Statistical physics of pairwise probability models

Yasser Roudi[1]*, Erik Aurell[2,3] and John A. Hertz[1,4]

[1] NORDITA, Stockholm, Sweden
[2] Department of Computational Biology, AlbaNova University Centre, Stockholm, Sweden
[3] ACCESS Linnaeus Centre, KTH-Royal Institute of Technology, Stockholm, Sweden
[4] The Niels Bohr Institute, Copenhagen University, Copenhagen, Denmark

Statistical models for describing the probability distribution over the states of biological systems are commonly used for dimensional reduction. Among these models, pairwise models are very attractive in part because they can be fit using a reasonable amount of data: knowledge of the mean values and correlations between pairs of elements in the system is sufficient. Not surprisingly, then, using pairwise models for studying neural data has been the focus of many studies in recent years. In this paper, we describe how tools from statistical physics can be employed for studying and using pairwise models. We build on our previous work on the subject and study the relation between different methods for fitting these models and evaluating their quality. In particular, using data from simulated cortical networks we study how the quality of various approximate methods for inferring the parameters in a pairwise model depends on the time bin chosen for binning the data. We also study the effect of the size of the time bin on the model quality itself, again using simulated data. We show that using finer time bins increases the quality of the pairwise model. We offer new ways of deriving the expressions reported in our previous work for assessing the quality of pairwise models.

Keywords: pairwise model, inverse ising problem, maximum entropy distribution, inference

INTRODUCTION

In biological networks the collective dynamics of thousands to millions of interacting elements, generating complicated spatiotemporal structures, is fundamental for the function. Until recently, our understanding of these structures was severely limited by technical difficulties for simultaneous measurements from a large number of elements. Recent technical developments, however, are making it more and more common that experimentalists record data from larger and larger parts of the system. A living example of this story is what is now happening in Neuroscience. Until a few years ago, neurophysiology meant recording from a handful of neurons at a time when studying early stages of sensory processing (e.g. the retina), and only single neurons when studying advanced cortical areas (e.g. inferotemporal cortex). This limit is now rapidly going away, and people are collecting data from larger and larger populations (see e.g. the contributions in Nicolelis, 2007). However, such data will not help us understand the system *per se*. It will only give us large collections of numbers, and most probably we will have the same problem that we had with the system, but now with the data set: we can only look at small parts of it at a time, without really making use of its collective structures. To use such data and make sense of it we need to find systematic ways to build mathematically tractable descriptions of the data, descriptions with lower dimensionality than the data itself, but involving relevant dimensions.

In recent years, binary pairwise models have attracted a lot of attention as parametric models for studying the statistics of spike trains of neuronal populations (Schneidman et al., 2006; Shlens et al., 2006, 2009; Tang et al., 2008; Roudi et al., 2009a,b), the statistics of natural images (Bethge and Berens, 2008), and inferring

neuronal functional connectivities (Shlens et al., 2006; Yu et al., 2008). These models are the simplest parametric models that can be used to describe correlated neuronal firing. To build them one only needs the knowledge of the distribution of spike probability over pairs of neurons, and, therefore, these models can be fit using reasonable amounts of data. When dealing with statistical models, there are two main questions that one needs to answer. The first question is the inference question, i.e. how one can find the best parameters for the model given the data. The second question is the question of model quality, i.e. once the best parameters are found, how good the fitted model is for describing the statistics of the data. These questions can be naturally studied in the framework provided by statistical physics. Starting from a probability distribution over the states of a number of elements, statistical physics deals with computing quantities such as correlation functions, entropy, free energy, energy minima, etc., through exact or carefully developed approximate methods. It allows one to give quantitative answers to questions such as: what is the entropy difference between real data and the parametric model? Is there a closed-form equation relating model parameters to the correlation functions? How well does a pairwise model approximate higher order statistics of the data?

