupling vector. Moreover, a simple (symmetric) mean field approximation to the maximum likelihood estimator is asymptotically optimal and can be computed without such prior

J. Stat. Mech. (2017) 063406

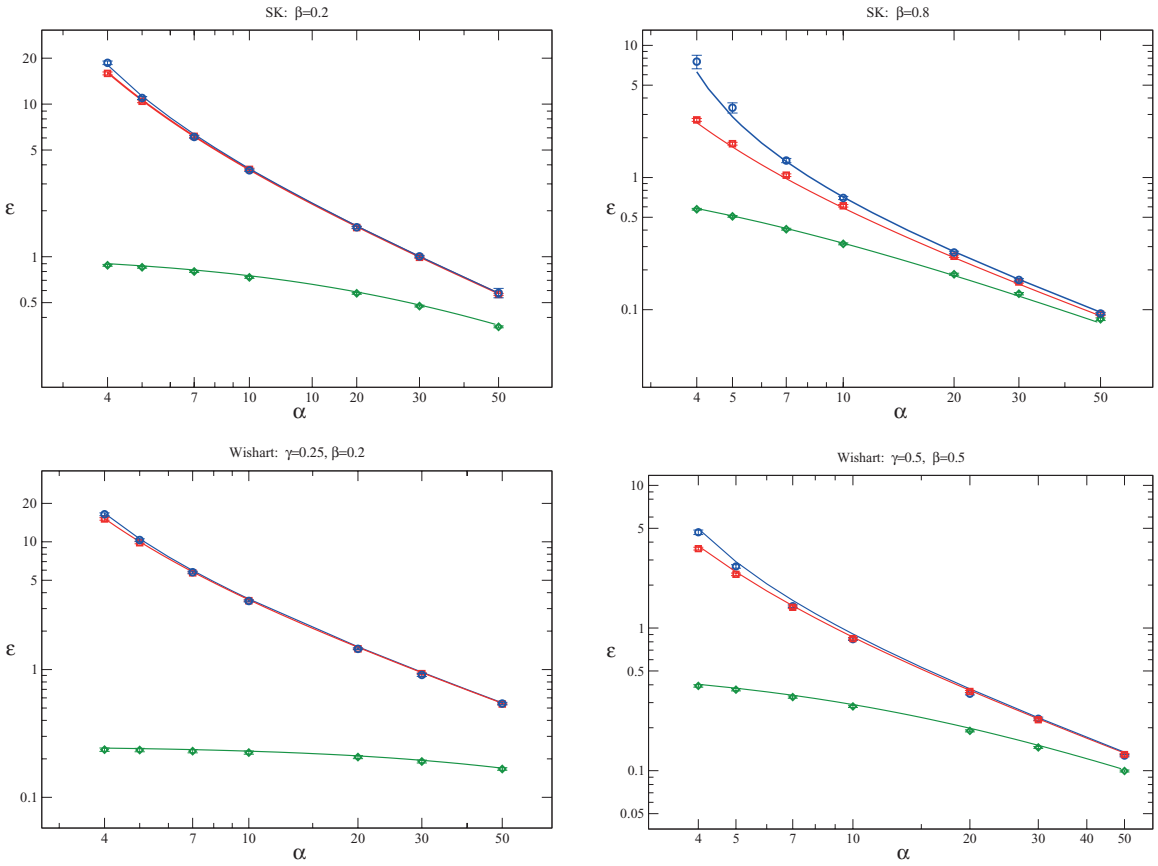

Figure 1. The mean squared error of the couplings inferred by using the pseudo-likelihood algorithm (blue dots) the optimal local algorithm (green dots) and the MF-ML algorithm (red dots) is compared to the corresponding average prediction error from the replica calculation (continuous lines). The error is plotted as a function of α. Four different systems are considered: SK model at $\beta = 0.2$ (top left), SK model at $\beta = 0.8$ (top right), Wishart model with $\gamma = 0.25$ at $\beta = 0.2$ (bottom left) and Wishart model with $\gamma = 0.5$ at $\beta = 0.5$ (bottom right). The algorithms were tested on a system of $N = 100$ spins and the results are averaged over 5 realizations of the network and 100 different datasets generated from each network. Error bars represent standard deviations of the means.

knowledge. In the case of the SK model, the region of small β in which these results hold covers the entire paramagnetic phase, where our simple cavity arguments are known to be valid. It would be interesting to work out analytically how well the mean field estimator approximates the exact maximum likelihood estimator in the thermodynamic limit.

Our work is only a first step to an understanding of the typical performance of learning algorithms for the inverse Ising problem. From a technical point of view our method could be generalised in several directions. We have restricted ourselves to models where data are sampled from the paramagnetic phase of a teacher network. While it is possible to generalise the analysis, the average over samples from a spin-glass phase would usually require more complex types of cavity arguments [38] which are related to the breaking of the replica symmetry of the teacher network. In such a case, the simple Gaussian

J. Stat. Mech. (2017) 063406

Table 1. The values of c_{PLM} (62) for pseudo-likelihood maximization and $c_{\mathrm{MF-ML}}$ (64) for the MF-ML algorithm are compared to the results 'c (simulations)' we obtained by fitting the function $\varepsilon = c/\alpha$ to the mean squared error ε of the inferred couplings obtained from simulations at large α. We considered two systems: SK model at $\beta = 0.6$ and Hopfield model with $\gamma = 0.15$ at $\beta = 0.6$. The algorithms were tested on a system of $N = 200$ spins for $\alpha = 900, 950, 1000$ and the results are averaged over 5 realizations of the network and 10 different datasets generated from each network. The errors on c are obtained by propagating the standard deviations of ε from simulations.

Model	Algorithm	c	c (simulations)
SK	PLM	5.199	5.16 ± 0.04
	MF-ML	5.137	5.14 ± 0.05
Wishart	PLM	3.582	3.64 ± 0.06
	MF-ML	3.578	3.60 ± 0.05

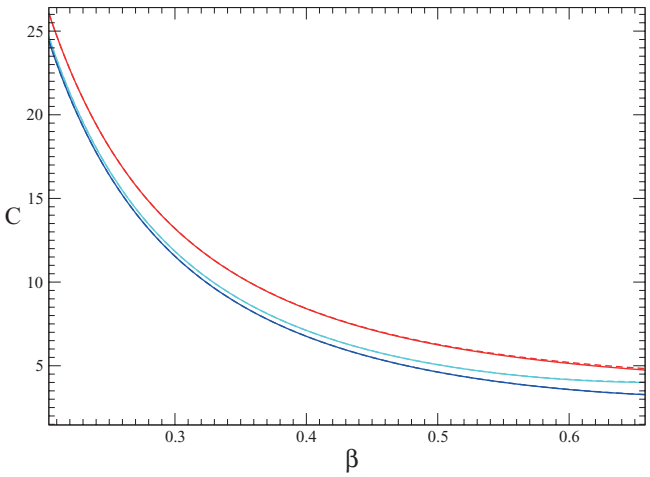

Figure 2. The values c_{PLM} (62) for pseudo-likelihood maximization (dotted lines) and $c_{\mathrm{MF-ML}}$ (64) for the MF-ML algorithm (continuous lines) are plotted as a function of β. The red lines correspond to the SK model, the blue lines to the Hopfield model with parameter $\gamma = 0.25$ (light blue) and $\gamma = 0.15$ (dark blue).

distribution of cavity fields on which our analysis strongly relies is no longer valid. One might expect that now the quadratic cost functions may no longer be optimal (and not even consistent) but could be outperformed by a pseudo-likelihood method.

We also expect that our cavity framework could be extended to sparse networks as long as the number of nonzero couplings per spin is large enough to allow for the application of the central limit arguments used in our work.

After finishing our work we became aware of a recent preprint [31] where similar learning problems (focussing on a teacher model with independent Gaussian couplings) were studied. The author applied a double replica calculation (the other set of replica are used for dealing with the partition function in the quenched average over the spins) instead of using cavity arguments. This results in a free energy function which agrees essentially with our result (20). However, the order parameters appearing in the free energy are not defined by (19) but by (7) instead, and the reconstruction error differs from ours. The major difference is that the result for the error in [31] does not contain

J. Stat. Mech. (2017) 063406

the spin-correlation matrix as in our equation (22). We believe that this could be related to an implicit approximation of the correlation matrix by a unit matrix.

Acknowledgments

We would like to thank Andrea Pagnani for fruitful discussions that significantly motivated our work.

Appendix A. Details of the replica calculation

From (11), (12) and (18) one can see that the free energy can be written as

$$F = -\lim_{n\to 0} N^{-1}\nu^{-1}\frac{\partial}{\partial n}\ln\int\prod_{a=1}^{n}\mathrm{d}\boldsymbol{W}^a\left\{\sum_{\sigma_0}\int \mathrm{d}h_*\prod_{a=1}^{n}\mathrm{d}h_a\frac{1}{Z_0}\exp\left[\beta\sigma_0 h_*\right]\right.$$
$$\left.\exp\left[-\nu\sum_{a=1}^{n}\Phi(\sigma_0 h_a)\right]\ p_{\mathrm{cav}}(h_*,h_1,\dots h_n)\right\}^{\alpha N}, \tag{A.1}$$

where $p_{\mathrm{cav}}(h_*,h_1,\dots h_n)$ is a multivariate Gaussian density with zero mean and covariance given by (17). The average over the Gaussian fields yields quadratic terms in $\boldsymbol{W}^*$ and $\{\boldsymbol{W}^a\}_{a=1}^n$, that can be simplified by introducing the order parameters $\{R,q,q_0\}$ (19), that have to be defined via integrals over delta functions. One finds that free energy decouples into two terms:

$$F(R,q,q_0) = F_0(R,q,q_0) + F_1(R,q,q_0). \tag{A.2}$$

The first one contains the integrals over the couplings and measures the density of the networks with order parameters R,q,q_0:

$$F_0(R,q,q_0) = -\lim_{n\to 0}\nu^{-1}\frac{\partial}{\partial n}\frac{1}{N}\ln Z_{\mathrm{coup}}, \tag{A.3}$$

with

$$Z_{\mathrm{coup}} = \int\prod_a \mathrm{d}\boldsymbol{W^a}\prod_a\delta\left(\sum_{ij}W_i^a C_{ij}W_j^* - NR\right)$$
$$\prod_a\delta\left(\sum_{ij}W_i^a C_{ij}W_j^a - Nq_0\right)\prod_{a<b}\delta\left(\sum_{ij}W_i^a C_{ij}W_j^b - Nq\right). \tag{A.4}$$

For notational simplicity here we have dropped the '0' from the correlation matrix $\boldsymbol{C}^{\setminus 0}$. F_0 can be computed following our derivation in [19]: we introduce the orthogonal matrix $\boldsymbol{U}$ that diagonalizes $\boldsymbol{C} = \boldsymbol{U}^\top\Lambda\boldsymbol{U}$,

$$Z_{\mathrm{coup}} = \int\prod_a \mathrm{d}\boldsymbol{W^a}\prod_a\delta(\sum_{ijk}U_{ij}W_j^a\Lambda_i U_{ik}W_k^* - NR)$$
$$\prod_a\delta(\sum_{ijk}U_{ij}W_j^a\Lambda_i U_{ik}W_k^a - Nq_0)\prod_{a<b}\delta(\sum_{ijk}U_{ij}W_j^a\Lambda_i U_{ik}W_k^b - Nq) \tag{A.5}$$

J. Stat. Mech. (2017) 063406

and transform the student coupling vector into new variables $\boldsymbol{U}^\top \boldsymbol{W}^a \to \boldsymbol{W}^a$, which we give the same name. We then express the delta functions as integrals over the auxiliary parameters $\{\hat{R}, \hat{q}, \hat{q}_0\}$. The integration gives

$$F_0(R,q,q_0) = \underset{\hat{R},\hat{q},\hat{q_0}}{\text{extr}} \frac{1}{\nu}\left\{ \mathrm{i}\hat{q}_0 q_0 + \mathrm{i}\hat{R}R - \frac{i}{2}\hat{q}q + \frac{1}{2N}\sum_i \frac{\hat{q} + \mathrm{i}\hat{R}^2(\sum_j U_{ij}W_j^*)^2\Lambda_i}{2\hat{q}_0 - \hat{q}} \right.$$
$$\left. + \frac{1}{2N}\sum_i \ln\left[\Lambda_i(\mathrm{i}\hat{q} - 2\mathrm{i}\hat{q}_0)\right] - \frac{1}{2}\ln(2\pi) \right\} \tag{A.6}$$

and the extremum over the conjugate order parameters yields

$$F_0(R,q,q_0) = \frac{1}{2\nu}\left[\frac{q_0 - R^2/V}{q - q_0} - \ln(q_0 - q) + \frac{1}{N}Tr\ln\overline{\boldsymbol{C}}\right], \tag{A.7}$$

where V, representing the cavity variance of the teacher field h_*, was introduced in (19). The second term of (A.2) contains the integration over the cavity fields h_* and $\{h_a\}_{a=1}^n$:

$$F_1(R,q,q_0) = -\lim_{n\to 0} N^{-1}\nu^{-1}\frac{\partial}{\partial n}\ln\left\{ 2\int \mathrm{d}h_* \prod_{a=1}^n \mathrm{d}h_a \frac{1}{Z_0}\exp\left[\beta h_*\right] \right.$$
$$\left. \exp\left[-\nu\sum_{a=1}^n \Phi(h_a)\right]\ p_{\text{cav}}(h_*, h_1, \dots h_n)\right\}^{\alpha N}, \tag{A.8}$$

where we applied the change of variables $\sigma_0 h_* \to h_*$ and $\sigma_0 h_a \to h_a$. The integration gives

$$F_1(R,q,q_0) = -\frac{\alpha}{\nu}\int \frac{\mathrm{d}v}{\sqrt{2\pi q}}\,\mathrm{e}^{-\frac{(v-\beta R)^2}{2q}}\ \ln\int \frac{\mathrm{d}y}{\sqrt{2\pi(q_0-q)}}\,\mathrm{e}^{-\frac{(y-v)^2}{2(q_0-q)}}\mathrm{e}^{-\nu\Phi(y)}. \tag{A.9}$$

Hence the free energy (A.2) becomes

$$F(R,q,q_0) = -\frac{1}{\nu}\left\{\frac{1}{2}\frac{q_0 - R^2/V}{q_0 - q} + \frac{1}{2}\ln(q_0 - q) - \frac{1}{2N}Tr\ln\overline{\boldsymbol{C}}\right.$$
$$\left. + \alpha\int \mathrm{d}v\ G_{\beta R,q}(v)\ \ln\int \mathrm{d}y\ G_{v,q_0-q}(y)\mathrm{e}^{-\nu\Phi(y)}\right\}, \tag{A.10}$$

where $G_{\mu,\omega}(v)$ denotes a scalar Gaussian density with mean μ and standard deviation ω.

Appendix B. Saddle point equations for the order parameters

We rewrite (20) as

$$F = -\underset{q,R,x}{\text{extr}}\left\{\frac{1}{2}\frac{q - R^2/V}{x} + \alpha\int \mathcal{D}v\ \max_y\left[-\frac{(y - \sqrt{q}v - \beta R)^2}{2x} - \Phi(y)\right]\right\}, \tag{B.1}$$

where $\mathcal{D}v = \mathrm{e}^{-v^2/2}/\sqrt{2\pi}$. The extremum over the order parameters gives the following set of equations:

J. Stat. Mech. (2017) 063406

$$
\begin{aligned}
0 &= \frac{1}{x} - \frac{\alpha}{\sqrt{q}} \int \mathcal{D}v \; v \;\; \frac{\partial \Phi(y)}{\partial y}\bigg|_{y=\hat{y}} \\
0 &= -\frac{R}{Vx} - \alpha\beta \int \mathcal{D}v \;\; \frac{\partial \Phi(y)}{\partial y}\bigg|_{y=\hat{y}} \\
0 &= -\frac{1}{x^2}\left(q - \frac{R^2}{V}\right) + \alpha \int \mathcal{D}v \;\; \left(\frac{\partial \Phi(y)}{\partial y}\bigg|_{y=\hat{y}}\right)^2 ,
\end{aligned}
\tag{B.2}
$$

where

$$
\hat{y} = \arg\max_{y} \left[-\frac{(y - \sqrt{q}v - \beta R)^2}{2x} - \Phi(y) \right] . \tag{B.3}
$$

If we consider the pseudo-likelihood algorithm with $\Phi(y) = -\beta y + \ln 2\cosh(\beta y)$ (see the the definition of Φ (13) and the cost function (5)) we obtain the following equations for the order parameters:

$$
0 = \frac{1}{x} + \frac{\alpha\beta}{\sqrt{q}} \int \mathcal{D}v \; v \;\; (1 - \tanh(\beta\hat{y})) \tag{B.4a}
$$

$$
0 = -\frac{R}{Vx} + \alpha\beta^2 \int \mathcal{D}v \, (1 - \tanh(\beta\hat{y})) \tag{B.4b}
$$

$$
0 = -\frac{1}{x^2}\left(q - \frac{R^2}{V}\right) + \alpha\beta^2 \int \mathcal{D}v \;\; (1 - \tanh(\beta\hat{y}))^2 , \tag{B.4c}
$$

where $\hat{y}$ is defined by

$$
\hat{y} = \sqrt{q}v + \beta R + \beta x(1 - \tanh(\beta\hat{y})). \tag{B.5}
$$

Appendix C. Relation between order parameters

We introduce the auxiliary variables $\{\eta_1, \eta_2\}$ in the free energy $F = F_0 + F_1$ as follows:

$$
\begin{aligned}
F_0(R, q, q_0, \eta_1, \eta_2) = &- \lim_{n\to 0} \nu^{-1} N^{-1} \frac{\partial}{\partial n} \ln \int \prod_a \mathrm{d}\boldsymbol{W}^{\boldsymbol{a}} \; \mathrm{d}\hat{q}_0 \; \mathrm{d}\hat{R} \; \mathrm{d}\hat{q} \\
&\prod_a \mathrm{e}^{\mathrm{i}\hat{R}(\sum_{ijk} U_{ij} W_j^a (\Lambda_i + \eta_1) U_{ik} W_k^* - NR)} \prod_a \mathrm{e}^{\mathrm{i}\hat{q}_0(\sum_{ijk} U_{ij} W_j^a \Lambda_i U_{ik} W_k^a - Nq_0)} \\
&\prod_{a<b} \mathrm{e}^{\mathrm{i}\hat{q}(\sum_{ijk} U_{ij} W_j^a (\Lambda_i + \eta_2) U_{ik} W_k^b - Nq)} .
\end{aligned}
\tag{C.1}
$$

The integration gives

J. Stat. Mech. (2017) 063406

$$F_0(R,q,q_0,\eta_1,\eta_2) = \underset{\hat{R},\hat{q},\hat{q_0}}{\text{extr}} \frac{1}{\nu}\left\{ \mathrm{i}\hat{q_0}q_0 + \mathrm{i}\hat{R}R - \frac{i}{2}\hat{q}q \right.$$
$$+\frac{1}{2N}\sum_i \frac{\hat{q}(\Lambda_i+\eta_2)+\mathrm{i}\hat{R}^2(\sum_j U_{ij}W_j^*)^2(\Lambda_i+\eta_1)^2}{2\hat{q_0}-\hat{q}(\Lambda_i+\eta_2)}$$
$$\left.+\frac{1}{2N}\sum_i \ln\left[\mathrm{i}\hat{q}(\Lambda_i+\eta_2)-2\mathrm{i}\hat{q_0}\right]-\frac{1}{2}\ln(2\pi)\right\}. \tag{C.2}$$

From (C.1) it is easy to see that the parameters $\{\rho, Q\}$ can be derived by derivatives of the free energy:

$$\rho = N^{-1}\overline{\boldsymbol{W}^*\cdot\langle\boldsymbol{W}\rangle} = \frac{\nu}{\mathrm{i}\hat{R}}\frac{\partial F_0}{\partial\eta_1},$$
$$Q = N^{-1}\overline{\langle\boldsymbol{W}\rangle^2} = -\frac{2\nu}{\mathrm{i}\hat{q}}\frac{\partial F_0}{\partial\eta_2} \tag{C.3}$$

in the limit $\{\eta_1\to 0,\eta_2\to 0\}$, where $\hat{R}$ and $\hat{q}$ are the values extremizing (C.2) in the limit $\{\eta_1\to 0,\eta_2\to 0\}$:

$$\hat{q} = \mathrm{i}\frac{R^2-Vq}{V(q_0-q)^2},$$
$$\hat{R} = \mathrm{i}\frac{R}{V(q-q_0)}, \tag{C.4}$$
$$\hat{q_0} = \mathrm{i}\frac{R^2+V(q_0-2q)}{2V(q_0-q)^2}.$$

From (C.2)–(C.4) we recover (21).

Appendix D. Replica result for quadratic cost functions

If the cost function has the simple quadratic form (27), computing the maximum and the integrals in (20) can be done analytically and the free energy is

$$F = -\frac{q-R^2/V}{2x}+\alpha\frac{q+\beta R(\beta R-2\eta)-\eta^2 x}{2(1+x)}, \tag{D.1}$$

where the order parameters obey to the following saddle point equations:

$$q = \frac{\eta^2(1+\alpha\beta^2V)}{(\alpha-1)(1+\beta^2V)},$$
$$R = \frac{\eta\beta V}{1+\beta^2V}, \tag{D.2}$$
$$x = \frac{1}{(\alpha-1)}.$$

With this result, the reconstruction error (22) for the linear estimator becomes (28). Moreover, we can compute the parameter Δ defined in (33) as follows. If the 'training

energy' per degree of freedom N becomes self-averaging we can use the relation (see also (11))

$$E(\boldsymbol{W}^{\mathrm{ML}}) = -\lim_{\nu\to\infty} \nu^{-1} \ln Z \tag{D.3}$$

to explicitly evaluate the minimum training energy as

$$E(\boldsymbol{W}^{\mathrm{ML}}) = \frac{N}{\alpha} F(\boldsymbol{W}^{\mathrm{ML}}) = -\frac{N\eta^2}{2\alpha}\frac{(1+\alpha\beta^2 V)}{(1+\beta^2 V)}, \tag{D.4}$$

where the second equality follows from (D.1) with the order parameters fixed to their saddle point values (D.2). From (32) and (D.4), one finds (34).