This paper has three main purposes: (a) to provide a summary of the previous results, derived using statistical physics, on the parameter inference and quality of pairwise models, (b) to study how the choice of the time bin size affects the inference and model quality problems and, (c) to report new and somewhat simpler derivations of our previous results.

In building binary pairwise models a crucial step is binning the spike trains into small time bins and assigning −1 or 1 to each bin depending on whether there is a spike in it or not. The answers to

the inference and model quality problems may thus be affected by the choice of this time bin size. In what follows, we first deal with the effect of the size of the time bin on the answer of the inference question. We review a number of approximate methods described in Roudi et al. (2009b) for inferring the parameters of the pairwise model and study how their quality depends on the choice of the bin size. We also describe a new approximation, the independent-triplet (IT) approximation, as an extension to the independent-pair (IP) approximation described in Roudi et al. (2009b). We show that although IT slightly outperforms IP, it cannot beat the more powerful approximations that we describe, i.e. the Thouless–Anderson–Palmer (TAP) inversion, and the Sessak–Monasson (SM) approximation. We also derive the TAP inversion approximation from the celebrated Belief Propagation algorithm. After this, we study the effect of the size of the time bin on the answer to the model quality question. This issue has been studied analytically in Roudi et al. (2009a), using a perturbative expansion. They derived equations for the difference between the entropies of the pairwise model and the true distribution of the data, as well as between an independent-neuron model and the data, when $N\bar{\nu}\delta t \to 0$ (the so called "perturbative regime" or the "low-rate regime"). Here, N is the number of neurons, $\bar{\nu}$ is the average firing rate and δt is the size of the time bin. They concluded that the quality of the pairwise model increases by decreasing δt or N. In this paper, we confirm these analytical results on data generated from a model cortical network.

A crucial step in the derivation of entropy differences in Roudi et al. (2009a) was to express the pairwise distribution in the so called Sarmanov–Lancaster representation (Lancaster, 1958; Sarmanov, 1963). Here we show that one can derive the same expressions by approximating the partition function of a Gibbs distribution (Eq. 2 in Section "The Binary Pairwise Model"). In addition to recovering the low-rate expansion of the entropies, we use the results to find the difference between the probability of synchronous spikes according to the true distribution, the pairwise model and the independent model. We also show that one can derive the results of Roudi et al. (2009a) by extending the idea behind the IP approximation to triplets of neuron, i.e. the IT approximation. By taking the limit $\nu_i\delta t \to 0$ of the IT approximation to the entropy of a given distribution, where ν_i is the firing rate of the ith neuron, we show that the results of Roudi et al. (2009a) can be recovered in a considerably simpler way.

In the last part of the paper, we discuss two possible extensions of the binary pairwise models and mention a number of important questions that they raise. We first describe the extensions to non-binary variables useful for studying the statistics of modular models of cortical networks. We then describe an extension of the pairwise model to a model with asymmetric connections which gives promising results for discovering the synaptic connectivity from neural spike trains.

THE BINARY PAIRWISE MODEL

In a binary pairwise model, starting from the spikes recorded from N neurons, one first divides the spike trains into small time bins. One then builds a binary representation of the spike trains by assigning a binary spin variable $s_i(t)$ to each neuron i and each time bin t, with $s_i(t) = -1$ if neuron i has not emitted any spikes

in that time bin, and $s_i(t) = -1$ if it has emitted one spike or more. From this binary representation, the mean values and correlations between the neurons are computed as

$$\left\langle s_i \right\rangle_{\text{data}} \equiv \frac{1}{T}\sum_t s_i(t) \tag{1a}$$

$$\left\langle s_i s_j \right\rangle_{\text{data}} \equiv \frac{1}{T}\sum_t s_i(t)s_j(t), \tag{1b}$$

where T is the total number of time bins.