Appendix E. Asymptotics from the replica approach

In the large α limit, we know that the parameter x gets small and the parameters q and R both converge to V. For the pseudo-likelihood estimator we find the following relation, starting from (B.4*b*) in the limit $R \to V$ and (B.4*c*):

$$q - \frac{R^2}{V} = \frac{1}{\alpha\beta^2} \frac{\int \mathcal{D}v \ \left(1-\tanh(\beta\hat{y})\right)^2}{\left[\int \mathcal{D}v \left(1-\tanh(\beta\hat{y})\right)\right]^2} \tag{E.1}$$

where $\hat{y}$ is given by (B.5), that in the limit of small x becomes $\hat{y} \simeq \sqrt{q}v + \beta R$. Via a change of variable, we find the following result, in the limit $R \to V, q \to V$:

$$\begin{aligned} q - \frac{R^2}{V} &= \frac{1}{\alpha\beta^2} \frac{\int \mathrm{d}v \ G_{\beta V,V}(v) \left(1-\tanh(\beta v)\right)^2}{\left[\int \mathrm{d}v \ G_{\beta V,V}(v) \left(1-\tanh(\beta v)\right)\right]^2} \\ &\simeq \frac{1}{\alpha\beta^2} \frac{1}{\int \mathrm{d}v \ G_{\beta V,V}(v) \left(1-\tanh^2(\beta v)\right)}, \end{aligned} \tag{E.2}$$

where in the last equality we exploited the relation $\int \mathrm{d}v \ G_{\beta V,V}(v) \tanh(\beta v) = \int \mathrm{d}v \ G_{\beta V,V}(v) \tanh^2(\beta v)$. Hence, the error (22) for large α scales as

$$\varepsilon \simeq \left(q - \frac{R^2}{V}\right) \frac{1}{N}\mathrm{Tr}\overline{C^{-1}} \simeq \frac{1}{\alpha\beta^2} \frac{1}{\int \mathrm{d}v \ G_{\beta V,V}(v) \left(1-\tanh^2(\beta v)\right)} \frac{1}{N}\mathrm{Tr}\overline{C^{-1}}. \tag{E.3}$$

Appendix F. Asymptotic error for pseudo-likelihood estimator

We show that $U_{ij} = -B_{ij}$ for the pseudo-likelihood case assuming that the model is matched to the true data generating distribution. We start from the relations

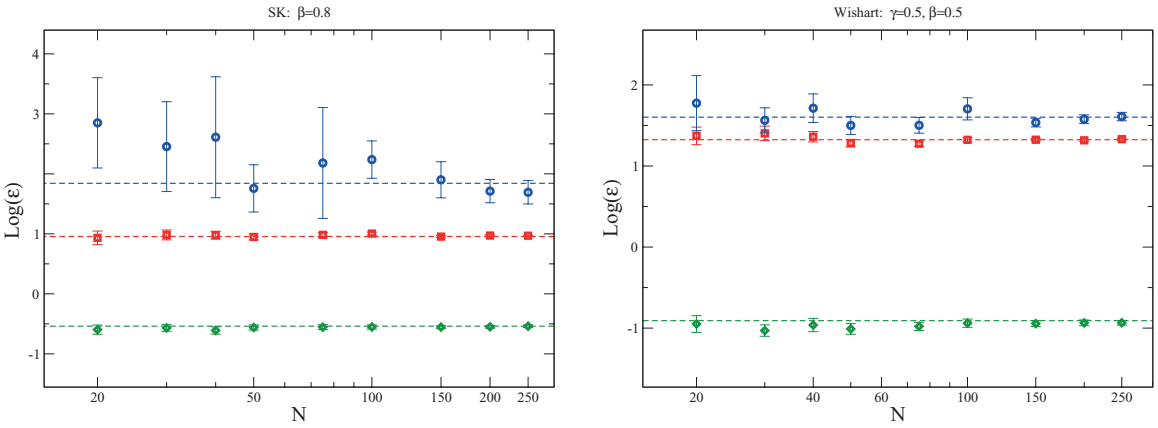

Figure G1. Mean squared error of the couplings inferred by using the pseudo-likelihood algorithm (blue dots) the optimal local algorithm (green dots) and the MF-ML algorithm (red dots) as a function of the system size N, for fixed $\alpha = 5$. The dotted lines represent replica results. Two different systems are considered: SK model at $\beta = 0.8$ (left) and Wishart model with $\gamma = 0.5$ at $\beta = 0.5$ (right). The results are averaged over 5 realizations of the network and 20 different datasets generated from each network. Error bars represent standard deviations of the means.

$$U_{ij} = \sum_{\boldsymbol{\sigma}_{\setminus 0}} P(\boldsymbol{\sigma}_{\setminus 0}) \sum_{\sigma_0} P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}) \partial_i \partial_j \ln P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}) \tag{F.1}$$

and

$$B_{ij} = \sum_{\boldsymbol{\sigma}_{\setminus 0}} P(\boldsymbol{\sigma}_{\setminus 0}) \sum_{\sigma_0} P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}) \partial_i \ln P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}) \partial_j \ln P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}). \tag{F.2}$$

We next perform the inner expectation over $P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0})$. The result follows from

$$\partial_i \partial_j \ln P = -\partial_i \ln P \, \partial_j \ln P + \frac{1}{P} \partial_i \partial_j P$$

and the fact that, by normalisation of P, one gets $\sum_\sigma \partial_i \partial_j P(\sigma|\boldsymbol{\sigma}_{\setminus 0}) = 0$.

Appendix G. Error dependence on the system size

In figure 1 we showed that the reconstruction error in systems with $N = 100$ spins well agrees with the replica result, which is valid in the thermodynamic limit. However, it is relevant for applications to show an example of how the system size can affect the reconstruction error. In figure G1 we show results obtained by fixing α and varying N. First of all we notice that finite size effects are much stronger for the the pseudo-likelihood algorithm than for the other two methods. Moreover, while the optimal local estimator always seems to outperform the other two methods, the performance difference between MF-ML and pseudo-likelihood algorithms depends more strongly on the system parameters (α, β, teacher coupling distribution), if N is small. For instance, we see in figure G1 that, for systems of $N = 20,30$ spins with couplings drawn from the Wishart distribution, the error of the MF-ML and pseudo-likelihood algorithms are compatible.

J. Stat. Mech. (2017) 063406

References

[1] Schneidman E, Berry M J, Segev R and Bialek W 2006 Weak pairwise correlations imply strongly correlated network states in a neural population *Nature* **440** 1007–12

[2] Roudi Y, Tyrcha J and Hertz J 2009 Ising model for neural data: model quality and approximate methods for extracting functional connectivity *Phys. Rev.* E **79** 051915

[3] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 Identification of direct residue contacts in protein–protein interaction by message passing *Proc. Natl Acad. Sci.* **106** 67–72

[4] Lezon T R, Banavar J R, Cieplak M, Maritan A and Fedoroff N V 2006 Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns *Proc. Natl Acad. Sci.* **103** 19033–8

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

[6] Kappen H J and de Borja Rodríguez F 1998 Efficient learning in Boltzmann machines using linear response theory *Neural Comput.* **10** 1137–56

[7] Tanaka T 1998 Mean-field theory of Boltzmann machine learning *Phys. Rev.* E **58** 2302

[8] Roudi Y, Aurell E and Hertz J 2009 Statistical physics of pairwise probability models *Frontiers Comput. Neurosci.* **3** 22

[9] Sessak V and Monasson R 2009 Small-correlation expansions for the inverse Ising problem *J. Phys. A: Math. Theor.* **42** 055001

[10] Besag J 1974 Spatial interaction and the statistical analysis of lattice systems *J. R. Stat. Soc. Ser.* B **36** 192–236

[11] Aurell E and Ekeberg M 2012 Inverse Ising inference using all the data *Phys. Rev. Lett.* **108** 090201

[12] Decelle A and Ricci-Tersenghi F 2014 Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models *Phys. Rev. Lett.* **112** 070603

[13] Mozeika A, Dikmen O and Piili J 2014 Consistent inference of a general model using the pseudolikelihood method *Phys. Rev.* E **90** 010101

[14] Tyagi P, Marruzzo A, Pagnani A, Antenucci F and Leuzzi L 2016 Regularization and decimation pseudolikelihood approaches to statistical inference in $x\ y$ spin models *Phys. Rev.* B **94** 024203

[15] Opper M and Kinzel W 1996 Statistical mechanics of generalization *Models of Neural Networks* III (Berlin: Springer) pp 151–209

[16] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)

[17] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: an Introduction* vol 111 (Oxford: Clarendon)

[18] Kuva S M, Kinouchi O and Caticha N 1998 Learning a spin glass: Determining Hamiltonians from metastable states *Physica A: Statistical Mechanics and its Applications* **257** 28–35

[19] Bachschmid-Romano L and Opper M 2015 Learning of couplings for random asymmetric kinetic ising models revisited: random correlation matrices and learning curves *J. Stat. Mech.* P09016

[20] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory, Beyond: an Introduction to the Replica Method and its Applications* vol 9 (Singapore: World Scientific)

[21] Schervish M J 2012 *Theory of Statistics* (Berlin: Springer)

[22] Cocco S and Monasson R 2011 Adaptive cluster expansion for inferring Boltzmann machines with noisy data *Phys. Rev. Lett.* **106** 090601

[23] Sohl-Dickstein J, Battaglino P B and DeWeese M R 2011 New method for parameter estimation in probabilistic models: minimum probability flow *Phys. Rev. Lett.* **107** 220601

[24] Ricci-Tersenghi F 2012 The bethe approximation for solving the inverse Ising problem: a comparison with other inference methods *J. Stat. Mech.* P08015

[25] Nguyen H C and Berg J 2012 Bethe–Peierls approximation and the inverse Ising problem *J. Stat. Mech.* P03004

[26] Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev.* A **45** 6056

[27] Sherrington D and Kirkpatrick S 1975 Solvable model of a spin-glass *Phys. Rev. Lett.* **35** 1792

[28] Bray A J and Moore M A 1980 Metastable states in spin glasses *J. Phys. C: Solid State Phys.* **13** L469

[29] Kinouchi O and Caticha N 1992 Optimal generalization in perceptions *J. Phys. A: Math. Gen.* **25** 6243

[30] Advani M and Ganguli S 2016 Statistical mechanics of optimal convex inference in high dimensions *Phys. Rev.* X **6** 031034

J. Stat. Mech. (2017) 063406

[31] Berg J 2016 Statistical mechanics of the inverse Ising problem and the optimal objective function (arXiv: 1611.04281)
[32] Opper M and Winther O 2001 Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling *Phys. Rev.* E **64** 056131
[33] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl Acad. Sci.* **79** 2554–8
[34] Parisi G and Potters M 1995 Mean-field equations for spin models with orthogonal interaction matrices *J. Phys. A: Math. Gen.* **28** 5267
[35] Opper M, Cakmak B and Winther O 2016 A theory of solving tap equations for ising models with general invariant random matrices *J. Phys. A: Math. Theor.* **49** 114002
[36] Abadir K M and Magnus J R 2005 *Matrix Algebra* (Cambridge: Cambridge University Press)
[37] Amit D J, Gutfreund H and Sompolinsky H 1985 Storing infinite numbers of patterns in a spin-glass model of neural networks *Phys. Rev. Lett.* **55** 1530
[38] Mézard M and Parisi G 2001 The Bethe lattice spin glass revisited *Eur. Phys. J.* B **20** 217–33

J. Stat. Mech. (2017) 063406

5.3 Further results

In Paper 3, we saw that both the result for the estimation error and the explicit form of the optimal local estimator depend on two order parameters: the the squared length of the teacher coupling vector, Y; the cavity variance of the teacher field, V. In this paragraph, we show that these parameters can be estimated from data, if the inverse temperature β is known.

The result for V is obtained from the empirical correlation matrix, according to equations (Paper 3, 33 - 34); Figure 5.1 shows a good agreement between the estimated and the true parameters for the SK model.

The length of the teacher vector Y can be computed via a linear estimator minimizing the cost function (Paper 3, 24). By combining the results for the order parameters (Paper 3, 21) and (Paper 3, D2), one can express Y in terms of the length of the student vector Q:

$$Y = \frac{(1+\beta^2 V)^2 (Q(\alpha-1)-\eta^2)}{(\alpha-1)\beta^2\eta^2}, \tag{5.1}$$

where we also used (Paper 3, 53) for the trace of the inverse correlation matrix. Now, if we set the free parameter η to 1, we can estimate the couplings $\boldsymbol{W}^{(\eta=1)}$ using (Paper 3, 26); hence, we can compute the length of the inferred coupling vector $Q_{\text{exp}}^{(\eta=1)} = N^{-1}(\boldsymbol{W}^{(\eta=1)})^2$. The parameter Y is then obtained from (5.1) by setting $\eta = 1$ and $Q = Q_{\text{exp}}^{(\eta=1)}$, assuming that we know β and that we already computed V from data. Figure 5.1 shows a good agreement between the true length of the teacher couplings and the one estimated from data for a SK model. Moreover, Tabel 5.1 shows that the estimation error predicted using the parameters estimated from data is in good agreement with the one predicted with the true value of the parameters.

We conclude that the optimal local estimator can be fully constructed from the data. Moreover, notice that both the explicit optimal local estimator and the explicit mean field ML estimator obey to the same equation (Paper 3, 26), where the difference is given by the parameter η. From the equations for η_{opt} (Paper 3, 29) and from $\eta_{MF} = \phi_0/\beta$, where ϕ_0 is given in (Paper 3, 59), it is evident that the parameter η is a self-averaging quantity; hence, it does not depend on the central spin σ_0. It follows that the couplings estimated using the two methods will be proportional to each other. Since the mean field ML estimator is symmetric, also the optimal local quadratic estimator must be symmetric.

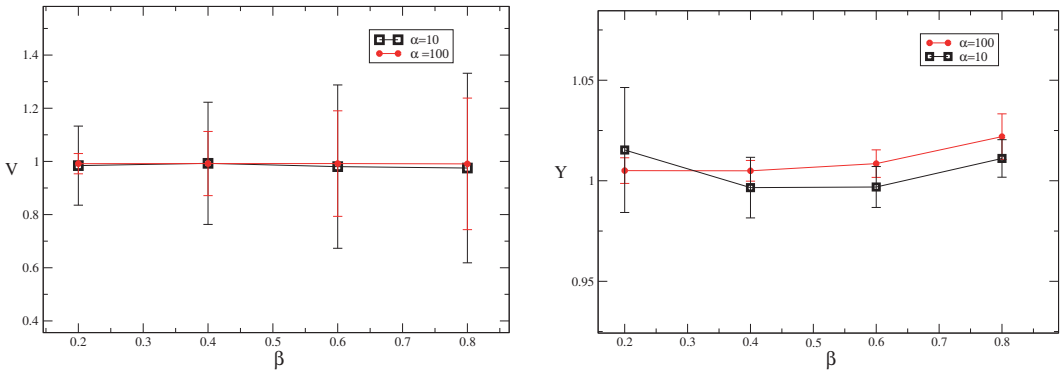

Figure 5.1: The parameter V (left) and Y (right) are estimated from data and plotted as a function of the inverse temperature. We consider a SK model of $N = 100$ spins, where the true value of the parameters is $Y = V = 1$. Results are averaged over 10 instances of the network.

β	ε_{ture}	ε_{est}
0.2	0.751	0.75 ± 0.02
0.8	0.318	0.315 ± 0.003

Table 5.1: The error ε_{ture} (Paper 3, 30) predicted by the replica calculation with the true value of the parameters $Y = V = 1$, is compared to the error obtained using the parameters V and Y estimated from data (see Figure 5.1). For ε_{ture}, we used the results of Paper 3, Figure 1.

5.4 Conclusions

In this chapter, we have discussed the inverse Ising problem and studied the average error of estimating the couplings based on local convex cost functions. We worked in a teacher-student scenario, combining techniques of statistical mechanics: to deal with the intractable distribution of the data we used cavity arguments, which are valid for dense and weakly interacting systems and become exact in the in the thermodynamic limit for certain statistical ensembles of network couplings; to performe the quenched averages we used the replica symmetric formalism, which is correct when the learning algorithms are based on convex cost functions.

The teacher network is fixed, and the analytic result for the error depends on two order parameters: the length of the teacher coupling vector Y, the

variance of the cavity field V, and the inverse of the spin correlation matrix. We expressed the latter quantity in terms of V and showed that both Y and V can be estimated from the data.

The optimal local cost function is of quadratic form, and the estimator associated to it is simply computed by inverting the empirical correlation matrix. This optimal local estimator is symmetric and its parameters can be estimated from data. Moreover, a mean-field approximation of the maximum likelihood estimator is also symmetric and asymptotically optimal, and outperforms the widely used pseudo-likelihood estimator (notice that the difference between the asymptotic error of those two methods gets smaller for higher temperatures). We found a very good agreement between theory and simulations for the SK model and a model where the matrix of couplings has a Gaussian distribution, for high enough temperatures (paramagnetic phase).

It would be interesting to carry out such a comparison in the small T region. At low temperatures, the disordered Ising model can undergo a phase transition to a glassy phase, where the free energy has multiple valleys separated by high barriers. A first limitation when performing inference is that data might not come from a uniform sampling over the equilibrium configuration but rather from a subset of possible states, as the system is non-ergodic. Another problem is that, as already mentioned, our simple cavity arguments for treating the Ising distribution of the spin configurations would no longer apply and would have a more complex distribution [MP01].

In addition, reconstruction based on mean-field estimators (as well as other estimators based on self-consistent equations for the magnetisation) will typically fail at low temperatures. Such equations indeed provide correct solutions for the magnetisation of single thermodynamic states, while the data could come from many thermodynamics states. A way to circumvent this problem is proposed in [NB12b], where the authors use a clustering algorithm to identify clusters of data generated from the same thermodynamic state. Then, magnetisations and correlations are computed separately for each cluster, in which the self-consistent equations for the magnetisations are still valid. Note that, if couplings are reconstructed from single thermodynamics states, they are systematically underestimated. Instead, by collecting the equations of different thermodynamic states and jointly solving them (using the Moore-Penrose pseudo-inverse methods), the quality of the reconstruction is significantly improved.

As a future direction, it would be also interesting to add a regularising term to the cost function, encoding prior information on the student-coupling distribution. This would help to avoid the divergence of the error in the small α region; moreover, it would allow us to study reconstruction in sparse networks.

A preliminary analysis shows that, in the presence of a regularising term, the replica calculation for the free energy becomes more complicated. The eigenvalues of the correlation matrix and the order parameters of the cavity fields cannot be easily disentangled as in Paper 3, and further investigation is needed.

Appendix

5.A The likelihood and pseudo-likelihood functions

We consider the problem of inferring external local fields and pairwise couplings J_{ij} in an equilibrium Ising model (paper 4, 1), based on a set of M independent configurations $\{\boldsymbol{\sigma}^k\}_{k=1}^M$. A widely used estimator is maximum likelihood, which consists in maximizing the log-likelihood function with respect to the parameters:

$$\mathcal{L}(\boldsymbol{H}, \boldsymbol{J}) = \beta \sum_{i<j} J_{ij} \langle \sigma_i \sigma_j \rangle_D + \beta \sum_i h_i \langle \sigma_i \rangle_D - \ln Z(\boldsymbol{h}, \boldsymbol{J}), \tag{5.2}$$

where the brakets $\langle \dots \rangle_D$ define averaging over the data. The parameters can be inferred in two ways. One can find the maximum of (5.2) by gradient ascent, which implies computing the partition function at every step of the iteration. Alternatively, one observes that the log-likelihood (5.2) is maximized when the data-averaged magnetizations and correlations match -respectively- the first and second moments of the distribution (paper 4, 1):

$$\langle \sigma_i \rangle = \langle \sigma_i \rangle_D \tag{5.3}$$
$$\langle \sigma_i \sigma_j \rangle = \langle \sigma_i \sigma_j \rangle_D, \tag{5.4}$$

These maximum likelihood conditions are reached by a gradient descent algorithm denoted as Boltzmann machine learning [AHS85], in which the parameters are updated according to the following rule:

$$h_i^{new} = h_i^{old} + \gamma(\langle \sigma_i \rangle_D - \langle \sigma_i \rangle) \tag{5.5}$$
$$J_{ij}^{new} = J_{ij}^{old} + \gamma(\langle \sigma_i \sigma_j \rangle_D - \langle \sigma_i \sigma_j \rangle). \tag{5.6}$$

where γ is the learning rate of the algorithm. Both the computation of the partition function in the first case, and the computation of the thermal averages in the second case, require the sum over 2^N terms, which should be performed at every step of the iteration. Even resorting to Monte Carlo sampling techniques, the exact inference remains intractable.

A much faster method to estimate the parameters is the pseudolikelihood method, where the log-likelihood function to be maximized is approximated by a simpler function. One isolates one 'central' spin σ_0 and infers the field H_0 and coupling vector $J_0 = \{J_{j0}\}_{j\neq 0}$ starting from the conditional probability

$$P_{\{h_0,J_0\}}(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}) = \frac{e^{\beta\sigma_0[\sum_{j\neq 0} J_{j0}\sigma_j + H_0]}}{2\cosh\beta[\sum_{j\neq 0} J_{j0}\sigma_j + H_0]}, \tag{5.7}$$

where, $\boldsymbol{\sigma}_{\setminus 0}$ denotes the set of all the spins but σ_0. The pseudolikelihood estimator of H_0 and J_0 minimizes the local cost function

$$f_0(H_0, J_0) = -\frac{1}{M}\sum_{k=1}^{M} \ln P_{\{h_0,J_0\}}(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}), \tag{5.8}$$

which does not require the computation of the partition function and can be estimated in polynomial time in the system size and number of samples. A derivation of the method in the context of the inverse Ising problem can be found in [NZB17b], while a proof that the pseudolikelihood estimator is consistent in [MDP14]. Note that in general the estimation of $J_{ij}^{(i)}$, where we fixed the spin σ_i, is different from the estimation $J_{ij}^{(j)}$ where we fixed σ_j. Typically at the end of the inference procedure one symmetrizes the estimator by considering $J_{ij} = (J_{ij}^{(i)} + J_{ij}^{(j)})/2$; alternatively, one could minimize the sum $\sum_i f_i$ while imposing $J_{ij}^{(i)} = J_{ij}^{(j)}$.

6 Learning and inference in presence of hidden units

6.1 Introduction

We have introduced the kinetic Ising model as a benchmark model to study network reconstruction from large-scale data. Data collection techniques are rapidly improving in many scientific fields, especially biology, allowing us to detect the activity of many system components at the same time. For example, we can now have access to neural recordings from populations of hundreds of thousands of neurons [ODB+14, Nic08], and the largest public repository for high-throughput gene expression data [BTW+08] today comprises 300000 samples for over 500 organisms submitted by laboratories from around the world. Still, the number of recorded units remains small if compared to the total number of units typically involved in carrying out a biological function. Hence, variables whose activity is recorded will also interact with variables not directly detectable, hereafter addressed as hidden variables. The existence of such variables can influence the dynamics of the observed ones, their effect being non-negligible. For instance, we could incorrectly identify direct couplings between observed nodes when they share a common input from unrecorded units. This problem is receiving considerable attention within the statistical mechanics community, and it is being addressed from two complementary perspectives [BDR17].

On one side, recent works [MMR13, HM15] considered the problem of getting as much information as possible from data in the highly under-sampled regime, namely, when the number of observed samples is much smaller than the dimensionality of the model. They propose a method called critical variable selection that allows to detect the degrees of freedom that are most relevant to the function of the system, without knowing much about the function, itself. Working in an information theoretical setting, the authors propose a measure of the amount of information encoded in a sample to define the notion of the most informative samples. This helps to analyse large datasets in the under-sampled regime, as it gives a method to find insightful structure in the data and avoid fitting noise, thus providing an effective alternative to current

dimensionality reduction schemes [GFM16].

A second line of research [TH13,DR13,Hua15,BHTR15,RT15,DB16] considers models in which the presence of hidden variables is introduced explicitly, to study in more detail how hidden units can influence the observed ones. Probabilistic models with hidden variables (typically denoted as latent variables) are indeed widely used in machine learning, as they allow complex distributions over observed variables to be written as more tractable joint distributions over the expanded space of observed and latent variables [Bis06].

A widely used method to estimate parameters in latent variable models is the expectation maximisation algorithm (EM) [MK07], which determines the parameters that maximise the expected log-likelihood under the posterior distribution of latent variables. It is a two-step iterative algorithm. In the E step, one fixes the model parameters and computes the conditional expectation of the log-likelihood given the observations. In the M step, one updates the parameters to the values that maximise the expected log-likelihood computed in the previous step. This iteration is guaranteed to converge to the maximum likelihood solution (or, in the case of multiple local maxima, to one of the local maxima of the likelihood) under mild regularity conditions [Wu83]. However, the computation of posterior averages in the E step is often intractable for large systems, and various approximate techniques have been proposed ([TH13, DR13, BHTR15]), which are based on approximations to the posterior moments of hidden spins. Hence, in this scenario, we find it interesting to investigate the theoretically optimal performance for predicting hidden spins.

As a paradigmatic model, we consider an extension of the kinetic Ising model, composed of two sets of variables: the observed ones, whose state can be measured at each time step, and the hidden ones, whose trajectory is unknown. There are couplings between the observed units, between the hidden ones, from hidden to observed, and from observed to hidden units. In a Bayesian setting, if the probabilistic model that generated the data is known, the sign of the posterior expectation of a hidden spin gives us the best possible prediction of that spin state. Hence, our goal is to determine the average error of the Bayes optimal prediction of hidden variables, where the averages are over the set of model parameters (the couplings), over the data (observed spin trajectories), and over the configuration of hidden spins. We can interpret those averages by drawing an analogy with the statistical physics of disordered systems. The model parameters and the data represent quenched variables, which change on a time scale much larger than the time scale characterising the changes in the hidden spins states, which play the role of thermal variables. If we consider the couplings to be Gaussian random variables with variance $1/N$, where N is the system size, and work in the thermodynamic limit of $N \to \infty$, we can

compute those averages exactly using the replica method of statistical physics. We will consider the replica symmetric ansatz and show that our analytical results agree well with simulations, both for posterior averages of the hidden spins and for the average prediction error.

After discussing the average case scenario, in the second part of this chapter, we will present a novel approximate technique to compute marginal moments of hidden spins for a single instance of the network. Our approach is based on the extended Plefka expansion, the weak coupling expansion we developed in chapter 3 to derive a mean-field description of the dynamics for the kinetic Ising model. In this chapter, we will show how to extend it to the case of a network with hidden units. The same approach was developed in parallel and proved to be successful [BS16] for the simpler case of stochastic linear dynamics of continuous degrees of freedom, in a system composed of a sub-network of observed nodes embedded into a larger bulk of unknown (i.e., hidden) nodes.

As a final point, we will discuss the applicability of the Plefka expansion in the E step of an EM algorithm aimed at finding the maximum likelihood estimation of the couplings. The problem of network reconstruction in a kinetic Ising model with hidden spins was recently addressed in [TH13, DR13, BHTR15]. The authors of [BHTR15] study a model in which hidden factors are conditionally independent of each other given the observed ones (i.e., there are no couplings among the hidden units) and use a message passing algorithm to learn the couplings. In [TH13] and [DR13] , the learning problem is addressed with algorithms of the EM type, which prove to achieve relatively good performances on systems were the hidden units were not connected to each others (or were present in the generative model and ignored during learning). Only in [DR13] , connections among hidden units are considered; in this case, the algorithm converges only when the number of hidden units is relatively small, namely, 10% of the total units. Moreover, reconstruction is satisfactory for sparse networks, but not as good for dense ones. Hence, those works remarkably contributed in addressing the problem, but there are still many questions left to be resolved. We will further discuss the challenges linked to an expectation maximisation approach and outline future directions.

6.2 Paper 4.

Author's contribution: I performed the analytical and numerical calculations, prepared the figures and contributed to writing the paper.

Inferring hidden states in a random kinetic Ising model: replica analysis

J. Stat. Mech. (2014) P06013

(http://iopscience.iop.org/1742-5468/2014/6/P06013)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Inferring hidden states in a random kinetic Ising model: replica analysis

Ludovica Bachschmid-Romano and Manfred Opper

Department of Artificial Intelligence, Technische Universität Berlin, Marchstraße 23, Berlin 10587, Germany
E-mail: ludovica.bachschmidromano@tu-berlin.de and manfred.opper@tu-berlin.de

Received 15 December 2013
Accepted for publication 24 April 2014
Published 18 June 2014

Online at stacks.iop.org/JSTAT/2014/P06013
doi:10.1088/1742-5468/2014/06/P06013

Abstract. We consider the problem of predicting the spin states in a kinetic Ising model when spin trajectories are observed for only a finite fraction of sites. In a Bayesian setting, where the probabilistic model of the spin dynamics is assumed to be known, the optimal prediction can be computed from the conditional (posterior) distribution of unobserved spins given the observed ones. Using the replica method, we compute the error of the Bayes optimal predictor for parallel discrete time dynamics in a fully connected spin system with non-symmetric random couplings. The results, exact in the thermodynamic limit, agree very well with simulations of finite spin systems.

Keywords: disordered systems (theory), statistical inference, learning theory, stochastic processes (theory)

J. Stat. Mech. (2014) P06013

1742-5468/14/P06013+15$33.00

Contents

1. Introduction 2
2. The model and Bayes optimal inference 3
3. Replica analysis 4
4. Distribution of local magnetization 9
5. Results 9
6. A comment on symmetric networks 12
7. Outlook 12
Acknowledgments 13
Appendix A. Self consistent solution for the two time order parameters 13
Appendix B. Boundary conditions 14
Appendix C. Forward-backward algorithm 14
References 15

1. Introduction

The problem of statistical inference in kinetic Ising models has recently attracted considerable interest in the statistical physics community, see e.g. [1–5]. These systems can be viewed as simple models of networks of spiking neurons and provide a prototype model for which a reconstruction of the network from dynamical data can be studied. Based on a temporal sequence of observed spin variables, a major goal is to estimate the couplings between sites. This task gets more complicated when at some sites the spin trajectories are not observed. Besides the problem of inferring the couplings it is then also interesting to predict the states of the non-observed spins when the couplings are known. In fact, an iterative solution to the maximum likelihood problem for estimating the couplings is the *Expectation Maximization* (EM) algorithm [6] which would iterate between estimating hidden spin states (given the last estimate of the couplings) and reestimating the couplings. Unfortunately, exact inference of hidden states is not tractable for large networks, but algorithms which are based on statistical physics approximations have recently been discussed [7, 8]. Hence, it will be interesting and important to study a scenario for which the theoretically optimal performance for predicting hidden spins can be computed exactly. In this paper, we will show that such a solution can be found in the thermodynamic limit of an infinitely large network when the couplings are random. Our approach will be based on the replica method of disordered systems which enables us to compute quenched averages over the random couplings for thermodynamic quantities of the model. These thermodynamic quantities

doi:10.1088/1742-5468/2014/06/P06013

J. Stat. Mech. (2014) P06013

are themselves functions of posterior averages (e.g. local magnetizations) of the hidden spins. The replica approach has been successfully applied in the past to a large variety of statistical learning problems for *static* network models (for a summary see [9–11]). We will restrict ourselves to a model where the couplings are mutually independent random variables, i.e. where no symmetry between in-and outgoing connections are assumed. For such type of models (without the observations) various exact solutions for the non-equilibrium dynamics have been computed, see e.g. [1, 5] and [12, 13] for soft spin models. From the point of view of equilibrium statistical physics the case of symmetric couplings might be interesting. Such a spin model would obey detailed balance and allow for a stationary Gibbs distribution. Unfortunately, for the Ising case, the exact computation of time dependent correlation functions which are necessary for our analysis seems not possible. On the other hand, from a point of view of neural modeling, the assumption of symmetric couplings is not realistic [1, 14], as synaptic connections in biological networks are known to be strongly asymmetric. Hence, we believe that our restriction to asymmetric couplings is justified both from a modeling and a computational perspective.

2. The model and Bayes optimal inference

We will consider a model with N Ising spins which are divided into two groups: a group of spins $s_i(t)$ at sites i = 1, ..., $N_{\rm obs} = \lambda\, N$ which are observed during a time interval of T time steps and a group of hidden, i.e. unobserved spins, denoted by $\sigma_a(t)$ at sites $a = 1, \ldots, N_{\rm hid} = (1-\lambda)N$. We assume parallel Markovian dynamics for the entire spin system, which is governed by the transition probability

$$P[\{s,\,\sigma\}(t+1)|\{s,\,\sigma\}(t)] = \prod_i \frac{\mathrm{e}^{s_i(t+1)g_i(t)}}{2\cosh[\mathrm{g_i}(t)]} \prod_a \frac{\mathrm{e}^{\sigma_a(t+1)g_a(t)}}{2\cosh[\mathrm{g_a}(t)]}\,, \tag{1}$$

where the fields are defined as

$$g_i(t) = \sum_j J_{ij}s_j(t) + \sum_b J_{ib}\sigma_b(t)\,,\; g_a(t) = \sum_j J_{aj}s_j(t) + \sum_b J_{ab}\sigma_b(t)\,, \tag{2}$$

in terms of the couplings J and $\{s,\,\sigma\}$ denotes all the possible spin vector configurations; when the time index is not specified we are considering the whole time series, $t = 0 \ldots T$. The total probability for a spin trajectory is given by

$$P(\{s,\,\sigma\}) = \frac{1}{2^N} \prod_{t=0}^{T-1} P[\{s,\,\sigma\}(t+1)|\{s,\,\sigma\}(t)]\,, \tag{3}$$

where we have considered completely random initial condition $P_0[\{s,\,\sigma\}(0)] = 1/2^N$.

To make predictions on the unobserved spins $\sigma_a(t)$, we assume that the model given by the couplings J is perfectly known and the posterior, i.e. conditional probability of the hidden spins defined by

$$P(\{\sigma\}|\{s\}) = \frac{P(\{s,\,\sigma\})}{P(\{s\})}\,, \tag{4}$$

J. Stat. Mech. (2014) P06013

gives the complete information for an optimal inference of hidden spins. Based on this probabilistic information, the best possible prediction $\sigma_a^{\rm opt}(t)$ for the hidden spin at site a and at time t is computed by

$$\sigma_a^{\rm opt}(t) = {\rm sign}\left[m_a(t)\right] , \tag{5}$$

where the local magnetization is defined as the posterior expectation

$$m_a(t) = \sum_{\{\sigma\}} \sigma_a(t)\, P(\{\sigma\}|\{s\}) . \tag{6}$$

Note that this does not correspond to the most likely spin *configuration* $\{\sigma\}$, because we have averaged out the configurations of spins $\sigma_b(t')$ for $b \neq a$ and $t' \neq t$.

Given a true 'teacher' sequence $\{\sigma^*\}$ of unobserved spins, we are interested in the total quality of the Bayes optimal prediction, i.e. in the expected probability of *wrongly* predicting a spin at site a and time t, given by the Bayes error

$$\varepsilon = \sum_{\{s,\sigma^*\}} P(\{s,\,\sigma^*\})\,\Theta(-\sigma_a^*(t)\,m_a(t)) = \sum_{\{s\}} P(\{s\}) \sum_{\{\sigma^*\}} P(\{\sigma^*\}|\{s\})\, \Theta(-\sigma_a^*(t)\,m_a(t)) , \tag{7}$$

where the step function $\Theta(x) = 1$ for $x > 0$ and 0 else. In the next section we will use the replica method to compute the error in the thermodynamic limit $N \to \infty$, when the couplings J are assumed to be mutually independent Gaussian random variables, with zero mean and variance of the order $1/N$.

3. Replica analysis

The posterior statistics of the hidden spins can be obtained from the following partition function

$$P(\{s\}) = \frac{1}{2^N} \sum_{\{\sigma\}} \prod_t P[\{s,\,\sigma\}(t+1)|\{s,\,\sigma\}(t)], \tag{8}$$

which equals the total probability of the observed spin configurations and is also the normalizer of the posterior probability. Typical performance in the thermodynamic limit for random couplings are then computed from the quenched average of the free energy $F = -\langle \ln P(\{s\})\rangle_{J,s}$, where the average is taken over the the couplings J and over the observed spin configurations with their weights $P(\{s\})$. Hence, the averaged free energy is given by

$$F = -\sum_{\{s\}} \left\langle P(\{s\}) \log P(\{\rm s\}) \right\rangle_J . \tag{9}$$

This average can be computed by the replica trick [9–11] in the following way:

$$F = -\lim_{n\to 1} \frac{\rm d}{{\rm d}n} \log \sum_{\{\rm s\}} \left\langle P^n(\{\rm s\}) \right\rangle . \tag{10}$$

For integer n, we have

$$\sum_{\{s\}} \left\langle P^n(\{s\}) \right\rangle_J = \frac{1}{2^{nN}} \sum_{\{s\}} \sum_{\{\sigma^{(1)}\}} \ldots \sum_{\{\sigma^{(n)}\}} \left\langle \left[\prod_{\alpha=1}^{n} \exp\left\{ \sum_{it} s_i(t+1)\, g_i^\alpha(t) \right.\right.\right.$$
$$\left.\left.\left. + \sum_{at} \sigma_a^\alpha(t+1)\, g_a^\alpha(t) - \sum_{it} \log 2\, \cosh[g_i^\alpha(t)] - \sum_{at} \log 2 \cosh[g_a^\alpha(t)] \right\} \right] \right\rangle_J , \quad (11)$$

with

$$g_i^\alpha(t) = \sum_j J_{ij} s_j(t) + \sum_b J_{ib} \sigma_b^\alpha(t), \qquad g_a^\alpha(t) = \sum_j J_{aj} s_j(t) + \sum_b J_{ab} \sigma_b^\alpha(t) \, . \quad (12)$$

To perform the average over the couplings J_{ij}, J_{ib}, J_{aj} and J_{ab}, which are assumed to be mutually independent Gaussian random variables with zero mean and variance k^2/N, we note that the fields $g_i^\alpha(t)$ and $g_a^\alpha(t)$ are also Gaussian, which are independent for different sites i and a, but will be dependent for different replica index α and β and also possibly for different times. This yields

$$\begin{aligned} \left\langle g_i^\alpha(t)\, g_i^\beta(t') \right\rangle &= \left\langle g_a^\alpha(t)\, g_a^\beta(t') \right\rangle = k^2(\lambda S(t,\, t') + (1-\lambda)\, Q^{\alpha\beta}(t,\, t')), \\ \left\langle g_i^\alpha(t)\, g_i^\alpha(t') \right\rangle &= \left\langle g_a^\alpha(t)\, g_a^\alpha(t') \right\rangle = k^2(\lambda S(t,\, t') + (1-\lambda)\, C^\alpha(t,\, t')), \end{aligned} \quad (13)$$

where we have defined the following order parameters

$$\begin{aligned} C^\alpha(t,\, t') &= \frac{1}{N_{\rm hid}} \sum_a \sigma_a^\alpha(t)\, \sigma_a^\alpha(t') \quad \text{for } t < t', \\ Q^{\alpha\beta}(t,\, t') &= \frac{1}{N_{\rm hid}} \sum_a \sigma_a^\alpha(t)\, \sigma_a^\beta(t') \quad \text{for } \alpha < \beta, \\ S(t,\, t') &= \frac{1}{N_{\rm obs}} \sum_i s_i(t)\, s_i(t') \qquad \text{for } t < t'. \end{aligned} \quad (14)$$

Introducing these definitions within δ functions and expressing the δ functions using conjugate (hatted) integration parameters, we get the following expression:

$$\sum_{\{s\}} \left\langle P^n(\mathbf{s}) \right\rangle_J = \frac{c}{2^{nN}} \int \prod_{t,t'} \prod_{\alpha<\beta} \left({\rm d}Q^{\alpha\beta}(t,\, t') {\rm d}\widehat{Q}^{\alpha\beta}(t,\, t') \right)$$
$$\times \prod_{t<t'} \prod_{\alpha} \left({\rm d}C^\alpha(t,\, t') {\rm d}\widehat{C}^\alpha(t,\, t') \right) \prod_{t<t'} \left({\rm d}S(t,\, t') {\rm d}\widehat{S}(t,\, t') \right)$$
$$\times \exp\left(iN_{\rm hid} \sum_\alpha \sum_{t<t'} C^\alpha(t,\, t')\, \widehat{C}^\alpha(t,\, t') + iN_{\rm hid} \sum_{\alpha<\beta} \sum_{tt'} Q^{\alpha\beta}(t,\, t')\, \widehat{Q}^{\alpha\beta}(t,\, t') \right.$$
$$+ iN_{\rm obs} \sum_{t<t'} S(t,\, t')\, \widehat{S}(t,\, t') + N_{\rm obs} \log \mathcal{E}_{\rm obs}(C,\, Q)$$
$$\left. + N_{\rm hid} \log \mathcal{E}_{\rm hid}(C,\, \widehat{C},\, Q,\, \widehat{Q}) \right), \quad (15)$$