The binary pairwise model of the data is then built by considering the following distribution over a set of N binary variable $\mathbf{s} = (s_1, s_2, \ldots, s_N)$

$$p_{\text{pair}}(\mathbf{s}) = \frac{1}{Z}\exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right], \tag{2}$$

and choosing the parameters, h_i and J_{ij} such that the mean values and pairwise correlations under this distribution matches those of the data defined in Eq. 1, that is

$$\left\langle s_i \right\rangle_{\text{pair}} \equiv \sum_{\mathbf{s}} p_{\text{pair}}(\mathbf{s})s_i = \left\langle s_i \right\rangle_{\text{data}} \tag{3a}$$

$$\left\langle s_i s_j \right\rangle_{\text{pair}} \equiv \sum_{\mathbf{s}} p_{\text{pair}}(\mathbf{s})s_i s_j = \left\langle s_i s_j \right\rangle_{\text{data}}. \tag{3b}$$

Following statistical physics terminology, the parameters h_i are usually called the *external fields* and J_{ij}, the *pairwise couplings* or *pairwise interactions*. One important property of the pairwise distributions in Eq. 2 is that it has the maximum amount of entropy among all the distribution that have the same mean and pairwise correlations as the data, and it is thus usually called the maximum entropy pairwise model. It is also called the Ising model, in accordance with the Ising model introduced in statistical physics as a simple model of magnetic materials.

Although typically written in terms of ±1 spin variables, it is sometimes useful to write the pairwise distribution of Eq. 2 in terms of Boolean variables $r_i = (s_i + 1)/2$. In the rest of the paper, we sometimes use such Boolean representation as some of the calculations and equations become considerably simpler in this representation. We denote the external fields and the couplings in the Boolean representations by $\mathcal{H}_i$ and $\mathcal{J}_{ij}$.

The external fields and pairwise couplings can be found using both exact and approximate methods as discussed in details in Roudi et al. (2009b) and reviewed briefly in the following sections. Before reviewing these methods, we first review the experimental studies that have used the maximum entropy pairwise model to study the statistics of spike trains.

REVIEW OF EXPERIMENTAL RESULTS

The maximum entropy pairwise model as a model for describing the statistics of neural firing patterns was introduced in Schneidman et al. (2006) and Shlens et al. (2006). This work used the Ising model to study the response of retinal ganglion cells to natural movies (Schneidman et al., 2006), steady spatially uniform illumination and white noise (Shlens et al., 2006), as well as the spontaneous activity of cultured cortical networks (Schneidman et al., 2006).

The main goal of these studies was to find out how close the fitted pairwise model is to the true experimentally computed distribution over **s**.

As a measure of distance, these studies compared the entropy difference between the true distribution and the pairwise model and compared it with the entropy difference between the true distribution and an independent model. The independent model is a distribution that has the same mean as the data but assumes the firing of each neuron is independent from the rest, i.e.

$$p_{ind}(\mathbf{s}) = \frac{1}{Z_{ind}} \prod_i \left[\langle s_i \rangle_{data} \delta_{s_i,1} + \left(1 - \langle s_i \rangle_{data}\right) \delta_{s_i,-1} \right]$$

$$= \frac{\exp\left[\sum_i h_i^{ind} s_i\right]}{\prod_i 2\cosh\left(h_i^{ind}\right)}. \tag{4a}$$