J. Stat. Mech. (2014) P06013

where c is a trivial constant not depending on N,

$$\mathcal{E}_{\rm obs}(C,\,Q)=\sum_{\{s\}}\left\langle \exp\left(\sum_{t\alpha} s(t+1)\,g^{\alpha}(t)-\sum_{t\alpha}\log 2\,\cosh[g^{\alpha}(t)]\right.\right.$$
$$\left.\left.-i\sum_{t<t'}\widehat{S}(t,\,t')\,s(t)\,s(t')\right)\right\rangle_{g}$$
$$\mathcal{E}_{\rm hid}(C,\,\widehat{C},\,Q,\,\widehat{Q})=\sum_{\{\sigma^{(1)}\}}\ldots\sum_{\{\sigma^{(n)}\}}\exp\left\langle\left(\sum_{t}\sigma^{\alpha}(t+1)\,g^{\alpha}(t)\right.\right.$$
$$-\sum_{t\alpha}\log 2\,\cosh[g^{\alpha}(t)]-i\sum_{\alpha}\sum_{t<t'}\widehat{C}^{\alpha}(t,\,t')\sigma^{\alpha}(t)\sigma^{\alpha}(t')$$
$$\left.\left.-i\sum_{\alpha<\beta}\sum_{tt'}\widehat{Q}^{\alpha\beta}(t,\,t')\sigma^{\alpha}(t)\sigma^{\beta}(t')\right)\right\rangle_{g}, \tag{16}$$

and the average is over the Gaussian fields with statistics given by (14). In the limit $N\to\infty$, keeping the ratio $\lambda=N_{\rm obs}/N$ fixed, the integrals over the order parameters can be performed using the saddle point method, where we assume replica symmetry, i.e. $C^{\alpha}(t,\,t')=C(t,\,t')\;\forall\alpha,\;t<t'$ and, $Q^{\alpha\beta}(t,\,t')=Q(t,\,t')\;\forall\alpha<\beta,\;t,\;t'$. We get

$$\lim_{N\to\infty}\frac{1}{N}\log\sum_{\{s\}}\langle P^{n}(\{s\})\rangle={\rm Extr}\,f_{n}(C,\,S,\,\ldots),$$

where we have to take the extremum with respect to the order parameters in the expression

$$f_{n}(C,\,S,\,\ldots)=i(1-\lambda)\,n\sum_{t<t'}C(t,\,t')\,\widehat{C}(t,\,t')+\frac{i}{2}(1-\lambda)(n^{2}-n)$$
$$\times\sum_{tt'}Q(t,\,t')\,\widehat{Q}(t,\,t')+i\lambda\sum_{t<t'}S(t,\,t')\,\widehat{S}(t,\,t')$$
$$+\lambda\log\sum_{\{s\}}\left\langle\left\langle\prod_{t}V(t)\right\rangle_{\zeta}^{n}e^{\sum_{t}s(t)\nu(t)}\right\rangle_{\psi,\nu}$$
$$+(1-\lambda)\log\left\langle\left\langle\Gamma_{0}\prod_{t}Z(t)\right\rangle_{\xi,\zeta}^{n}\right\rangle_{\phi,\psi}-n\log 2, \tag{17}$$

where we have introduced

$$V(t)=\frac{{\rm e}^{s(t+1)(\psi(t)+\zeta(t))}}{2\cosh(\psi(t)+\zeta(t))},\qquad Z(t)=\frac{\cosh[\psi(t)+\zeta(t)+\phi(t+1)+\xi(t+1)]}{\cosh(\psi(t)+\zeta(t))}, \tag{18}$$

in terms of Gaussian independent random fields $\psi(t)$, $\zeta(t)$, $\nu(t)$, $\xi(t)$ and $\phi(t)$, with zero mean and covariances given by the following set of equations:

$$\begin{aligned}\left\langle \psi(t)\psi(t')\right\rangle &= k^2(\lambda S(t,\,t')+(1-\lambda)\,Q(t,\,t'))\,,\\ \left\langle \zeta(t)\zeta(t')\right\rangle &= k^2(1-\lambda)(C(t,\,t')-Q(t,\,t'))\,,\end{aligned} \tag{19}$$

$$\left\langle \nu(t)\nu(t')\right\rangle = -i\widehat{S}(t,\,t'),$$

$$\left\langle \xi(t)\xi(t')\right\rangle = -i(\widehat{C}(t,\,t')-\widehat{Q}(t,\,t')), \tag{20}$$

$$\left\langle \phi(t)\phi(t')\right\rangle = -i\widehat{Q}(t,\,t'), \tag{21}$$

for $t'\neq t$ and

$$\langle\psi(t)\psi(t)\rangle = k^2(\lambda+(1-\lambda)\,Q(t,\,t))\,,\qquad \langle\zeta(t)\zeta(t)\rangle = k^2(1-\lambda)(1-Q(t,\,t))\,, \tag{22}$$

$$\left\langle \nu(t)\nu(t)\right\rangle = 0,\qquad \left\langle \xi(t)\xi(t)\right\rangle = i\widehat{Q}(t,\,t), \tag{23}$$

$$\left\langle \phi(t)\phi(t)\right\rangle = -i\widehat{Q}(t,\,t), \tag{24}$$

for $t'=t$. The term Γ_0 contains the initial condition for the fields ϕ, ξ (appendix B). The three sets of Gaussian variables in (20), (21), (23) and (24) have been introduced to linearize the quadratic forms in equation (17). We can now perform the continuation to noninteger n and obtain the free energy per spin $\lim_{N\to\infty}F/N$ as the stationary value of

$$\begin{aligned}f(C,\,S,\,\ldots) = &-\tfrac{i}{2}\,(1-\lambda)\sum_{t<t'}C(t,\,t')\,\widehat{C}(t,\,t')-\tfrac{i}{2}(1-\lambda)\sum_{tt'}Q(t,\,t')\,\widehat{Q}(t,\,t')\\ &-\lambda\frac{\sum_{\{s\}}\left\langle\left\langle\prod_t V(t)\right\rangle_\zeta\log\left\langle\prod_t V(t)\right\rangle_\zeta \mathrm{e}^{-i\sum_t s(t)\nu(t)}\right\rangle_{\psi,\nu}}{\sum_{\{s\}}\left\langle\left\langle\prod_t V(t)\right\rangle_\zeta \mathrm{e}^{\sum_t s(t)\nu(t)}\right\rangle_{\psi,\nu}}\\ &-(1-\lambda)\frac{\left\langle\left\langle\Gamma_0\prod_t Z(t)\right\rangle_{\xi,\zeta}\log\left\langle\Gamma_0\prod_t Z(t)\right\rangle_{\xi,\zeta}\right\rangle_{\phi,\psi}}{\left\langle\Gamma_0\prod_t Z(t)\right\rangle_{\xi,\zeta,\phi,\psi}}.\end{aligned} \tag{25}$$

From equation (25) we can compute the selfaverageing values of the order parameters and their conjugates. Previous studies [1, 5, 15] of spin models with asymmetric couplings have shown that spin correlations $S(t,\,t')$ decay after one time step. Hence, we expect that also for our model the other two time order parameters are zero for $t\neq t'$. Indeed, we can show (for an example, see appendix A) that the results

$$C(t,\,t')=Q(t,\,t')=\widehat{C}(t,\,t')=\widehat{Q}(t,\,t')=\widehat{S}(t,\,t')=0$$

are self-consistent solutions of the order parameter equations for $t'\neq t$ and this solution is also supported by simulations. In this case, only the terms with $t'=t$ give non-zero contribution in equation (16) and the free energy of the system simplifies to

J. Stat. Mech. (2014) P06013

$$
\begin{aligned}
f(Q,\widehat{Q}) = &-\frac{i}{2}(1-\lambda)\sum_{t=0}^{T} Q(t)\,\widehat{Q}(t) - \frac{i}{2}(1-\lambda)\sum_{t=0}^{T}\widehat{Q}(t) \\
&-\lambda\sum_{t=0}^{T-1}\sum_{\{s\}(t+1)}\left\langle \left\langle V(t)\right\rangle_{\zeta t}\log\left\langle V(t)\right\rangle_{\zeta t}\right\rangle_{\psi_t} \\
&-(1-\lambda)\sum_{t=0}^{T-1}\frac{\left\langle \left\langle \widetilde{Z}(t)\right\rangle_{\zeta_t}\log\left\langle \widetilde{Z}(t)\right\rangle_{\zeta_t}\right\rangle_{\phi_{t+1},\psi_t}}{\left\langle \widetilde{Z}(t)\right\rangle_{\zeta_t,\phi_{t+1},\psi_t}} - (1-\lambda)\widetilde{\Gamma}_0,
\end{aligned}
\tag{26}
$$

where

$$
\widetilde{Z}(t) = \frac{\cosh[\psi(t)+\zeta(t)+\phi(t+1)]}{\cosh(\psi(t)+\zeta(t))},
$$

$Q(t) \equiv Q(t,t)$ and the initial condition $\widetilde{\Gamma}_0$ is given in appendix B. The order parameter $Q(t)$ gives the typical overlap of two independent spin configurations at time t drawn at random from the posterior distribution. By symmetry, it also describes the expected overlap of the hidden spins drawn from the posterior with the true 'teacher' spins of the model from which the observation data were generated. Hence the limit $Q(t) = 0$ describes a situations where the posterior gives no information on the hidden spins. On the other hand, $Q(t) = 1$, means that we can predict the hidden spins perfectly. We obtain the following equations for the order parameters:

$$
Q(t) = \frac{1}{\left\langle \widetilde{Z}(t-1)\right\rangle_{\zeta_{t-1},\phi_t,\psi_{t-1}}}\left\langle \frac{\left\langle \tanh\widetilde{A}(t-1)\;\widetilde{Z}(t-1)\right\rangle^2_{\zeta_{t-1}}}{\left\langle \widetilde{Z}(t-1)\right\rangle_{\zeta_{t-1}}}\right\rangle_{\phi_t,\psi_{t-1}}, \quad t=1\ldots T \tag{27}
$$

$$
\begin{aligned}
\widehat{Q}(t) = &\frac{ik^2(1-\lambda)}{\left\langle \widetilde{Z}(t)\right\rangle_{\zeta_t,\phi_{t+1},\psi_t}}\left\langle \frac{\left\langle [\tanh\widetilde{A}(t)-\tanh\widetilde{B}(t)]\widetilde{Z}(t)\right\rangle^2_{\zeta_t}}{\left\langle \widetilde{Z}(t)\right\rangle_{\zeta_t}}\right\rangle_{\phi_{t+1},\psi_t} \\
&+ ik^2\lambda\sum_{\{s\}(t+1)}\left\langle \frac{\left\langle [s(t+1)-\tanh\widetilde{B}(t)]\,V(t)\right\rangle^2_{\zeta_t}}{\left\langle V(t)\right\rangle_{\zeta_t}}\right\rangle_{\psi_t}, \quad t=0\ldots T-1
\end{aligned}
\tag{28}
$$

where

$$
\widetilde{A}(t) = \psi(t)+\zeta(t)+\phi(t+1), \qquad \widetilde{B}(t) = \psi(t)+\zeta(t).
$$

The equations for the initial and final conditions, $Q(0)$ and $\widehat{Q}(T)$, are given in in appendix B.

J. Stat. Mech. (2014) P06013

4. Distribution of local magnetization

It is easy to extend the replica approach to the computation of other thermodynamic quantities such as functions of the local magnetizations. We find that, in the thermodynamic limit, hidden spins can be viewed as mutually independent random variables which are coupled to random fields. The spins have local magnetizations

$$m(t|\psi,\,\phi)=\frac{\left\langle \tanh A(t-1)\,\widetilde{Z}(t-1)\right\rangle_{\zeta_{t-1}}}{\left\langle \widetilde{Z}(t-1)\right\rangle_{\zeta_{t-1}}}, \tag{29}$$

where the 'inner' averages over ζ reflect the averaging out of the other spins. The magnetizations depend on the random fields $\psi(t-1)$, $\phi(t)$. These Gaussian fields reflect the disorder originating from the random couplings. In computing expectations they get an extra statistical weight given by

$$w(\psi,\,\phi)=\frac{\left\langle \widetilde{Z}(t-1)\right\rangle_{\zeta_{t-1}}}{\left\langle \widetilde{Z}(t-1)\right\rangle_{\zeta_{t-1},\psi_{t-1},\phi_t}} \tag{30}$$

in the 'outer' average. Hence, the distribution of local magnetizations at an arbitrary site and at time t is given by

$$p_t(m)=\langle w(\psi,\,\phi)\,\delta(m-m(t|\psi,\,\phi))\rangle_{\psi_{t-1},\phi_t}, \tag{31}$$

from which the overlap Q is recovered as $Q(t)=\int_{-1}^{1}p_t(m)\,m^2\mathrm{d}m$. Finally, to get the Bayes error we note that the (posterior) probability of a spin σ equals $\frac{1}{2}(m\sigma+1)$. Hence equation (7) is translated into

$$\varepsilon=\frac{1}{2}\sum_{\sigma=\pm1}\int_{-1}^{1}p_t(m)\,(m\sigma+1)\,\Theta(-\sigma m)\,\mathrm{d}m=\frac{1}{2}\left(1-\int_{-1}^{1}p_t(m)\,|m|\,\mathrm{d}m\right), \tag{32}$$

where the last equality follows easily from the fact that $p_t(m)=p_t(-m)$.

5. Results

We have solved the order parameter equations (27) and (29) by iterating equations (22), (24), (27) and (29) for different values of the load parameter λ and coupling strength k (for an example see figure 1). We start the recursion from the prior initial condition $Q(0)=0$ and then iterate the equations forward and backward, updating the boundary conditions at each iteration according to equation (B.2). The overlap is smallest at the boundary $t=0$ and $t=T$, because there the information flow is only from one direction and is also expected to decay over the time T.

When the length T of the spin trajectories grows, the order parameters $Q(t)$ and $\widehat{Q}(t)$ for times t away from the boundaries, i.e. $0\ll t\ll T$, converge to stationary

J. Stat. Mech. (2014) P06013

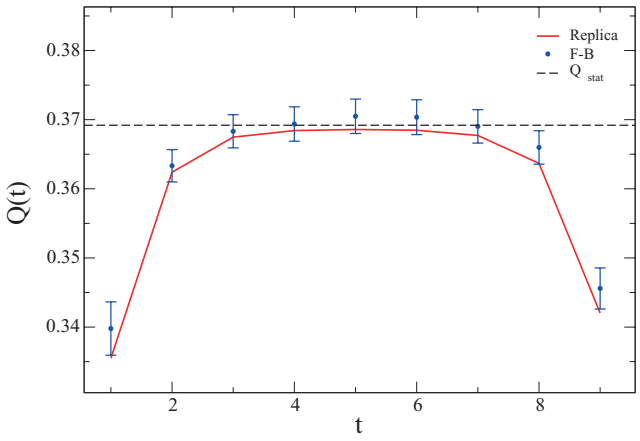

Figure 1. Order parameter Q as a function of time, for a system with $\lambda = 0.4$ and $k = 1$. Red line: solution of the order parameter equations. Black dashed line: stationary value $Q_{\rm stat}$ of the order parameter. Blue points: Q from numerical simulation of a system with $N_{\rm hid} = 10\,000$ hidden spins, averaged over 10 000 samples; the error bars represent the standard deviation.

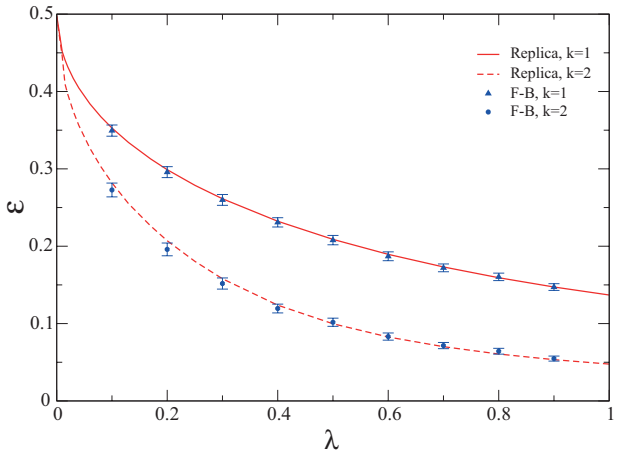

Figure 2. Bayes error as a function of the load factor for $k = 1$ (solid red line, blue triangles) and $k = 2$ (dashed red line, blue circles). Red lines: replica result, computed with the stationary values of the order parameters. Blue points: numerical simulation of a system with $N_{\rm hid} = 8$ hidden spins, averaged over 2500 samples; the error bars represent the standard deviation; the Bayes error is computed at time $t = T/2$, with $T = 20$.

values $Q_{\rm stat}$ and $\widehat{Q}_{\rm stat}$. These can be directly computed from equations (27–29) by setting $Q(t) = Q(t-1) = Q_{\rm stat}$ and $\widehat{Q}(t) = \widehat{Q}(t+1) = \widehat{Q}_{\rm stat}$. For given stationary order parameters we have then computed the distribution of local magnetizations and the Bayes error. The Bayes error ε is shown in figure 2 as a function of the load factor λ. In the limit of no observations, $\lambda = 0$, the prediction on the the state of hidden spins is completely random and the error has the trivial value $\varepsilon = 0.5$. The error rapidly decreases as λ gets larger, but remains nonzero for $\lambda = 1$, indicating the presence of a residual error in almost fully observed systems due to the stochasticity of the Markov process. Since the couplings are responsible for the propagation of information between spin sites, the Bayes error decreases as the coupling strength increases; in particular

J. Stat. Mech. (2014) P06013

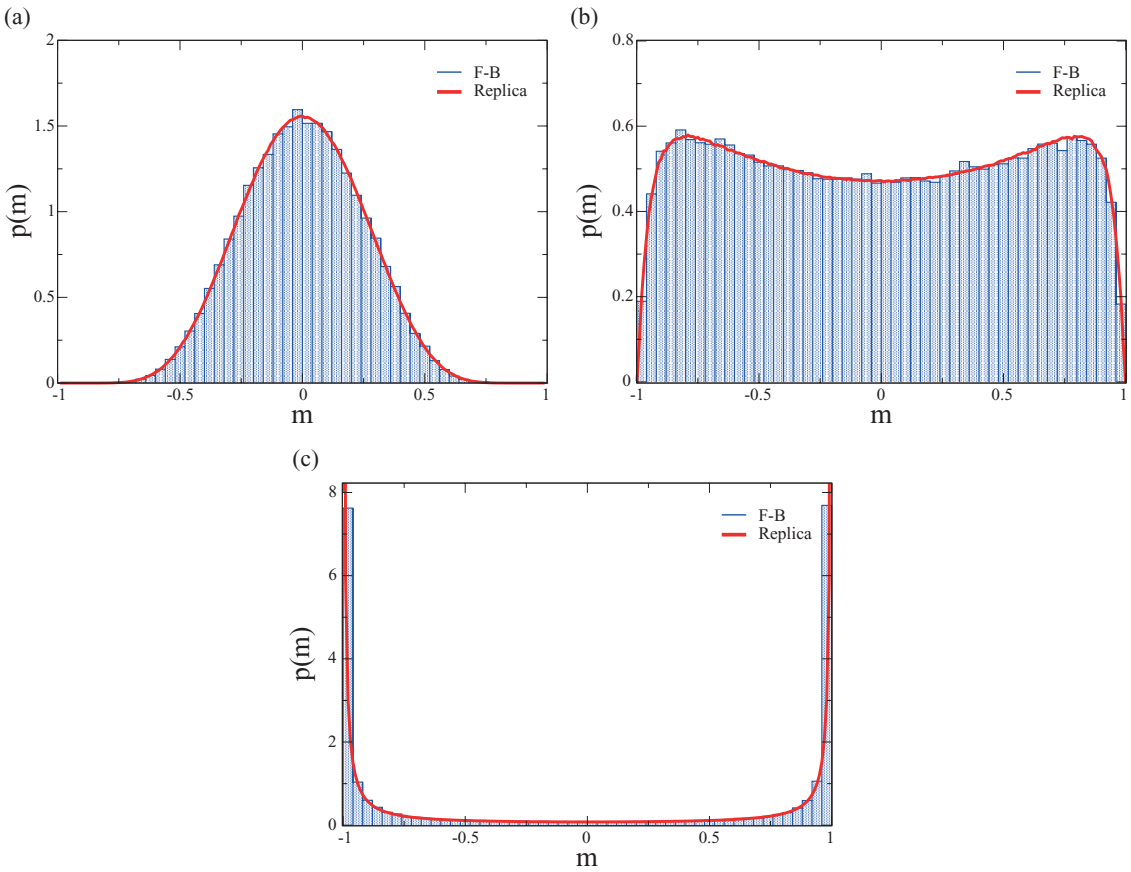

Figure 3. Distribution of local magnetization for load factor $\lambda = 0.8$ and coupling strengths $k = 0.2$ (a), $k = 0.6$ (b) and $k = 2$ (c). Red line: analytical result (equation (31)) assuming stationary values of the order parameters. Blue histogram: numerical simulations averaged over 80 000 samples for a system with eight hidden spins. The magnetization is computed at time $t = T/2$, with $T = 20$.

we find that $\varepsilon \to 0$ for $k \to \infty$. This behaviour is illustrated in figure 3, where the distribution $p(m)$ of the local magnetization (equation (31)) is shown. For small k the distribution is close to a Gaussian centered at zero, with vanishing variance as $k \to 0$, meaning (see equation (32)) that nontrivial prediction on the magnetization can be made. As k grows larger the distribution broadens and above a critical value the curve becomes bimodal. For large k, the distribution $p(m)$ concentrates at $m = \pm 1$, allowing for a perfect prediction of hidden spins.

Our analytical results agree very well with simulations of spin systems with relative small number of spins. For these systems we could compute local magnetizations $m_a(t)$ exactly by enumeration. The Markovian spin dynamics facilitated these computations with the use of a *forward–backward* algorithm [16] well known for hidden Markov models (appendix C). We compute $Q(t)$ using

$$Q(t) = \frac{1}{N_{\text{hid}}} \sum_{a=1}^{N_{\text{hid}}} E_{s,J} m_a^2(t),$$

where $E_{s,J}$ denotes the expectation over all possible observed spins and over the set of random couplings.

6. A comment on symmetric networks

From the point of view of equilibrium statistical physics a corresponding analysis for symmetric couplings $J_{ij} = J_{ji}$ might be of interest. In this case our approach would lead to additional order parameters (e.g. response functions). More important, order parameters would be usually non-zero for $t \neq t'$. Take for example the order parameter

$$\begin{aligned} C(t,\, t') &= E_J\left[\sum_{\{s\}} P(\{s\}) \sum_{\{\sigma\}} P(\{\sigma\}|\{s\})\sigma_a(t)\sigma_a(t')\right] \\ &= E_J\left[\sum_{\{s\},\{\sigma\}} P(\{\sigma\},\, \{s\})\sigma_a(t)\sigma_a(t')\right], \end{aligned} \tag{33}$$

where the last line follows from Bayes theorem. Hence, $C(t,\, t')$ equals the usual spin correlation in a system of $N_{\rm hid} + N_{\rm obs} = N$ spins, where there is no difference between hidden and observed spins (because there is *no conditioning* on the latter ones). Unfortunately, even for this simpler, more standard type of spin-glass model (studied extensively in the 1990s), exact analytical results for two time correlations (except for the case of uncorrelated couplings and Gaussian or spherical spin models) were not possible. A Monte Carlo approach to the effective non-Markovian single spin dynamics [15, 17] could be adapted to our model but it would require extensive nontrivial numerical simulations with an increasing complexity when the time window T grows. Moreover, this method cannot be easily extended to the stationary case.

To circumvent this problem, one might be tempted to resort to equilibrium techniques instead. In fact, for the case of symmetric couplings, the Markovian dynamics of the joint system of s and σ spins has a well known stationary equilibrium distribution. This static distribution is usually known as the Little model [18–20] and was frequently discussed in in the framework of Hopfield type neural networks with parallel dynamics. On might then calculate learning properties of the static model by using again the replica approach. While this should indeed be feasible (when replica symmetry breaking effects are neglected), one should note that this approach would consider a quite different statistical ensemble. The equilibrium case would deal with the probability $P(\sigma\,(t) \mid s(t))$ of spins at fixed large time t, whereas our dynamic ensemble is concerned with $P\,(\{\sigma\} \mid \{s\})$ with a conditioning on information $\{s\}$ from the *time history* of past and future observations.

Hence, the problem of solving the model with symmetric couplings is far different from the asymmetric case studied in this paper and will be postponed to future work.

7. Outlook

In this paper we have presented a first step in analyzing optimal Bayesian inference for kinetic Ising models with observed and unobserved spins valid for large random systems. The replica analysis revealed a fairly simple statistical picture of the posterior trajectories of hidden spins. Spins at different time steps (and sites) are statistically independent, but their local magnetizations depend on the propagation of information from past and future spins which is expressed through order parameters.

J. Stat. Mech. (2014) P06013

One can expect that this simple picture derived for the disorder averaged system can be translated into equations for the local magnetizations of hidden spins which are valid for a typical *single* system with fixed couplings and observations. In fact, such mean field equations generalizing the results of [1] to the case of observations can be derived from cavity arguments and could be used as an efficient algorithm for the computation of local magnetizations in large random networks. This could then be used as an approximation in the E-Step of an EM algorithm [6] which aims at computing the maximum likelihood estimator of the network couplings J_{ij}, averaging out unobserved spins. We will discuss such an approach in a forthcoming paper.

It will be interesting to extend this replica approach to other dynamical models. As long as we restrict ourselves to asymmetric random couplings one can expect that the case of continuous time (at least for the stationary limit) models could be treated. This would include e.g. continuous time Glauber dynamics and coupled stochastic differential equations (soft spin models).

Acknowledgments

This work is supported by the Marie Curie Training Network NETADIS (FP7, grant 290038).

Appendix A. Self consistent solution for the two time order parameters

Let us consider, as an example, the stationary value of the order parameter Q. From the saddle point equation $\dfrac{\partial f}{\partial \widehat{Q}} = 0$ we find:

$$Q(t,\,t') = \frac{1}{\left\langle I_0 \prod_\tau Z(\tau)\right\rangle_{\xi,\zeta,\phi,\psi}} \times \left\langle \frac{\left\langle I_0 \tanh A(t-1) \prod_\tau \mathrm{Z}(\tau)\right\rangle_{\xi,\zeta} \left\langle I_0 \tanh A(t'-1) \prod_\tau \mathrm{Z}(\tau)\right\rangle_{\xi,\zeta}}{\left\langle I_0 \prod_\tau Z(\tau)\right\rangle_{\xi,\zeta}} \right\rangle_{\phi,\psi}, \tag{A.1}$$

where

$$A(t) = \psi(t) + \zeta(t) + \phi(t+1) + \xi(t+1). \tag{A.2}$$

We want to show that $Q(t,\,t') = 0$ for $t \neq t'$ is a self consistent solution. If our assumption holds for the order parameters on the right hand side of equation (A.1), the averages over the gaussian fields factorize over time, yielding:

$$Q(t,\,t') = \frac{\left\langle \tanh A(t-1)\mathrm{Z}(t-1)\right\rangle_{\xi_{t-1},\zeta_{t-1},\phi_t,\psi_t}}{\left\langle Z(t-1)\right\rangle_{\xi_{t-1},\zeta_{t-1},\phi_t,\psi_t}} \frac{\left\langle \tanh A(t'-1)\mathrm{Z}(t'-1)\right\rangle_{\xi_{t'-1},\zeta_{t'-1},\phi_{t'},\psi_{t'}}}{\left\langle Z(t'-1)\right\rangle_{\xi_{t'-1},\zeta_{t'-1},\phi_{t'},\psi_{t'}}}. \tag{A.3}$$

The first two terms in the numerator of the above equation can be written in terms of the independent random variables $x = \psi(t-1) + \zeta(t-1)$ and $y = \varphi(t) + \xi(t)$ as

$$\left\langle \frac{\sinh(x+y)}{\cosh(x)} \right\rangle_{x,y} = \left\langle \frac{\sinh(x)\cosh(y)+\cosh(x)\sinh(y)}{\cosh(x)} \right\rangle_{x,y} = \langle \tanh(x) \rangle_x \langle \cosh(y) \rangle_y + \langle \sinh(y) \rangle_y = 0. \tag{A.4}$$

Using a similar procedure, this argument can be extended to all the other order parameters.

Appendix B. Boundary conditions

The parameters Γ_0 and $\widetilde{\Gamma_0}$ containing the initial conditions have the following expression:

$$\Gamma_0 = 2\cosh[\phi(0)+\xi(0)], \qquad \widetilde{\Gamma_0} = \frac{\langle \cosh(\phi_0)\log(2\cosh(\phi_0)) \rangle_{\phi_0}}{\langle \cosh(\phi_0) \rangle_{\phi_0}}. \tag{B.1}$$

The initial and final condition for the order parameter are:

$$Q(0) = \frac{\langle \tanh(\phi_0)\sinh(\phi_0) \rangle_{\phi_0}}{\langle \cosh(\phi_0) \rangle_{\phi_0}}, \qquad \widehat{Q}(T) = 0. \tag{B.2}$$

Appendix C. Forward-backward algorithm

In order to compute the local magnetizations of hidden spins at each time t, we need the posterior distribution $P[\{\sigma\}(t)|\{s\}]$ $1 \leqslant t < T$ of the hidden spins at time t, given the obserserved spins at *all times*.

It is convenient to divide the computation of $P[\{\sigma\}(t)|\{s\}]$ in two parts, one involving the spins up to time $t+1$, the other the spins from $t+2$ to T:

$$\begin{aligned} P[\{\sigma\}(t)|\{s\}] &= P[\{\sigma\}(t)|\{s\}_{1:t+1}, \{s\}_{t+2:T} \\ &\propto P[\{\sigma\}(t)|\{s\}_{1:t+1}] P[\{s\}_{t+2:T}|\{\sigma\}(t), \{s\}(t+1)], \end{aligned} \tag{C.1}$$

where the last line follows from Bayes' rule and the conditional independence of $\{s\}_{t+2:T}$ and $\{s\}_{1:t}$ given $\{s\}(t+1)$ and $\{\sigma\}(t)$. The two terms in the right hand side of equation (C.2) can be computed by recursion through time. In particular, it can be shown [16] that the first term, referred to as the 'forward message', $\mathrm{fm}[\{\sigma\}(t)] = P[\{\sigma\}(t)|\{s\}_{1:t+1}]$, is obtained by a forward recursion form 1 to t governed by the following equation:

$$\begin{aligned} \mathrm{fm}[\{\sigma\}(t)] &\propto P[\{s\}(t+1)|\{\sigma, s\}(t)] \\ &\times \sum_{\{\sigma\}(t-1)} P[\{\sigma\}(t)|\sigma(t-1), \{s\}(t-1)]\ \mathrm{fm}[\{\sigma\}(t-1)]. \end{aligned} \tag{C.2}$$

J. Stat. Mech. (2014) P06013

The second term, or 'backward message' $\mathrm{bm}[\{\sigma\}(t+1)] = P[\{s\}_{t+2:T}|\{\sigma\}(t), \{s\}(t+1)]$, is obtained by a backward recursion, running from T to $t+1$ and obeying:

$$\mathrm{bm}[\{\sigma\}(t+1)] = \sum_{\{\sigma\}(t+1)} P[\{s\}(t+2)|\{\sigma, s\}(t+1)]\,\mathrm{bm}[\{\sigma\}(t+2)] \times P[\{\sigma\}(t+1)|\{\sigma, s\}(t)]. \tag{C.3}$$

References

[1] Mézard M and Sakellariou J 2011 *Exact mean-field inference in asymmetric kinetic Ising systems* *J. Stat. Mech.* L07001
[2] Roudi Y and Hertz J 2011 *Mean field theory for nonequilibrium network reconstruction* Phys. Rev. Lett. **106** 048702
[3] Zeng H-L, Aurell E, Alava M and Mahmoudi H 2011 *Network inference using asynchronously updated kinetic Ising model* Phys. Rev. E **83** 041135
[4] Sakellariou J, Roudi Y, Mézard M and Hertz J 2012 *Effect of coupling asymmetry on mean-field solutions of the direct and inverse Sherrington–Kirkpatrick model* Phil. Mag. **92** 272–9
[5] Huang H and Kabashima Y 2013 *Dynamics of asymmetric kinetic Ising systems revisited* (arXiv:1310.5003)
[6] Dempster A P, Laird N M and Rubin D B 1977 *Maximum likelihood from incomplete data via the EM algorithm* J. R. Stat. Soc. Ser. B **39** 1–38
[7] Dunn B and Roudi Y 2013 *Learning and inference in a nonequilibrium Ising model with hidden nodes* Phys. Rev. E **87** 022127
[8] Tyrcha J and Hertz J 2014 *Network inference with hidden units* Math. Biosci. Eng. **11** 149
[9] Opper M and Kinzel W 1996 *Statistical mechanics of generalization* Models of Neural Networks III ed E Domany *et al* (Berlin: Springer)
[10] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford: Oxford University Press)
[11] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[12] Crisanti A and Sompolinsky H 1987 *Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model* Phys. Rev. A **36** 4922–39
[13] Sompolinsky H, Crisanti A and Sommers H J 1988 *Chaos in random neural networks* Phys. Rev. Lett. **61** 259–62
[14] Parisi G 1986 *Asymmetric neural networks and the process of learning* J. Phys. A: Math. Gen. **19** L675
[15] Eissfeller H and Opper M 1994 *Mean-field Monte Carlo approach to the Sherrington–Kirkpatrick model with asymmetric couplings* Phys. Rev. E **50** 709–20
[16] Russel S and Norvig P 1995 *Artificial Intelligence: a Modern Approach* (Upper Saddle River, NJ: Pearson Education)
[17] Scharnagl A, Opper M and Kinzel W 1995 *On the relaxation of infinite-range spin glasses* J. Phys. A: Math. Gen. **28** 5721
[18] Little W 1974 *The existence of persistent states in the brain* Math. Biosci. **19** 101–20
[19] Little W and Shaw G L 1975 *A statistical theory of short and long term memory* Behav. Biol. **14** 115–33
[20] Little W and Shaw G L 1978 *Analytic study of the memory storage capacity of a neural network* Math. Biosci. **39** 281–90

J. Stat. Mech. (2014) P06013

6.3 Further results

6.3.1 Inferring hidden states in a kinetic Ising model via the extended Plefka expansion

In this section we continue to study the kinetic Ising model with hidden nodes, as introduced in the previous section, and provide an approximate method to estimate the posterior expectation of hidden spins given the observed ones. This is achieved by finding a suitable approximation to the log-likelihood of the data, that corresponds to the generating function of the posterior moments. We will work in a generating functional approach and provide a mean field approximation to the log-likelihood, and test it on a fully connected dense network with Gaussian independent random couplings. In the previous chapter, we showed that accurate mean field equations for the kinetic Ising model with dense Gaussian random couplings are found via the Extended Plefka expansion, namely a weak coupling expansion of the Legendre transform of the log-likelihood, at fixed first and second moment over time. We now extend this formalism to the case in which a finite fraction of the trajectories of the spins are observed and the rest are hidden. We thus consider a model composed of two sets of variables: the observed spins $\{s_i(t)\}$, $i = 1 \dots N_{\text{OBS}}$ and the hidden spins $\{\sigma_a(t)\}$, $a = 1 \dots N_{\text{HID}}$. The dynamics is defined by the following transition probability:

$$p[\{s,\sigma\}(t+1)|\{s,\sigma\}(t)] = \frac{\exp\left[\sum_i s_i(t+1)g_i(t) + \sum_a \sigma_a(t+1)g_a(t)\right]}{\prod_{i,a} 2\cosh[g_i(t)]2\cosh[g_a(t)]}, \quad (6.1)$$

where $g_i(t)$ and $g_a(t)$ are the fields acting on observed spin i and hidden spin a at time t:

$$\begin{aligned} g_i(t) &= \sum_j J_{ij} s_j(t) + \sum_b J_{ib}\sigma_b(t), \\ g_a(t) &= \sum_j J_{aj} s_j(t) + \sum_b J_{ab}\sigma_b(t)\,, \end{aligned} \quad (6.2)$$

and $J_{ij}, J_{ia}, J_{ai}, J_{ab}$ are the observed-to-observed, observed-to-hidden, hidden-to-observed and hidden-to-hidden couplings. The likelihood of the observed spin configuration under the dynamics defined by (6.1) is

$$p[\{\mathbf{s}(t)\}_{0:T}] = Tr_\sigma \prod_t p[\{s,\sigma\}(t+1)|\{s,\sigma\}(t)], \quad (6.3)$$

and its computation involves the trace over the hidden spins trajectories $\{\sigma(1)\dots\sigma(T)\}$. To perform the calculation, we introduce a set of the auxiliary fields, $\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}$, and consider the following functional:

$$\mathcal{L}_\alpha[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}] = \log \int DG\ Tr_\sigma \exp\{\Omega_\alpha[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}]\}, \tag{6.4}$$

where

$$\begin{aligned}
\Omega_\alpha[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}] = &\sum_{it} s_i(t+1) g_i(t) + \sum_{at} \sigma_a(t+1) g_a(t) \\
&- \sum_{it} \log 2\cosh[g_i(t)] - \sum_{at} \log 2\cosh[g_a(t)] \\
&\sum_{it} i\hat{g}_i(t) \left[g_i(t) - \sum_j J_{ij} s_j(t) - \alpha \sum_b J_{ib} \sigma_b(t) \right] \\
&+ \sum_{at} i\hat{g}_a(t) \left[g_a(t) - \sum_j J_{aj} s_j(t) - \alpha \sum_b J_{ab} \sigma_b(t) \right] \\
&+ \sum_{at} \psi_a(t) \sigma_a(t) - i \sum_{it} H_i(t) \hat{g}_i(t) - i \sum_{at} \mathcal{H}_a(t) \hat{g}_a(t) \\
&+ \sum_{at} \hat{B}_i(t)\, \hat{g}_i^2(t) + \sum_{at} \hat{\mathcal{B}}_a(t)\, \hat{g}_a^2(t),
\end{aligned} \tag{6.5}$$

and where we have defined $G = \{g_i, \hat{g}_i, g_a, \hat{g}_a\}$. Note that the log likelihood of the data is recovered in the limit of auxiliary fields going to zero: $\psi \to 0, \hat{\mathcal{C}} \to 0, \hat{B} \to 0, \hat{\mathcal{B}} \to 0, \hat{\mathcal{G}} \to 0$. We have introduced the parameter α in the above functional to control the magnitude of the observed-to-hidden and hidden-to-hidden couplings. When $\alpha = 0$ the hidden units are not interacting among themselves nor are they influenced by the dynamics of the observed units and, due to the normalizing term over the field $g_a(t)$, also the hidden-to-observed couplings cancel from the likelihood, i.e. all the random variables decouple. The Plefka expansion will allow us to analytically compute an approximation to the likelihood by adding first and second order corrections to the non–interacting description. This is achieved by considering the Legendre transform of the log–likelihood (6.4) at fixed first and second local moments. These

moments can be defined by derivatives of the log–lilelihood as follows:

$$\mu_a(t) = \frac{\partial \mathcal{L}_\alpha}{\partial \psi_a(t)} = \langle \sigma_a(t) \rangle_\alpha \tag{6.6}$$

$$-i\hat{m}_i(t) = \frac{\partial \mathcal{L}_\alpha}{\partial H_i(t)} = -i\langle \hat{g}_i(t) \rangle_\alpha \tag{6.7}$$

$$-i\hat{\mu}_a(t) = \frac{\partial \mathcal{L}_\alpha}{\partial \mathcal{H}_a(t)} = -i\langle \hat{g}_a(t) \rangle_\alpha \tag{6.8}$$

$$B_i(t) \equiv \frac{\partial \mathcal{L}_\alpha}{\partial \hat{B}_i(t)} = \langle \hat{g}_i^2(t) \rangle_\alpha \tag{6.9}$$

$$\mathcal{B}_a(t) \equiv \frac{\partial \mathcal{L}_\alpha}{\partial \hat{\mathcal{B}}_a(t)} = \langle \hat{g}_a^2(t) \rangle_\alpha \tag{6.10}$$

where $\langle \cdots \rangle_\alpha$ denotes averaging over the distribution defined by the measure inside the functional (6.4). Namely, for any function $F(\sigma)$ of the trajectory of hidden spins σ we define:

$$\langle F \rangle_\alpha = \frac{\int D\mathcal{G}\, \mathrm{Tr}_\sigma\ F(\sigma)\ \exp\Big(\Omega_\alpha[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}]\Big)}{\int D\mathcal{G}\, \mathrm{Tr}_\sigma \exp\Big(\Omega_\alpha[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}]\Big)}. \tag{6.11}$$

By setting the auxiliary fields to zero and $\alpha = 1$, one recovers the posterior moments of the original dynamical system (6.1). Note that we are fixing only local moments, as this will allow us to derive marginal distributions of spin trajectories. We are also neglecting two times moments, as we are considering an asymmetric network, for which we know (Paper 4) that two times correlations of the form $C(t, t')$ decay to zero if $|t - t'| > 1$. By computing the Legendre transform of $\mathcal{L}_\alpha$ with respect to the fields $\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}$ and performing an extended Plefka expansion of the resulting functional, as explained in Paper 1 (for details, see Section 6.A), we are able to derive an effective single-site log-likelihood:

$$\mathcal{L}_0[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}] = \log \left\{ \prod_{it} \left\langle \frac{e^{s_i(t+1)h_i(t)}}{2\cosh\ h_i(t)} \right\rangle_{h_i(t)} \prod_{at} \left\langle \frac{\cosh\left[h_a(t) + u_a(t+1)\right]}{2\cosh\ h_a(t)} \right\rangle_{h_a(t)} \right\} \tag{6.12}$$

where we have introduced the gaussian variables $h_i(t) \sim \mathcal{N}(\gamma_i(t), V_i(t))$ and $h_a(t) \sim \mathcal{N}(\gamma_a(t), V_a(t))$:

$$\begin{aligned}
u_a(t) &= \psi_a(t) - i\sum_i J_{ia}\hat{m}_i(t) - i\sum_b J_{ba}\hat{\mu}_b(t) + \sum_i J_{ia}^2\mu_a(t)\left(B_i(t) - \hat{m}_i^2(t)\right) \\
&\quad + \sum_b J_{ba}^2\mu_a(t)\left(\mathcal{B}_b(t) - \hat{\mu}_b^2(t)\right) \\
\gamma_i(t) &= \sum_j J_{ij}s_j(t) + \sum_b J_{ib}\mu_b(t) + i\sum_a J_{ia}^2\hat{m}_i(t)\left(1-\mu_a^2(t)\right) + H_i(t) \\
\gamma_a(t) &= \sum_j J_{aj}s_j(t) + \sum_b J_{ab}\mu_b(t) + i\sum_b J_{ab}^2\hat{\mu}_a(t)\left(1-\mu_b^2(t)\right) + \mathcal{H}_a(t) \\
V_i(t) &= \langle\phi_i(t)\phi_i(t)\rangle = \sum_a J_{ia}^2\left(1-\mu_a^2(t)\right) - 2\hat{B}_i(t) \\
V_a(t) &= \langle\phi_a(t)\phi_a(t)\rangle = \sum_b J_{ab}^2\left(1-\mu_b^2(t)\right) - 2\hat{\mathcal{B}}_a(t).
\end{aligned} \tag{6.13}$$

The auxiliary fields $\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}$ have to be set to zero in order to obtain the marginal moments of hidden spins for the original model (6.1, 6.3), within the approximate description. The TAP equations for these moments are obtained from the set of equations (6.6-6.10) as follows:

$$\mu_a(t) = \lim_{\{\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}\}\to 0} \frac{\partial\mathcal{L}_0}{\partial\psi_a(t)} = \frac{\left\langle \tanh[h_a(t-1)+u_a(t)]\frac{\cosh[h_a(t-1)+u_a(t)]}{\cosh h_a(t-1)}\right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t-1)+u_a(t)]}{\cosh h_a(t-1)}\right\rangle_{h_a}} \tag{6.14}$$

$$-i\hat{m}_i(t) = \lim_{\{\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}\}\to 0} \frac{\partial\mathcal{L}_0}{\partial H_i(t)} = s_i(t+1) - \frac{\left\langle \tanh h_i(t)\frac{e^{s_i(t+1)h_i(t)}}{2\cosh h_i(t)}\right\rangle_{h_i}}{\left\langle \frac{e^{s_i(t+1)h_i(t)}}{2\cosh h_i(t)}\right\rangle_{h_i}} \tag{6.15}$$

$$-i\hat{\mu}_a(t) = \lim_{\{\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}\}\to 0} \frac{\partial\mathcal{L}_0}{\partial\hat{\mathcal{H}}_a(t)} = \mu_a(t+1) - \frac{\left\langle \tanh h_a(t)\frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)}\right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)}\right\rangle_{h_a}} \tag{6.16}$$

$$B_i(t) = \lim_{\{\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}\}\to 0} \frac{\partial \mathcal{L}_0}{\partial \hat{B}_i(t)} = \frac{\left\langle \left(1-\tanh^2 h_i(t)\right) \frac{e^{s_i(t+1)h_i(t)}}{2\cosh h_i(t)} \right\rangle_{h_i}}{\left\langle \frac{e^{s_i(t+1)h_i(t)}}{2\cosh h_i(t)} \right\rangle_{h_i}} - \frac{\left\langle \left(s_i(t+1)-\tanh h_i(t)\right)^2 \frac{e^{s_i(t+1)h_i(t)}}{2\cosh h_i(t)} \right\rangle_{h_i}}{\left\langle \frac{e^{s_i(t+1)h_i(t)}}{2\cosh h_i(t)} \right\rangle_{h_i}} \tag{6.17}$$

$$\mathcal{B}_a(t) = \lim_{\{\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}\}\to 0} \frac{\partial \mathcal{L}_0}{\partial \hat{\mathcal{B}}_a(t)} = \frac{\left\langle \left(1-\tanh^2 h_a(t)\right) \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}} - \frac{Tr_{\sigma_a} \left\langle \left(\sigma_a(t+1)-\tanh h_a(t)\right)^2 \frac{e^{\sigma_a(t+1)[h_a(t)+u_a(t+1)]}}{2\cosh h_a(t)} \right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}} \tag{6.18}$$

It is interesting to notice the structure of time dependences in the set of TAP equations: the posterior average of hidden spins (magnetization) propagates forward the message from past spins, while the posterior averages of the auxiliary variables $\{\hat{g}_i, \hat{g}_a\}$ are responsible for the back–propagation of the information coming from observations at future times.