$$h_i^{ind} = \tanh^{-1}\left(\langle s_i \rangle_{data}\right) \tag{4b}$$

The measure of misfit can thus be defined as

$$\Delta = \overline{\frac{S_{pair} - S_{true}}{S_{ind} - S_{true}}}, \tag{5}$$

where the overline indicates averaging with respect to many samples of N neurons. The results of Schneidman et al. (2006) showed that for populations of size $N = 10$ or so, Δ was around 0.1. In other words, in terms of entropy difference, the pairwise model offered a ten fold improvement over the independent model. In the other study, Shlens et al. (2006) found $\Delta \sim 0.01$ for $N = 7$. These authors also considered a slightly different model in which only the pairwise correlation between the adjacent cells was used in the fit, and correspondingly the pairwise interactions between non-adjacent cells were set to 0. The results showed that this adjacent pairwise model also performed very well, with $\Delta \sim 0.02$ on average. It is important to note that in Schneidman et al. (2006) the stimulus induced long range correlations between cells, while in the data studied Shlens et al. the correlations extended only to nearby cells. Following these studies, Tang et al. (2008) reproduced these observations to a large extent in other cortical preparations and also concluded that the pairwise model can successfully approximate the multi-neuron firing patterns. In a very recent study (Shlens et al., 2009), Shlens et al. extended their previous analysis to population of up to 100 neurons concluding that the adjacent pairwise model performs very well even in this case, with $\Delta \sim 0.01$–0.02. The studies of Shlens et al. were done without stimulation or with white noise stimulus, situations in which neurons do not exhibit strong long range correlations. It is still unclear, from the existing experimental data, how the pairwise models of large size. (not necessarily adjacent) will perform for cases in which neuronal correlations exist between neurons separated by large distance, i.e. when stimulated by natural scenes.

The assumption that Δ is a good measure of distance rests upon the assumption that we are interested in finding out how different the true and model distributions are over the whole space of possible spike patterns. This can be appreciated when we note that the definition in Eq. 5 is equivalent to $\Delta = D_{KL}(p_{true} \| p_{pair})/D_{KL}(p_{true} \| p_{ind})$, where $D_{KL}(p \| q)$ is the Kullback–Leibler divergence

between two distributions $p(\mathbf{s})$ and $q(\mathbf{s})$ defined as $D_{KL}(p \| q) = \Sigma_\mathbf{s} p(\mathbf{s})\log[p(\mathbf{s})/q(\mathbf{s})]$. Using this observation, we can think of Δ as a weighted sum of the difference between the log probability of states according to the true distribution and according to the model distribution normalized by the distance to the independent model. Of course, we may not be interested in finding how different the two distributions are over the space of all possible states, but only how differently the model and true distributions assign probabilities to a subset of *important* states. For instance, in a particular setting, it may be important only to build a good model for the probability of all states in which some large number of neurons M fire simultaneously (i.e. when $\Sigma_i r_i = M$), regardless of how different the two distribution are on the rest of the states. It was found in Tkacik et al. (2006) that the pairwise model offers a significant improvement over the independent model in modelling the experimental probability of synchronous multi-neuron firing of up to $N = 15$ neurons.

APPROXIMATIONS FOR FITTING BINARY PAIRWISE MODELS

The commonly used Boltzmann Learning algorithm for fitting the parameters of the Ising model is a very slow process, particularly for large N. Although effort has been made to speed up the Boltzmann learning algorithm (Broderick et al., 2007), such modified Boltzmann algorithms still require many gradient descent steps and long Monte Carlo runs for each step. This fact motivates the development of fast techniques for calculating the model parameters which do not rely on gradient descent or Monte Carlo sampling. A number of such approximations have been studied in Roudi et al. (2009b). In what follows we describe these approximations, and in particular the relation between the simpler approximations [naive mean-field (nMF) approximation and the IP approximation] to the more advanced ones (TAP approximation and SM approximation).

NAIVE MEAN-FIELD APPROXIMATION AND THE INDEPENDENT-PAIR APPROXIMATIONS

The simplest method (Kappen and Rodriguez, 1998; Tanaka, 1998) for finding the parameters of the Ising models from the data uses mean-field theory:

$$\tanh^{-1} m_i = h_i + \sum_j J_{ij} m_j, \tag{6}$$

where $m_i = \langle s_i \rangle_{data}$. These equations express the effective field that determines the magnetization m_i as the external field plus the sum of the influences of other spins through their average values m_j, weighted by the couplings J_{ij}. Differentiating with respect to a magnetization m_j gives the inverse susceptibility (i.e. inverse correlation) matrix

$$(\mathbf{C}^{-1})_{ij} = -J_{ij} \tag{7}$$

for $i \neq j$, where $C_{ij} = \langle s_i s_j \rangle_{data} - m_i m_j$. Thus, if one knows the mean values m_i and correlations C_{ij} of the data, one can use Eq. 7 to find the J_{ij} and then solve Eq. 6 to find the h_i. We call this approximation "naive" mean-field theory (abbreviated nMFT) to distinguish it from the TAP approximation described below, which is also a mean-field theory.