To test the correctness of our results, we consider a network with couplings J_{ij}, J_{ib}, J_{aj} and J_{ab}, which are assumed to be mutually independent Gaussian random variables with zero mean and variance $1/N$. As a first check, we make a comparison with the average case scenario studied in Paper 4. Namely, we solve the system of equations (6.14-6.18) iteratively to compute the magnetization of hidden spins $\mu_a(t)$ at all sites a and times t, and use it to compute the order parameter

$$Q(t) = \frac{1}{N_{\rm HID}} \sum_{a=1}^{N_{\rm hid}} E_{s,J}\, \mu_a^2(t), \tag{6.19}$$

where $E_{s,J}$ denotes empirical averaging over different realizations of the network and different datasets generated from each realization. We compare the obtained result with the theoretical value of $Q(t)$ derived in Paper 4 by means of a replica calculation. Figure 6.1 shows a good agreement between the two values.

We then consider a single instance of the network, and compare the magnetisations predicted via the extended Plefka expansion with the exact ones, which can be computed with the forward-backward algorithm explained in

Appendix E of Paper 4. The result is shown in figure 6.2 , where we consider a network in which the 20% of the spins are hidden; in this case, according to our previous analysis (Paper 4, Figure 2), an exact calculation of the posterior average of hidden spins would predict the true magnetisation with relatively good accuracy. Hence, the good matching between exact and estimated magnetisations in figure 6.2 indicates that the extended Plefka expansion provides a good approximation to the true posterior averages.

To conclude this section, we would like to point out that the Plefka expansion can also be applied to study a scenario where all spins can be observed simultaneously, but observations are sparse in times; namely, in between observed time steps, there is a number T_u of time steps of the Glauber dynamics where the state of the system is not observed.

As a preliminary analysis to tackle the problem of inferring the state of the spins at unobserved time steps, we consider one set of spin trajectories, for which we can observe the spin state at the initial and final time steps. By concatenation of such trajectories, we can then describe the sparse observation scenario. In Appendix 6.B , we show how to compute the conditional expectation $\mu_a(t)$ of the spins at unobserved time steps given the observations, using the Plefka expansion. We can then estimate the values of the spins as

$$\sigma_a^{est}(t) = \text{sign}[\mu_a(t)]. \tag{6.20}$$

In figure 6.3 , we show the average percentage of correctly inferred spins as a function of T_u . It is interesting to notice that, even if there is only one unobserved time step, we cannot, on average, infer its state with a precision higher than 75% , due to the stochastic nature of the system. The analysis of sparse time-series data appears to be an interesting problem for future research. As a next step in the analysis, we should compare the results with the exact local magnetizations at unobserved time steps, which we could obtain from an algorithm of the forward-backward type.

6.3.2 Network reconstruction

In the previous section, we derived a mean-field approximation for the posterior statistics of hidden spins that could be used in the E step of an EM algorithm aimed at computing the maximum likelihood value of the couplings. In the E step, one estimates the posterior moments (6.14-6.18) and the log-likelihood (6.12) at fixed couplings. In the M step, one updates the couplings proportional to the derivatives of the log-likelihood with respect to them. We implemented this algorithm and applied it to a system with Gaussian independent couplings scaling as $1/N$. Two main problems emerge:

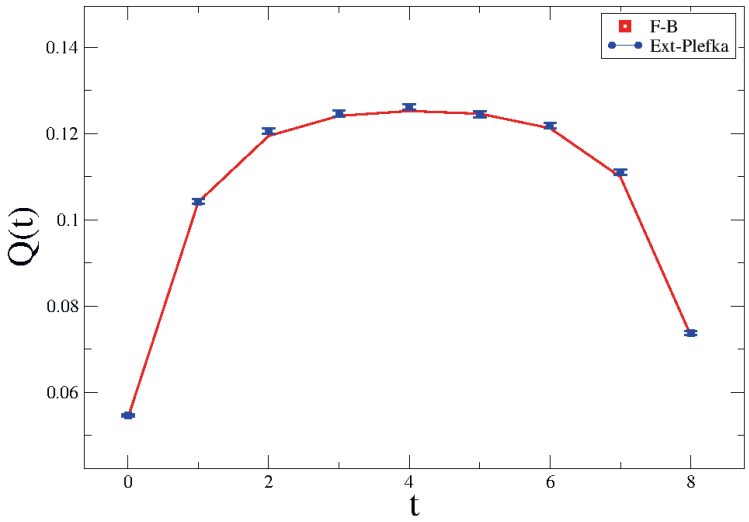

Figure 6.1: The blue dots show the order parameter Q (see 6.19) computed from the Extended Plefka's expansion, as a function of time. Averages are over 5 realizations of the networks and 10 datasets generated from each network. The red line corresponds to the theoretical result we obtained in Paper 4. The considered network is composed of 20 spins, the 10% of them being hidden. The teacher couplings are drawn as independent Gaussian random variables with zero mean and variance $1/N$.

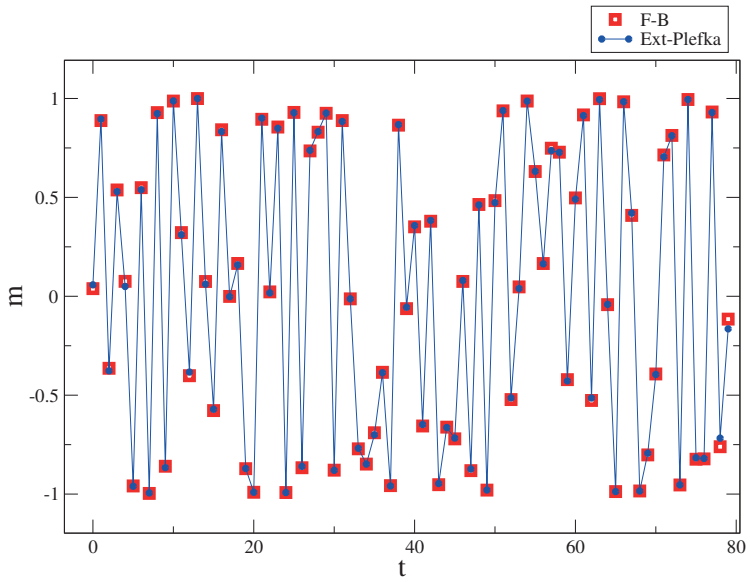

Figure 6.2: Magnetization of one hidden spin as a function of time. We compare the result we get from our Plefka's expansion (blue dots), with the theoretical result obtained by using the Forward-Backward algorithm explained in Appendix E of Paper 4, Appendix E (red dots). We consider a network of 20 spins where the hidden units are the 20% of the total number of spins. The teacher couplings are drawn as independent Gaussian random variables with zero mean and variance $1/N$.

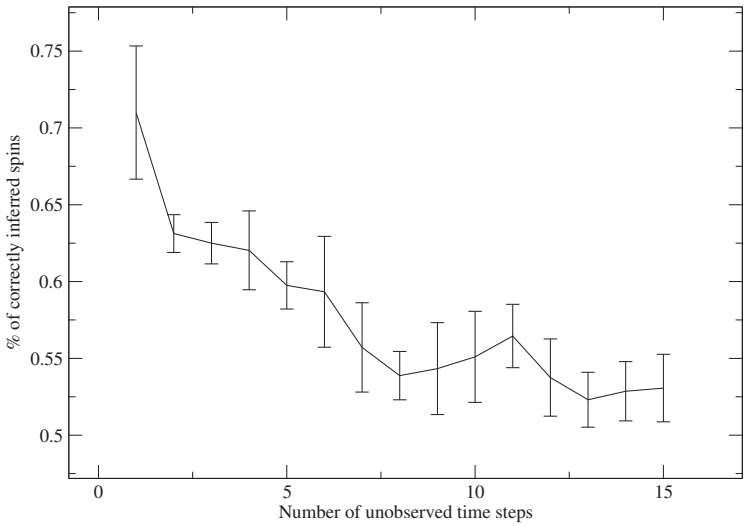

Figure 6.3: Percentage of correctly inferred spin values at unobserved time steps vs T_u. The spin values are estimated using (6.20). We consider a system of 20 spins. Results are averaged over 5 istances of the network. The teacher couplings are drawn as independent Gaussian random variables with zero mean and variance $1/N$. When the number of unobserved time steps exceeds $T_u = 12$, the probability of correctly inferring the spin value drops to 0.5.

- *Convergence to local maxima.* We know that the expectation maximisation algorithm is guaranteed to converge to the maximum of the log-likelihood, under mild regularity conditions [Wu83]. If the log-likelihood has multiple maxima, the algorithm could converge to different local maxima depending on the initial conditions. This is exactly the situation that we observe. If the initial condition for the couplings corresponds to the true value of the couplings plus a small Gaussian noise, we can reconstruct the network accurately, as shown in figure 6.4. On the contrary, if we start from random couplings, the algorithm converges to a different point in the parameter space, far from the true one.

- *Regions of the parameter space explored during the search.* The E step of the algorithm is based on the extended Plefka expansion that relies on the assumption that the couplings are weak and gives accurate results if the variance of the coupling distribution is a quantity of order $1/N$. During the search, the algorithm reaches regions of the parameter space where this assumption seems to be no longer valid; for instance, where most of the couplings are small and a few are significantly larger. It is not clear how accurately we are able to predict the posterior moments in those regions, with a large inaccuracy possibly compromising the upcoming steps of the algorithm.

By introducing a prior over the couplings (i.e., by adding a regularising term to the log-likelihood), the performances of the algorithm improve, but convergence to the desired result is still not obtained for most random choices of the initial conditions. As we mentioned in the introduction, recent works [TH13, DR13, BHTR15] showed that the presence of hidden-to-hidden connections seems to be one of the major sources of complication. One could then think to design an algorithm to find the maximum likelihood solution for the other three sets of couplings, where the likelihood has been marginalised out over the hidden-to-hidden couplings. Moreover, a theoretical study of how the average prediction error for inferring the couplings scales as a function of the length of available data trajectories could also be an important starting point to better understand this problem. In general, we think that further investigation is needed to design a better algorithm, and we postpone this for future work.

6.4 Conclusions

In this chapter, we studied the problem of inferring hidden states in a kinetic Ising model, where a fraction of the trajectories of spins is observed and a

fraction is hidden. Couplings are independent Gaussian random variables, and are present between the observed variables, between the hidden one, from hidden to observed and vice versa.

First, we computed the average error of the Bayes optimal predictor for the hidden spins, which can be determined from the posterior expectation of hidden spins, and studied its dependence on the fraction of the hidden units and on the stochasticity of the dynamics.

Then, we considered one particular realization of the couplings. The posterior average of hidden spins can be computed exactly using a forward-backward algorithm, but it becomes intractable if the number of hidden units exceeds a ten. An approximation of posterior expectations is found by applying the extended Plefka expansion to the log-likelihood of the data, which corresponds to the generating function of the posterior moments; this yields a set of equations for the posterior moments, which can be solved iteratively. The predicted magnetization of hidden spins agrees very well with the exact result from the forward-backward algorithm.

We also delineated two future directions of our analysis. In a scenario where all spins are measurable, but observations are sparse in times, one can predict the state of the spins at unobserved time steps using the extened Plefka expansion. Moreover, our mean-field method can be used in the E step if an EM type of algorithm is aimed at inferring the parameters of the model, namely, the couplings between the spins. We observed that, in a naive version of the algorithm, where the M step is performed by simple gradient ascent, several important questions remain open, mainly related to the algorithm getting stuck in local maxima where the hypothesis of the approximation used in the E step might be violated. While a lot of research has been devoted to the problem of escaping local maxima (see, e.g., [MM11]), it is less trivial to identify an efficient method to test the validity of the mean-field equations in the regions explored by the algorithm. We believe that the design of such an inference algorithm deserves further investigation. As a conclusive remark, we stress that, to apply this model to a real-world system, one should know beforehand the number of hidden spins (N_{HID}) , which is hardly realisable in practice. In [DR13] , the authors argue that the number of hidden units could also be inferred from the data, as the value that maximises the log-likelihood of observed spins considered a function of N_{HID}. An interesting alternative perspective is mentioned in [TH13] , where the authors propose to abandon the idea of modelling the details of the unrecorded units but instead model the activity of entire populations of hidden units by hidden variables. For instance, imagine we wanted to use our model to infer the effective connectivity of a group of excitatory neurons, whose activity can be simultaneously mea-

sured, coupled to a population of unrecorded inhibitory neurons. In this case, we could represent the activity of the whole unrecorded inhibitory population, which could be either active or inactive at a given time step, by a single hidden node.

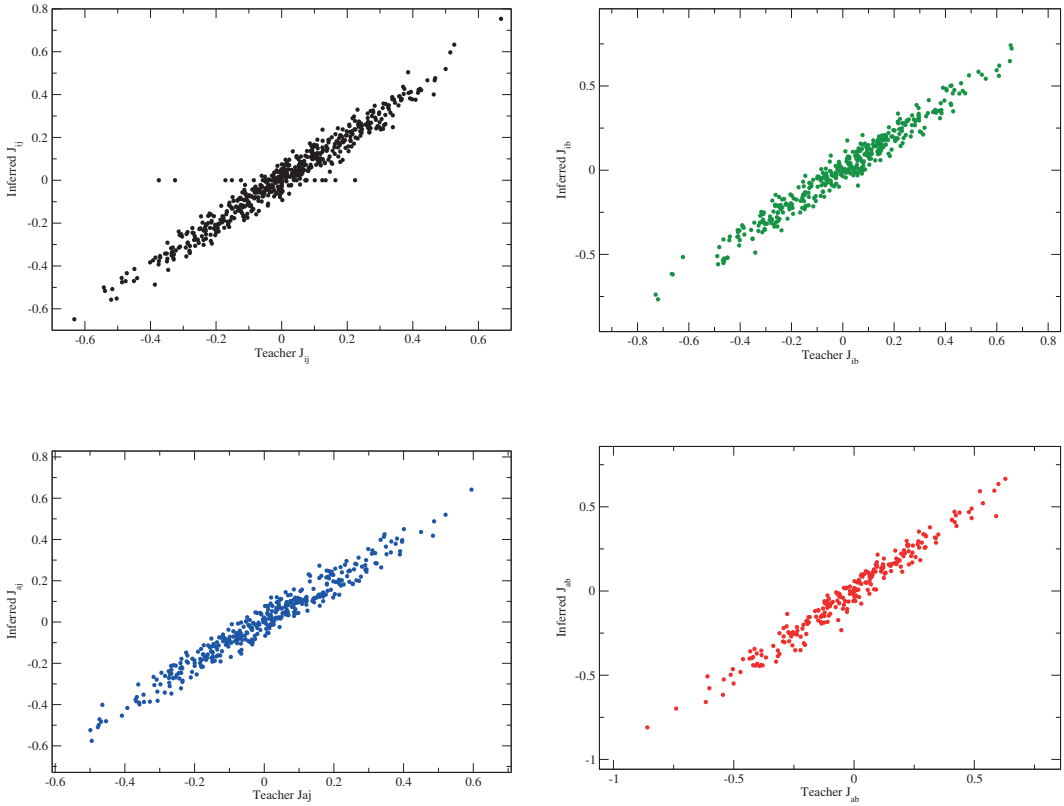

Figure 6.4: Scatter plots showing the inferred couplings versus the correct ones. We consider a network with $N = 40$ spins with 16 hidden units and a trajectory length of 40000 time steps and start from initial conditions close to the true values of the couplings. The teacher couplings are drawn as independent Gaussian random variables with zero mean and variance V given by $V = 1/N_{\mathrm{HID}}$ for J_{ib} and J_{ab}, $V = 1/N_{\mathrm{OBS}}$ for J_{ij} and J_{aj}. The initial value of the couplings was set to $J_{lm}^{\mathrm{initial}} = J_{lm}^{\mathrm{teacher}} + \mathcal{N}(0, \sqrt{V/5})$ for all indices l, m.

Appendix

6.A Details of the extended Plefka expansion

The Legendre transform of $\mathcal{L}_\alpha$ with respect to the fields $\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}$ is given by the following functional

$$\begin{aligned}\Gamma_\alpha[\mu, \hat{m}, \hat{\mu}, , B, \mathcal{B}] = \mathcal{L}_\alpha[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}] - \sum_{at} \psi_a(t)\mu_a(t) + i\sum_{it} H_i(t)\hat{m}_i(t) \\ + i\sum_{at} \mathcal{H}_a(t)\hat{\mu}_a(t) - \sum_{it} \hat{B}_i(t)B_i(t) - \sum_{at} \hat{\mathcal{B}}_a(t)\mathcal{B}_a(t),\end{aligned} \tag{6.21}$$

where $\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}$ are determined by extremizing Γ_α according to the following set of state equations:

$$\begin{aligned}\frac{\partial \Gamma_\alpha}{\partial \mu_a(t)} &= -\psi_a[\mu, \hat{m}, \hat{\mu}, , B, \mathcal{B}](t) \\ \frac{\partial \Gamma_\alpha}{\partial \hat{m}_i(t)} &= iH_i[\mu, \hat{m}, \hat{\mu}, , B, \mathcal{B}](t) \\ \frac{\partial \Gamma_\alpha}{\partial \hat{\mu}_a(t)} &= i\mathcal{H}_a[\mu, \hat{m}, \hat{\mu}, , B, \mathcal{B}](t) \\ \frac{\partial \Gamma_\alpha}{\partial B_i(t)} &= -\hat{B}_i[\mu, \hat{m}, \hat{\mu}, , B, \mathcal{B}](t) \\ \frac{\partial \Gamma_\alpha}{\partial \mathcal{B}_a(t)} &= -\hat{\mathcal{B}}_a[\mu, \hat{m}, \hat{\mu}, , B, \mathcal{B}](t).\end{aligned} \tag{6.22}$$

It is convenient for the following calculation to rewrite Γ_α as

$$\Gamma_\alpha = \log \int DG\, Tr_\sigma \exp[\Xi_\alpha], \tag{6.23}$$

where

$$\begin{aligned}\Xi_\alpha =& \sum_{it} s_i(t+1)g_i(t) + \sum_{at} \sigma_a(t+1)g_a(t) - \sum_{it} \log 2\cosh[g_i(t)] \\ & - \sum_{at} \log 2\cosh[g_a(t)] \\ & \sum_{it} i\hat{g}_i(t) \left[g_i(t) - \sum_j J_{ij}s_j(t) - \alpha \sum_b J_{ib}\sigma_b(t) \right] \\ & + \sum_{at} i\hat{g}_a(t) \left[g_a(t) - \sum_j J_{aj}s_j(t) - \alpha \sum_b J_{ab}\sigma_b(t) \right] \\ & + \sum_{at} \psi_a(t)(\sigma_a(t) - \mu_a(t)) \\ & - i\sum_{it} H_i(t)(\hat{g}_i(t) - \hat{m}_i(t)) - i\sum_{at} \mathcal{H}_a(t)(\hat{g}_a(t) - \hat{\mu}_a(t)) \\ & + \sum_{it} \hat{B}_i(t)\,(\hat{g}_i^2(t) - B_i(t)) + \sum_{at} \hat{\mathcal{B}}_a(t)\,(\hat{g}_a^2(t) - \mathcal{B}_a(t)).\end{aligned} \tag{6.24}$$

We then Taylor expand Γ_α around $\alpha = 0$ up to the second order:

$$\Gamma_\alpha = \Gamma_0 + \alpha\Gamma^{(1)} + \frac{\alpha}{2}\Gamma^{(2)}, \tag{6.25}$$

where we have defined $\Gamma^{(k)} = \partial\Gamma_\alpha/\partial\alpha^k|_{\alpha=0}$. The 0-th term of the expansion is given by

$$\begin{aligned}&\Gamma_0[\mu, \hat{m}, \hat{\mu}, B, \mathcal{B}] = \mathcal{L}_0[\psi^0, H^0, \mathcal{H}^0, \hat{B}^0, \hat{\mathcal{B}}^0] - \sum_{at} \psi_a^0(t)\mu_a(t) \\ &+ i\sum_{it} H_i^0(t)\hat{m}_i(t) + i\sum_{at} \mathcal{H}_a^0(t)\hat{\mu}_a(t) - \sum_{it} \hat{B}_i^0(t)B_i(t) - \sum_{at} \hat{\mathcal{B}}_a^0(t)\mathcal{B}_a(t),\end{aligned} \tag{6.26}$$

where

$$\begin{aligned}\mathcal{L}_0[\psi^0, H^0, \mathcal{H}^0, \hat{B}^0, \hat{\mathcal{B}}^0]s = \log \int DG\, Tr_\sigma \exp\Bigg\{ & \sum_{it} s_i(t+1)g_i(t) + \sum_{at} \sigma_a(t+1)g_a(t) \\ & - \sum_{it} \log 2\cosh[g_i(t)] - \sum_{at} \log 2\cosh[g_a(t)] \\ & \sum_{it} i\hat{g}_i(t)[g_i(t) - \sum_j J_{ij}s_j(t)] + \sum_{at} i\hat{g}_a(t)[g_a(t) - \sum_j J_{aj}s_j(t)] \\ & + \sum_{at} \psi^0_a(t)\sigma_a(t) - i\sum_{it} H^0_i(t)\hat{g}_i(t) - i\sum_{at} \mathcal{H}^0_a(t)\hat{g}_a(t) \\ & + \sum_{it} \hat{B}^0_i(t)\, \hat{g}^2_i(t) + \sum_{at} \hat{\mathcal{B}}^0_a(t)\, \hat{g}_a(t)^2 \Bigg\}.\end{aligned} \tag{6.27}$$

The fields $\psi^0, H^0, \mathcal{H}^0, \hat{B}^0, \hat{\mathcal{B}}^0$ are the ones that satisfy the set (6.22) for $\Gamma_\alpha = \Gamma_0$. The first derivative of Γ is computed from (6.23) as

$$\Gamma^{(1)} = \left\langle \frac{\partial \Xi_\alpha}{\partial \alpha} \right\rangle \Bigg|_{\alpha=0}. \tag{6.28}$$

Hence, from

$$\begin{aligned}\frac{\partial \Xi_\alpha}{\partial \alpha} = & -i\sum_{ibt} J_{ib}\hat{g}_i(t)\sigma_b(t) - i\sum_{abt} J_{ab}\hat{g}_a(t)\sigma_b(t) + \sum_{at} \frac{\partial \psi_a(t)}{\partial \alpha}[\sigma_a(t) - \mu_a(t)] \\ & - i\sum_{it} \frac{\partial H_i(t)}{\partial \alpha}[\hat{g}_i(t) - \hat{m}_i(t)] - i\sum_{at} \frac{\partial \mathcal{H}_a(t)}{\partial \alpha}[\hat{g}_a(t) - \hat{\mu}_a(t)] \\ & + \sum_{at} \frac{\partial \hat{B}_i(t)}{\partial \alpha}\,[\hat{g}^2_i(t) - B_i(t)] + \sum_{at} \frac{\partial \hat{\mathcal{B}}_a(t)}{\partial \alpha}\,[\hat{g}^2_a(t) - \mathcal{B}_a(t)]\end{aligned} \tag{6.29}$$

we obtain

$$\begin{aligned}\Gamma^{(1)} &= -i\sum_{ibt} J_{ib}\langle \hat{g}_i(t)\sigma_b(t)\rangle_0 - i\sum_{abt} J_{ab}\langle \hat{g}_a(t)\sigma_b(t)\rangle_0 \\ &= -i\sum_{ibt} J_{ib}\hat{m}_i(t)\mu_b(t) - i\sum_{abt} J_{ab}\hat{\mu}_a(t)\mu_b(t),\end{aligned} \tag{6.30}$$

where the last equality follows from equations 6.6 - 6.10 evaluated at zero couplings. The second order correction is computed from

$$\frac{\partial^2 \Gamma_\alpha}{\partial \alpha^2} = \left\langle \frac{\partial^2 \Xi_\alpha}{\partial \alpha^2} \right\rangle_\alpha + \left\langle \left(\frac{\partial \Xi_\alpha}{\partial \alpha} \right)^2 \right\rangle_\alpha - \left(\left\langle \frac{\partial \Xi_\alpha}{\partial \alpha} \right\rangle_\alpha \right)^2. \tag{6.31}$$

The first term on the right hand side of the above equation is zero, as one can see from (6.29) and (6.6 - 6.10). One thus finds

$$\Gamma^{(2)} = \left\langle \left(\frac{\partial \Xi_\alpha}{\partial \alpha} - \left\langle \frac{\partial \Xi_\alpha}{\partial \alpha} \right\rangle \right)^2 \right\rangle \Bigg|_{\alpha=0} . \tag{6.32}$$

To compute the above term we start from the set of equations

$$\begin{aligned}
\frac{\partial \psi_a(t)}{\partial \alpha}\Bigg|_{\alpha=0} &= -\frac{\partial}{\partial \mu_a(t)} \frac{\partial \Gamma_\alpha}{\partial \alpha}\Bigg|_{\alpha=0} = i \sum_i J_{ia} \hat{m}_i(t) + i \sum_b J_{ba} \hat{\mu}_b(t) \\
i \, \frac{\partial H_i(t)_a}{\partial \alpha}\Bigg|_{\alpha=0} &= \frac{\partial}{\partial \hat{m}_i(t)} \frac{\partial \Gamma_\alpha}{\partial \alpha}\Bigg|_{\alpha=0} = -i \sum_b J_{ib} \mu_b(t) \\
i \, \frac{\partial \mathcal{H}_a(t)}{\partial \alpha}\Bigg|_{\alpha=0} &= \frac{\partial}{\partial \hat{\mu}_a(t)} \frac{\partial \Gamma_\alpha}{\partial \alpha}\Bigg|_{\alpha=0} = -i \sum_b J_{ab} \mu_b(t),
\end{aligned} \tag{6.33}$$

and insert it in (6.29) to find:

$$\left[\frac{\partial \Xi_\alpha}{\partial \alpha} - \left\langle \frac{\partial \Xi_\alpha}{\partial \alpha} \right\rangle \right]_{\alpha=0} = -i \sum_a J_{ia} \delta \hat{g}_i(t) \delta \sigma_a(t) - i \sum_b J_{ab} \delta \hat{g}_a(t) \delta \sigma_b(t), \tag{6.34}$$

where we have defined $\delta \hat{g}_i(t) = \hat{g}_i(t) - \hat{m}_i(t)$, $\delta \hat{g}_a(t) = \hat{g}_a(t) - \hat{\mu}_a(t)$, $\delta \sigma_a(t) = \sigma_a(t) - \mu_a(t)$. From this result, using (6.32) and the definitions (6.6-6.10) we find:

$$\begin{aligned}
\Gamma^{(2)} = &- \sum_{iat} J_{ia}^2 \left(B_i(t) - \hat{m}_i^2(t) \right) \left(1 - \mu_a^2(t) \right) \\
&- \sum_{abt} J_{ab}^2 \left(\mathcal{B}_a(t) - \hat{\mu}_a^2(t) \right) \left(1 - \mu_b^2(t) \right) .
\end{aligned} \tag{6.35}$$

The idea of the extended Plefka Expansion is to describe the system in terms of an effective non-interacting log–likelihood, where the degrees of freedom are coupled to effective local fields. The equations for the effective local fields can be found by using (6.26, 6.28, 6.32) to compute Γ_α (6.25) expanded to the second order, and insert the latter quantity in the set of equations (6.22),

obtaining:

$$\begin{aligned}
\psi_a(t) &= \psi_a^0(t) + i\sum_i J_{ia}\hat{m}_i(t) + i\sum_b J_{ba}\hat{\mu}_b(t) - \sum_i J_{ia}^2\mu_a(t)\left(B_i(t) - \hat{m}_i^2(t)\right) \\
&\quad - \sum_b J_{ba}^2\mu_a(t)\left(\mathcal{B}_b(t) - \hat{\mu}_b^2(t)\right) \\
iH_i(t) &= iH_i^0(t) - i\sum_b J_{ib}\mu_b(t) + \sum_a J_{ia}^2\hat{m}_i(t)\left(1-\mu_a^2(t)\right) \\
i\mathcal{H}_a(t) &= i\mathcal{H}_a^0(t) - i\sum_b J_{ab}\mu_b(t) + \sum_b J_{ab}^2\hat{\mu}_a(t)\left(1-\mu_b^2(t)\right) \\
\hat{B}_i(t) &= \hat{B}_i^0(t) + \frac{1}{2}\sum_b J_{ia}^2\left(1-\mu_a^2(t)\right) \\
\hat{\mathcal{B}}_a(t) &= \hat{\mathcal{B}}_a^0(t) + \frac{1}{2}\sum_b J_{ab}^2\left(1-\mu_b^2(t)\right).
\end{aligned} \tag{6.36}$$

We can now identify the effective local fields with the values $\psi^0, H^0, \mathcal{H}^0, \hat{B}^0, \hat{\mathcal{B}}^0$ in (6.36), where we have set the auxiliary fields $\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}$ to zero. Inserting (6.36) in $\mathcal{L}_0$ (6.27), we find the final result for the effective single site log likelihood. Setting $\alpha = 1$ we obtain:

$$\begin{aligned}
&\mathcal{L}_0[\psi, H, \mathcal{H}, \hat{B}, \hat{\mathcal{B}}] = \log\int DG\, Tr_\sigma \exp\left\{\sum_{it} s_i(t+1)g_i(t) + \sum_{at}\sigma_a(t+1)g_a(t)\right. \\
&- \sum_{it}\log 2\cosh[g_i(t)] - \sum_{at}\log 2\cosh[g_a(t)] \\
&+ \sum_{at}\sigma_a(t)\left[\psi_a(t) - i\sum_i J_{ia}\hat{m}_i(t) - i\sum_b J_{ba}\hat{\mu}_b(t) + \sum_i J_{ia}^2\mu_a(t)\left(B_i(t) - \hat{m}_i^2(t)\right)\right. \\
&\qquad \left. + \sum_b J_{ba}^2\mu_a(t)\left(\mathcal{B}_b(t) - \hat{\mu}_b^2(t)\right)\right] \\
&+ \sum_{it} i\hat{g}_i(t)\left[g_i(t) - \sum_j J_{ij}s_j(t) - \sum_b J_{ib}\mu_b(t) - i\sum_a J_{ia}^2\hat{m}_i(t)\left(1-\mu_a^2(t)\right) - H_i(t)\right] \\
&+ \sum_{at} i\hat{g}_a(t)\left[g_a(t) - \sum_j J_{aj}s_j(t) - \sum_b J_{ab}\mu_b(t) - i\sum_b J_{ab}^2\hat{\mu}_a(t)\left(1-\mu_b^2(t)\right) - \mathcal{H}_a(t)\right] \\
&\left. - \frac{1}{2}\sum_{it}\hat{g}_i^2(t)\left[\sum_a J_{ia}^2\left(1-\mu_a^2(t)\right) - 2\hat{B}_i(t)\right] - \frac{1}{2}\sum_{at}\hat{g}_a^2(t)\left[\sum_b J_{ab}^2\left(1-\mu_b^2(t)\right) - 2\hat{\mathcal{B}}_a(t)\right]\right\}.
\end{aligned} \tag{6.37}$$

To simplify the above equation, we can linearize the quadratic terms by introducing two sets of Gaussian random variables: $\phi_i(t)$, independent for each

i,and $\phi_a(t)$, independent for each a, with zero mean and covariance respectively given by

$$\langle \phi_i(t)\phi_i(t)\rangle = \sum_a J_{ia}^2 \left(1-\mu_a^2(t)\right) - 2\hat{B}_i(t),$$
$$\langle \phi_a(t)\phi_a(t)\rangle = \sum_b J_{ab}^2 \left(1-\mu_b^2(t)\right) - 2\hat{\mathcal{B}}_a(t).$$

Thus, $\mathcal{L}_0$ can be written in a form where each spin is coupled to an effective stochastic local field in the following way:

$$\begin{aligned}
\mathcal{L}_0[\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}] = \log \int DG\, Tr_\sigma \exp\Bigg\{ & \sum_{it} s_i(t+1)g_i(t) + \sum_{at}\sigma_a(t+1)g_a(t)] \\
& - \sum_{it}\log 2\cosh[g_i(t)] - \sum_{at}\log 2\cosh[g_a(t)]\Bigg\} \\
\prod_{a,t}\exp\Bigg\{ & \sigma_a(t)\Bigg[\psi_a(t) - i\sum_i J_{ia}\hat{m}_i(t) - i\sum_b J_{ba}\hat{\mu}_b(t) + \sum_i J_{ia}^2\mu_a(t)\left(B_i(t)-\hat{m}_i^2(t)\right) \\
& + \sum_b J_{ba}^2\mu_a(t)\left(\mathcal{B}_b(t)-\hat{\mu}_b^2(t)\right)\Bigg]\Bigg\} \\
\prod_{i,t}\Bigg\langle \exp\Bigg\{ & i\hat{g}_i(t)\Bigg[g_i(t) - \phi_i(t) - \sum_j J_{ij}s_j(t) - \sum_b J_{ib}\mu_b(t) \\
& - i\sum_a J_{ia}^2\hat{m}_i(t)\left(1-\mu_a^2(t)\right) - H_i(t)\Bigg]\Bigg\}\Bigg\rangle_{\phi_i} \\
\prod_{a,t}\Bigg\langle \exp\Bigg\{ & i\hat{g}_a(t)\Bigg[g_a(t) - \phi_a(t) - \sum_j J_{aj}s_j(t) - \sum_b J_{ab}\mu_b(t) \\
& - i\sum_b J_{ab}^2\hat{\mu}_a(t)\left(1-\mu_b^2(t)\right) - \mathcal{H}_a(t)\Bigg]\Bigg\}\Bigg\rangle_{\phi_a},
\end{aligned} \tag{6.38}$$

which corresponds to (6.12) with the set of relations (6.14-6.18).

6.B Sparse observations

Let us consider a model composed of one set of spins, whose configuration is observed at the initial and at the final time step: $\{s_a(0)\}_{a=1}^N, \{s_a(T)\}_{a=1}^N$. The log likelihood is given by

$$\mathcal{L}_\alpha[\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}] = \log\int DG\, Tr_\sigma \exp\{\Omega_\alpha[\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}]\}, \tag{6.39}$$

where

$$\begin{aligned}\Omega_\alpha[\psi,H,\mathcal{H},\hat{B},\hat{\mathcal{B}}] &= \sum_a s_a(T)g_a(T-1) + \sum_a \sum_{t=0}^{t=T-2} \sigma_a(t+1)g_a(t) \\ &- \sum_a \sum_{t=0}^{t=T-1} \log 2\cosh[g_a(t)] + \sum_a \sum_{t=1}^{t=T-1} i\hat{g}_a(t)\left[g_a(t) - \alpha\sum_b J_{ab}\sigma_b(t)\right] \\ &+ i\hat{g}_a(0)\left[g_a(0) - \sum_j J_{ab}s_j(0)\right] + \sum_a \sum_{t=1}^{t=T-1} \psi_a(t)\sigma_a(t) \\ &- i\sum_a \sum_{t=1}^{t=T-1} \mathcal{H}_a(t)\hat{g}_a(t) + \sum_a \sum_{t=1}^{t=T-1} \hat{\mathcal{B}}_a(t)\,\hat{g}_a^2(t).\end{aligned} \tag{6.40}$$

By following the same steps of the calculation in Appendix 6.A, one finds the following result for the effective single site pseudo–likelihood:

$$\begin{aligned}\mathcal{L}_0 = \log\Bigg\{ &\prod_a \left\langle \frac{e^{s_a(T)h_a(T-1)}}{2\cosh\, h_a(T-1)} \right\rangle_{h_a(T-1)} \prod_a \prod_{t=1}^{T-2} Tr_{\sigma_a} \left\langle \frac{e^{\sigma_a(t+1)[h_a(t)+u_a(t+1)]}}{2\cosh\, h_a(t)} \right\rangle_{h_a(t)} \\ &\frac{\cosh\left[h_a(0)+u_a(1)\right]}{\cosh\, h_a(0)} \Bigg\}\end{aligned} \tag{6.41}$$

where $h_a(0) = \sum_b J_{ab}s_b(0)$ and for $t = 1 \dots T-1$ we have introduced the gaussian variables $h_a \sim \mathcal{N}(\gamma_a, V_a)$:

$$\begin{aligned} u_a(t) &= \psi_a(t) - i\sum_b J_{ba}\hat{\mu}_b(t) + \sum_b J_{ba}^2\mu_a(t)\left(\mathcal{B}_b(t) - \hat{\mu}_b^2(t)\right), \\ \gamma_a(t) &= \sum_b J_{ab}\mu_b(t) + i\sum_b J_{ab}^2\hat{\mu}_a(t)\left(1-\mu_b^2(t)\right) + \mathcal{H}_a(t), \\ V_a &= \langle\phi_a(t)\phi_a(t)\rangle = \sum_b J_{ab}^2\left(1-\mu_b^2(t)\right) - 2\hat{\mathcal{B}}_a(t). \end{aligned}$$

The TAP equations obtained from (6.6-6.10) by setting to zero the auxiliary fields at the end of the calculations are :

$$\mu_a(1) = \tanh[h_a(0) + u_a(1)],$$

$$\mu_a(t) = \frac{\left\langle \tanh[h_a(t-1)+u_a(t)]\frac{\cosh[h_a(t-1)+u_a(t)]}{\cosh h_a(t-1)} \right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t-1)+u_a(t)]}{\cosh h_a(t-1)} \right\rangle_{h_a}}, \quad t = 2\dots T-1$$

$$-i\hat{\mu}_a(t) = \mu_a(t+1) - \frac{\left\langle \tanh h_a(t) \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}}, \quad t = 1 \dots T-2$$

$$-i\hat{\mu}_a(T-1) = s_a(T) - \frac{\left\langle \tanh h_a(T-1) \frac{e^{s_a(T)h_a(T-1)}}{\cosh h_a(T-1)} \right\rangle_{h_a(T-1)}}{\left\langle \frac{e^{s_a(T)h_a(T-1)}}{\cosh h_a(T-1)} \right\rangle_{h_a(T-1)}}$$

$$\begin{aligned}\mathcal{B}_a(t) =& \frac{\left\langle \left(1-\tanh^2 h_a(t)\right) \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}} \\ &- \frac{Tr_{\sigma_a} \left\langle (\sigma_a(t+1) - \tanh h_a(t))^2 \frac{e^{\sigma_a(t+1)[h_a(t)+u_a(t+1)]}}{2\cosh h_a(t)} \right\rangle_{h_a}}{\left\langle \frac{\cosh[h_a(t)+u_a(t+1)]}{\cosh h_a(t)} \right\rangle_{h_a}}, \quad t = 1 \dots T-2\end{aligned}$$

$$\begin{aligned}\mathcal{B}_a(T-1) =& \frac{\left\langle \left(1-\tanh^2 h_a(t)\right) \frac{e^{s_a(T)h_a(T-1)}}{\cosh h_a(T-1)} \right\rangle_{h_a(T-1)}}{\left\langle \frac{e^{s_a(T)h_a(T-1)}}{\cosh h_a(T-1)} \right\rangle_{h_a(T-1)}} \\ &- \frac{\left\langle (s_a(t+1) - \tanh h_a(t))^2 \frac{e^{s_a(T)h_a(T-1)}}{\cosh h_a(T-1)} \right\rangle_{h_a(T-1)}}{\left\langle \frac{e^{s_a(T)h_a(T-1)}}{\cosh h_a(T-1)} \right\rangle_{h_a(T-1)}}.\end{aligned}$$

7 Summary and outlook

In this thesis, we combined ideas of statistical physics and machine learning to study learning and inference in the Ising model, with a major focus on its Glauber dynamics with parallel update rule. We showed how methods borrowed from the statistical mechanics of spin glasses can be used both to study analytically the average case scenario and to implement solutions for single instances of the problem. We have restricted our analysis to large, densely connected systems, with weak random couplings. First, this restriction has allowed us to better understand the dynamics of the kinetic model, by deriving a new mean-field solution valid for any degree of symmetry of the network, via an extension of the Plefka expansion that includes second order statistics in the formalism. The novel feature of our result is a memory term appearing in the equation for the effective local fields, coupling each variable to all its past values. The computation of these complex fields via a Monte Carlo algorithm makes the method more computationally demanding with respect to other mean field solutions, but more accurate in predicting single site magnetizations. In fact, we conjecture that it provides the exact result for the marginal distribution of spin trajectories, in the thermodynamic limit of an infinitely large system. We leave the rigorous prove of this conjecture to future work; one strategy would be to average our equation for the magnetization over many instances of the network, and compare the result with the exact dynamical mean field theory for the disordered average dynamics [EO94]. An open question remains how to design an inference algorithm based on the extended Plefka expansion, for learning the couplings between the spins based on a set of observed trajectories.

We discussed the applicability of mean field approaches as inference tools for dynamic data in the case of an asymmetric network, where the exact mean field solution is known and a simple linear estimator based on this solution has recently been proposed.

We compared the performance of this linear estimator to the asymptotic performance of the maximum likelihood estimator, which has the property of asymptotic efficiency; we also analysed the Bayes optimal estimator, which is is asymptotically optimal if the prior corresponds to the distribution of the true parameters. Working in the student-teacher scenario, we computed the estimation error as a function of the size of the data set using the replica method. The

Markovian dynamics, where transition probabilities are normalized individually, and the fast decay of two-times correlations in the asymmetric network allow to treat the distribution of the data in a simplified scheme in the limit of a very large system: using arguments based on a central limit theorem, we remapped the problem of learning from one set of observed trajectories of T time steps into the problem of learning in T independent perceptrons; in each perceptron, however, the inputs are not independent but correlated through the equal-time correlation matrix. We derived an exact result for the statistics of the correlation matrix and consequently for the estimation error.

In the large $\alpha = T/N$ limit, the error of linear estimators is close to optimal only for weak couplings, whereas it deviates from optimality for stronger couplings. At finite values of α, the linear mean field estimator (the error of which diverges when $\alpha \to 1$) is outperformed by the optimal linear estimator. If the correct prior knowledge on the distribution of the couplings is introduced, the Bayes optimal estimator outperforms all the considered methods: we derived a novel approximate algorithm to implement it, of the expectation propagation type; it is computationally faster than maximum likelihood and it would be interesting to understand whether it gives the exact solution in the large N limit.