In the IP approximation, one solves the two-spin problem for neurons i and j, ignoring the rest of the network. This yields the following expressions for the parameters in terms of the mean values and correlations:

$$J_{ij}^{\text{IP}} = \frac{1}{4}\log\left[\frac{\left((1+m_i)(1+m_j)+C_{ij}\right)\left((1-m_i)(1-m_j)+C_{ij}\right)}{\left((1-m_i)(1+m_j)+C_{ij}\right)\left((1+m_i)(1-m_j)+C_{ij}\right)}\right] \quad (8a)$$

$$h_i^j = \frac{1}{2}\log\left[\frac{(1+m_i)(1-m_j)-C_{ij}}{(1-m_i)(1-m_j)+C_{ij}}\right] + J_{ij} \quad (8b)$$

where h_i^j is the external field acting on i when it is considered in a pair with j. It has been noted in Roudi et al. (2009b) that in the limit $m_i \rightarrow -1$ and $m_j \rightarrow -1$, J_{ij}^{IP} matches the leading order of the low-rate expansion derived in Roudi et al. (2009a).

Although the couplings found in the IP approximation can be directly used as an approximation to the true values of J_{ij}, relating the fields h_i^j found from the IP approximation to those of the model is slightly tricky. The reason is that the expression we find depends on which j we took to pair with neuron i. It is natural to think that we can sum h_i^j over j, i.e. over all possible pairings of cell i, to find the Ising model parameter h_i. In doing so, however, we should be careful. The expression in Eq. 8b has two types of terms, those that only depend on i, i.e. the first term in the following decomposition, and those that involve j, i.e. the second and third terms below

$$h_i^j = \frac{1}{2}\log\left[\frac{1+m_i}{1-m_i}\right] + \frac{1}{2}\log\left[\frac{\frac{1-m_j-C_{ij}}{(1+m_i)}}{\frac{1-m_j+C_{ij}}{(1-m_i)}}\right] + J_{ij}. \quad (9)$$

The first terms is the field that would have been acting on i if it were not connected to any other neuron, and the rest are contributions from interactions with j. By simply summing h_i^j over j, we will be overcounting this term, once for each pairing. In other words, the correct IP approximation for h_i will be

$$h_i^{\text{IP}} = \frac{1}{2}\log\left[\frac{1+m_i}{1-m_i}\right] + \frac{1}{2}\sum_{j\neq i}\log\left[\frac{\frac{1-m_j-C_{ij}}{(1+m_i)}}{\frac{1-m_j+C_{ij}}{(1-m_i)}}\right] + \sum_{j\neq i} J_{ij}. \quad (10)$$

Although simple in its derivation and intuition, in the limit $m_i \rightarrow -1$ for all i, Eq. 10 recovers both the leading term and the first order corrections of the low-rate expansion, as shown in Section "The Independent-pair Approximation for the External Fields in the Limit $\delta \rightarrow 0$" in Appendix.

The simple nMF and IP approximations have been shown to perform well in deriving the parameters of the Ising model when the population size is small.