Our analysis for the error can be used as a benchmark for applications when selecting a specific algorithm, and we found it relevant to extend it also to the equilibrium case, where the exact computation of statistically efficient estimators, such as the maximum likelihood estimator, is computationally intractable for large systems. The intractability of the partition function also introduces extra complexity in the replica calculation; for couplings that are learnt independently for each spin using local cost functions, we tackled the problem using cavity arguments that are valid as long as the Ising system is in the paramagnetic phase. As in the kinetic case, the result explicitly shows the influence of the correlation matrix on the estimation error. In contrast, the picture is surprisingly simpler, in that a mean field approximation to the maximum likelihood estimator is asymptotically optimal and outperforms the the widely used method of pseudo-likelihood. Moreover, the local cost function that achieves minimal error is of quadratic form. The explicit optimal local estimator based on its minimization is symmetric, which poses the question whether it also represent a global optimal estimator; it can be simply computed by inverting the empirical correlation matrix, and depends on a parameter that can be fully estimated from data. It would be interesting to study how the results change at small temperatures, namely the spin glass phase, where a more complex form of our cavity arguments has to be employed and the non-ergodic behavior of the system has to be taken into account (see

section 5.4).

Both for the kinetic and the equilibrium case, our results for the error are exact in the thermodynamic limit, but agree very well with simulations of finite systems. In order to provide a theoretical framework for comparing the performance of various learning algorithms, we have restricted our analysis to specific statistical ensembles of the teacher couplings, that allowed to derive an analytical solution for the average case scenario, by using the cavity and the replica methods of statistical physics. However, we have also proposed two new algorithms for inferring the couplings that are valid for the single instances of the network, namely the Bayes optimal estimator computed via an iterative algorithm for the kinetic case, and the optimal local estimator for the equilibrium case: while we have tested their performance on simulated data, the next step that we leave to future work is to apply such inference algorithms to real-world data, where we don't have prior knowledge on the distribution of the underlying network of couplings.

With the aim of applying our analysis to network reconstruction from real-world data, it would be also relevant to study sparse networks. Biological networks are indeed typically sparse, and using learning algorithms that are targeted to dense networks might lead to over-fitting, some of the inferred interactions reproducing noise rather than the true interactions. For the algorithm and the analysis performed in a Bayesian setting, our formalism can be applied to the sparse case by choosing a suitable prior over the couplings, for instance a spike-and-slab prior (see section 4.5). For algorithms based on a cost function minimization, the sparsity of the coupling matrix can be controlled by adding a regularization term to the cost function. The computation of the estimation error gets more complicated in this scenario: while, in the present analysis, the correlation matrix affects the result for the error via the trace of the inverse matrix, when a quadratic regularization term is added to the cost function preliminary results show that the error becomes a more complicated function of the matrix eigenvalues.

Another important aspect to bear in mind when addressing the problem of reconstructing real-world networks is that typically only a tiny part of the system is experimentally accessible. In the last part of the dissertation, we have studied a recently-proposed extension of the kinetic Ising model, where a subset of the spin trajectories is observed while the rest is hidden.

Current algorithms for learning the couplings in this model are of the expectation maximization (EM) type, and iterate between computing the expected value of the hidden units at fixed parameters (E-step), and updating the parameters to maximize the expected log-likelihood (M-step). Such algorithms typically fail if the number of hidden units is too large (i.e., more than the 10%

of the total units), presumably due to the error of predicting the hidden-spin values in the E-step.

We provided a theoretical analysis for this problem, which can be used as a yardstick for the applicability of this class of algorithms, by assessing the theoretically optimal performance for predicting hidden spins. This has been achieved by computing the average error of the Bayes optimal prediction of hidden spin states, as a function of the the fraction of hidden spins and the strength of the couplings, via a replica formalism.

We also studied single instances of the network: we applied our extended Plefka expansion to derive a mean field result for the magnetization of hidden spins that very well agrees with simulations of finite systems. This mean field result can be used to address two different problems. One is the analysis of time series data where observations are sparse in time, and the goal is to predict the state of the spins at unobserved time steps; the other is the design of a novel algorithm of the EM type to learn the coupling of a kinetic Ising model with hidden nodes. We have discussed the main challenges that one faces when addressing these problems, which appear to be interesting follow-up research directions of our work.

We believe that our multi-disciplinary approach - at the intersection between computer science, statistical physics and Bayesian statistics - contributed to give new insights in the theoretical understanding of inverse problems on random networks and provided new tools for implementing algorithmic solutions. Many new questions have also been raised in the dissertation, which appear to be fruitful avenues for future research.

Bibliography

[AE12] Erik Aurell and Magnus Ekeberg. Inverse Ising inference using all the data. *Physical review letters*, 108(9):090201, 2012.

[AG16] Madhu Advani and Surya Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

[AHS85] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[AM12] Erik Aurell and Hamed Mahmoudi. Dynamic mean-field and cavity methods for diluted Ising systems. *Physical Review E*, 85(3):031119, 2012.

[BBB+03] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Bayesian factor regression models in the large p, small n paradigm. *Bayesian statistics*, 7:733–742, 2003.

[BCG+12] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

[BCKM98] Jean-Philippe Bouchaud, Leticia F Cugliandolo, Jorge Kurchan, and Marc Mezard. Out of equilibrium dynamics in spin-glasses and other glassy systems. *Spin glasses and random fields*, pages 161–223, 1998.

[BDR17] Claudia Battistin, Benjamin Dunn, and Yasser Roudi. Learning with unknowns: analyzing biological data in the presence of hidden variables. *Current Opinion in Systems Biology*, 2017.

[BDT+07] Tamara Broderick, Miroslav Dudik, Gasper Tkacik, Robert E Schapire, and William Bialek. Faster solutions of the inverse pairwise Ising problem. *arXiv preprint arXiv:0712.2437*, 2007.

[Ber16] Johannes Berg. Statistical mechanics of the inverse Ising problem and the optimal objective function. *arXiv preprint arXiv:1611.04281*, 2016.

[Bes74] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[Bet35] Hans A Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 150(871):552–575, 1935.

[BHTR15] Claudia Battistin, John Hertz, Joanna Tyrcha, and Yasser Roudi. Belief propagation and replicas for inference and learning in a kinetic Ising model with hidden spins. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(5):P05021, 2015.

[Bir99] Giulio Biroli. Dynamical tap approach to mean field glassy systems. *Journal of Physics A: Mathematical and General*, 32(48):8365, 1999.

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[BM80] AJ Bray and MA Moore. Metastable states in spin glasses. *Journal of Physics C: Solid State Physics*, 13(19):L469, 1980.

[BS16] Barbara Bravi and Peter Sollich. Inference for dynamics of continuous variables: the Extended Plefka Expansion with hidden nodes. *arXiv preprint arXiv:1603.05538*, 2016.

[BSO16] Barbara Bravi, Peter Sollich, and Manfred Opper. Extended plefka expansion for stochastic dynamics. *Journal of Physics A: Mathematical and Theoretical*, 49(19):194003, 2016.

[BTW+08] Barrett Tanya, Troup Dennis B, Wilhite Stephen E, Ledoux Pierre, Rudnev Dmitry, Evangelista Carlos, Kim Irene F, Soboleva Alexandra, Tomashevsky Maxim, Marshall Kimberly A, Phillippy Katherine H, Sherman Patti M, Muertter Rolf N, and Edgar Ron. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(Database issue):D885–D890, oct 2008.

[Cav09] Andrea Cavagna. Supercooled liquids for pedestrians. *Physics Reports*, 476(4):51–124, 2009.

[CFG+15] Cristiano Capone, Carla Filosa, Guido Gigante, Federico Ricci-Tersenghi, and Paolo Del Giudice. Inferring synaptic structure in presence of neural interaction time scales. *PloS one*, 10(3):e0118412, 2015.

[CLM09] Simona Cocco, Stanislas Leibler, and Rémi Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

[CLS96] ACC Coolen, SN Laughton, and D Sherrington. Dynamical replica theory for disordered spin systems. *Physical Review B*, 53(13):8184, 1996.

[CM11] Simona Cocco and Rémi Monasson. Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Physical review letters*, 106(9):090601, 2011.

[CM12] Simona Cocco and Rémi Monasson. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. *Journal of Statistical Physics*, 147(2):252–314, 2012.

[Coo01] ACC Coolen. Statistical mechanics of recurrent neural networks ii:dynamics. *Handbook of biological physics*, 4:619–684, 2001.

[Cra16] Harald Cramér. *Mathematical Methods of Statistics (PMS-9)*, volume 9. Princeton university press, 2016.

[CS87] A Crisanti and Haim Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Physical Review A*, 36(10):4922, 1987.

[CS88] A Crisanti and Haim Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Ising spins and glauber dynamics. *Physical Review A*, 37(12):4865, 1988.

[CS93] ACC Coolen and D Sherrington. Dynamics of fully connected attractor neural networks near saturation. *Physical review letters*, 71(23):3886, 1993.

[CS94] ACC Coolen and D Sherrington. Order-parameter flow in the sk spin glass. i. replica symmetry. *Journal of Physics A: Mathematical and General*, 27(23):7687, 1994.

[Cug03] L Cugliandolo. Course 7: Dynamics of glassy systems. *Slow Relaxations and nonequilibrium dynamics in condensed matter*, pages 161–171, 2003.

[DAT78] JRL De Almeida and David J Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 1978.

[DB16] Benjamin Dunn and Claudia Battistin. The appropriateness of ignorance in the inverse kinetic Ising model. *arXiv preprint arXiv:1612.06185*, 2016.

[DDP78] C De Dominicis and L Peliti. Field-theory renormalization and critical dynamics above t c: Helium, antiferromagnets, and liquid-gas systems. *Physical Review B*, 18(1):353, 1978.

[DMR15] Benjamin Dunn, Maria Mørreaunet, and Yasser Roudi. Correlations and functional connections in a population of grid cells. *PLoS computational biology*, 11(2):e1004052, 2015.

[DR13] Benjamin Dunn and Yasser Roudi. Learning and inference in a nonequilibrium Ising model with hidden nodes. *Physical Review E*, 87(2):022127, 2013.

[EA75] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[Edw70] SF Edwards. Proceedings of the third international conference on amorphous materials. 1970.

[ELL+13] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.

[EO94] H. Eissfeller and M. Opper. Mean-field monte carlo approach to the sherrington-kirkpatrick model with asymmetric couplings. *Phys. Rev. E*, 50:709–720, Aug 1994.

[EVdB01] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[Fer16] Ulisse Ferrari. Learning maximum entropy models from finite-size data sets: A fast data-driven algorithm allows sampling from the posterior distribution. *Physical Review E*, 94(2):023301, 2016.

[FS88] Alan M Ferrenberg and Robert H Swendsen. New monte carlo technique for studying phase transitions. *Physical review letters*, 61(23):2635, 1988.

[Gar87] Elizabeth Gardner. Maximum storage capacity in neural networks. *EPL (Europhysics Letters)*, 4(4):481, 1987.

[Gar88] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[GATSS13] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-dependent maximum entropy models of neural population codes. *PLoS computational biology*, 9(3):e1002922, 2013.

[GBS09] Dongning Guo, Dror Baron, and Shlomo Shamai. A single-letter characterization of optimal noisy compressed sensing. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 52–59. IEEE, 2009.

[GFM16] Silvia Grigolon, Silvio Franz, and Matteo Marsili. Identifying relevant positions in proteins by Critical Variable Selection. *Molecular BioSystems*, 12(7):2147–2158, 2016.

[GKG+13] Felipe Gerhard, Tilman Kispersky, Gabrielle J Gutierrez, Eve Marder, Mark Kramer, and Uri Eden. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. *PLoS computational biology*, 9(7):e1003138, 2013.

[Gla63] Roy J Glauber. Time-dependent statistics of the Ising model. *Journal of mathematical physics*, 4(2):294–307, 1963.

[GM93] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[GS10] Surya Ganguli and Haim Sompolinsky. Statistical mechanics of compressed sensing. *Physical review letters*, 104(18):188701, 2010.

[GSS11] Elad Ganmor, Ronen Segev, and Elad Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences*, 108(23):9679–9684, 2011.

[Gue03] Francesco Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

[GV05] Dongning Guo and Sergio Verdú. Randomly spread cdma: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, 2005.

[HKP91] John A Hertz, Anders S Krogh, and Richard G Palmer. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.

[HLP34] GH Hardy, JE Littlewood, and G Pólya. Inequalities university press, 1934.

[HM15] Ariel Haimovici and Matteo Marsili. Criticality of mostly informative samples: a Bayesian model selection approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(10):P10013, 2015.

[Hua87] Kerson Huang. *Statistical Mechanics*. John Wiley & Sons, 2 edition, 1987.

[Hua15] Haiping Huang. Effects of hidden nodes on network structure inference. *Journal of Physics A: Mathematical and Theoretical*, 48(35):355002, 2015.

[Jay57] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[Kac68] M Kac. Trondheim theoretical physics seminar. *Nordita Publication*, (286), 1968.

[KKC98] SM Kuva, Osame Kinouchi, and Nestor Caticha. Learning a spin glass: Determining hamiltonians from metastable states. *Physica A: Statistical Mechanics and its Applications*, 257(1):28–35, 1998.

[KKM+16] Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016.

[KMS+12] Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

[KS00] HJ Kappen and JJ Spanjers. Mean field theory for asymmetric neural networks. *Physical Review E*, 61(5):5658, 2000.

[LBC+06] Timothy R Lezon, Jayanth R Banavar, Marek Cieplak, Amos Maritan, and Nina V Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences*, 103(50):19033–19038, 2006.

[LCRP14] Kenneth W Latimer, EJ Chichilnisky, Fred Rieke, and Jonathan W Pillow. Inferring synaptic conductances from spike trains with a biophysically inspired point process model. In *Advances in Neural Information Processing Systems*, pages 954–962, 2014.

[LL69] Lev Davidovich Landau and Evgenii M Lifshitz. *Statistical Physics: V. 5: Course of Theoretical Physics*. Pergamon press, 1969.

[LTS90] Esther Levin, Naftali Tishby, and Sara A Solla. A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, 78(10):1568–1574, 1990.

[MB88] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[MDP14] Alexander Mozeika, Onur Dikmen, and Joonas Piili. Consistent inference of a general model using the pseudolikelihood method. *Physical Review E*, 90(1):010101, 2014.

[Min01] Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[MK07] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[MM09] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MM11] Cristopher Moore and Stephan Mertens. *The nature of computation*. OUP Oxford, 2011.

[MMR13] Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. On sampling and modeling complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09003, 2013.

[MP01] Marc Mézard and Giorgio Parisi. The bethe lattice spin glass revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20(2):217–233, 2001.

[MPS+84] Marc Mézard, Giorgio Parisi, Nicolas Sourlas, G Toulouse, and Miguel Virasoro. Nature of the spin-glass phase. *Physical review letters*, 52(13):1156, 1984.

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Co Inc, 1987.

[MS11] M Mézard and J Sakellariou. Exact mean-field inference in asymmetric kinetic Ising systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(07):L07001, 2011.

[MS14] Hamed Mahmoudi and David Saad. Generalized mean field approximation for parallel dynamics of the Ising model. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(7):P07001, 2014.

[MSR73] Paul Cecil Martin, ED Siggia, and HA Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973.

[MV85] Marc Mézard and Miguel Angel Virasoro. The microstructure of ultrametricity. *Journal de Physique*, 46(8):1293–1307, 1985.

[MVK10] Enzo Marinari and Valery Van Kerrebroeck. Intrinsic limitations of the susceptibility propagation inverse inference for the mean field Ising spin glass. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(02):P02008, 2010.

[MY82] ND Mackenzie and AP Young. Lack of ergodicity in the infinite-range Ising spin-glass. *Physical Review Letters*, 49(5):301, 1982.

[NB12a] H Chau Nguyen and Johannes Berg. Bethe–peierls approximation and the inverse Ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03004, 2012.

[NB12b] H Chau Nguyen and Johannes Berg. Mean-field theory for the inverse Ising problem at low temperatures. *Physical review letters*, 109(5):050602, 2012.

[Nic08] Miguel A L Nicolelis. *Methods for neural ensemble recordings*. CRC Press, 2 edition, 2008.

[Nis01] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[NY96] Hidetoshi Nishimori and Michiko Yamana. Dynamical probability distribution function of the sk model at high temperatures. *Journal of the Physical Society of Japan*, 65(1):3–6, 1996.

[NZB17a] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

[NZB17b] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *arXiv preprint arXiv:1702.01522*, 2017.

[ODB+14] Obien Marie Engelene J, Deligkaris Kosmas, Bullmann Torsten, Bakkum Douglas J, and Frey Urs. Revealing neuronal function through microelectrode array recordings. *Frontiers in Neuroscience*, 8:423, dec 2014.

[OK96] Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. *Models of neural networks III*, pages 151–209, 1996.

[OS01] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[OW00] Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

[OW01a] Manfred Opper and Ole Winther. Adaptive and self-averaging thouless-anderson-palmer mean-field theory for probabilistic modeling. *Physical Review E*, 64(5):056131, 2001.

[OW01b] Manfred Opper and Ole Winther. From naive mean field theory to the tap equations. *Advanced Mean Field Methods: Theory and Practice*, page 7, 2001.

[Par80a] Giorgio Parisi. Magnetic properties of spin glasses in a new mean field theory. *Journal of Physics A: Mathematical and General*, 13(5):1887, 1980.

[Par80b] Giorgio Parisi. A sequence of approximated solutions to the sk model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.

[Par83] Giorgio Parisi. Order parameter for spin-glasses. *Physical Review Letters*, 50(24):1946, 1983.

[Par06] Giorgio Parisi. Spin glasses and fragile glasses: Statics, dynamics, and complexity. *Proceedings of the National Academy of Sciences*, 103(21):7948–7955, 2006.

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[Pei36] Rudolf Peierls. On Ising's model of ferromagnetism. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 477–481. Cambridge University Press, 1936.

[Per84] P Peretto. Collective properties of neural networks: a statistical physics approach. *Biological cybernetics*, 50(1):51–62, 1984.

[Ple82] T Plefka. Convergence condition of the tap equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, 1982.

[PP79] RG Palmer and CM Pond. Internal field distributions in model spin glasses. *Journal of Physics F: Metal Physics*, 9(7):1451, 1979.

[RGF09] Sundeep Rangan, Vivek Goyal, and Alyson K Fletcher. Asymptotic analysis of map estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems*, pages 1545–1553, 2009.

[RH11a] Yasser Roudi and John Hertz. Dynamical tap equations for non-equilibrium Ising spin glasses. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(03):P03031, 2011.

[RH11b] Yasser Roudi and John Hertz. Mean field theory for nonequilibrium network reconstruction. *Physical review letters*, 106(4):048702, 2011.

[RT15] Yasser Roudi and Graham Taylor. Learning with hidden variables. *Current opinion in neurobiology*, 35:110–118, 2015.

[SBIB06] Elad Schneidman, Michael J Berry, Ronen Segev II, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007, 2006.

[SDBD09] Jascha Sohl-Dickstein, Peter Battaglino, and Michael R DeWeese. Minimum probability flow learning. *arXiv preprint arXiv:0906.4779*, 2009.

[SFG+06] Jonathon Shlens, Greg D Field, Jeffrey L Gauthier, Matthew I Grivich, Dumitru Petrusca, Alexander Sher, Alan M Litke, and EJ Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *Journal of Neuroscience*, 26(32):8254–8266, 2006.

[SK75] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[SM09] Vitor Sessak and Rémi Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42(5):055001, 2009.

[Som87] Hans-Jürgen Sommers. Path-integral approach to Ising spin-glass dynamics. *Physical review letters*, 58(12):1268, 1987.

[SRMH12] Jason Sakellariou, Yasser Roudi, Marc Mezard, and John Hertz. Effect of coupling asymmetry on mean-field solutions of the direct and inverse sherrington-kirkpatrick model. *Philosophical Magazine*, 92(1-3):272–279, 2012.

[SST92] HS Seung, Haim Sompolinsky, and N Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056, 1992.

[Tal06] Michel Talagrand. The parisi formula. *Annals of mathematics*, pages 221–263, 2006.

[Tan00] Toshiyuki Tanaka. Information geometry of mean-field approximation. *Neural Computation*, 12(8):1951–1968, 2000.

[TAP77] David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[TH13] Joanna Tyrcha and John Hertz. Network inference with hidden units. *arXiv preprint arXiv:1301.7274*, 2013.

[TJH+08] Aonan Tang, David Jackson, Jon Hobbs, Wei Chen, Jodi L Smith, Hema Patel, Anita Prieto, Dumitru Petrusca, Matthew I Grivich, Alexander Sher, et al. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *Journal of Neuroscience*, 28(2):505–518, 2008.

[TMM+14] Gasper Tkacik, Thierry Mora, Olivier Marre, Dario Amodei, II Berry, J Michael, and William Bialek. Thermodynamics for a network of neurons: Signatures of criticality. *arXiv preprint arXiv:1407.5946*, 2014.

[TRMH13] Joanna Tyrcha, Yasser Roudi, Matteo Marsili, and John Hertz. The effect of nonstationarity on models inferred from neural data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03005, 2013.

[VIS15] Maxim Volgushev, Vladimir Ilin, and Ian H Stevenson. Identifying and tracking simulated synaptic inputs from neuronal firing: insights from in vitro experiments. *PLoS computational biology*, 11(3):e1004167, 2015.

[VMLC16] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[WRB93] Timothy LH Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

[Wu83] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[WWS+09] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[ZJ02] Jean Zinn-Justin. Quantum field theory and critical phenomena. 2002.