In Roudi et al. (2009b), we showed that for data binned at 10 ms, the nMF and the IP approximations perform well in deriving the parameters of the Ising model when the population size is small. In **Figure 1** we extend this study and evaluate how the quality of these approximations depend on the size of the time bin, δt. In this figure, we plot the R^2 value between the couplings found from nMFT and IP approximations and the results of long Boltzmann
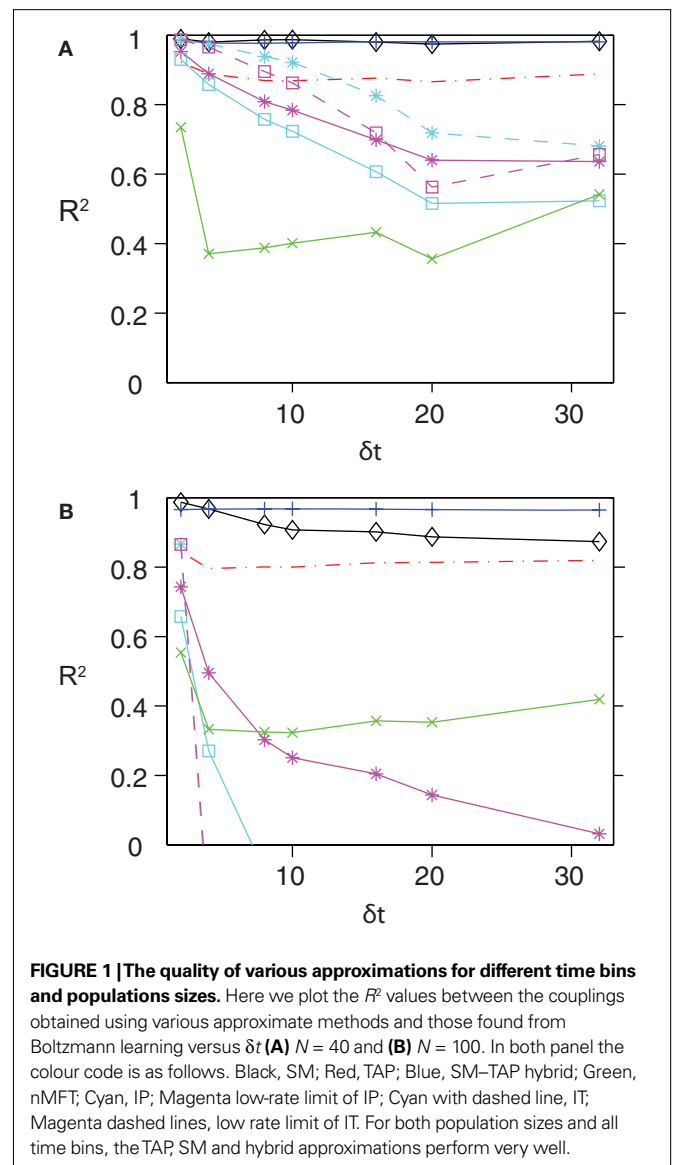
FIGURE 1 | The quality of various approximations for different time bins and populations sizes. Here we plot the R^2 values between the couplings obtained using various approximate methods and those found from Boltzmann learning versus δt **(A)** $N = 40$ and **(B)** $N = 100$. In both panel the colour code is as follows. Black, SM; Red, TAP; Blue, SM–TAP hybrid; Green, nMFT; Cyan, IP; Magenta low-rate limit of IP; Cyan with dashed line, IT; Magenta dashed lines, low rate limit of IT. For both population sizes and all time bins, the TAP, SM and hybrid approximations perform very well.

runs as a function of the time bin chosen to bin the data. Denoting the couplings found using the approximations and the Boltzmann learning by J_{ij}^{approx} and $J_{ij}^{\text{Boltzmann}}$, respectively, the R^2 is defined as

$$R^2 = 1 - \frac{\sum_{ij}\left(J_{ij}^{\text{approx}} - J_{ij}^{\text{Boltzmann}}\right)^2}{\sum_{ij}\left(J_{ij}^{\text{Boltzmann}} - \overline{J_{ij}^{\text{Boltzmann}}}\right)^2},$$

where $\overline{J_{ij}^{\text{Boltzmann}}} = [N(N-1)]^{-1}\sum_{i<j} J_{ij}^{\text{Boltzmann}}$. The quality of the approximations in **Figure 1** is shown for both $N = 40$ and 100. The simulations used to generate the spike trains and the Boltzmann learning procedure used in this figure are the same as those reported in Roudi et al. (2009b). Briefly, the simulated a cortical network was composed of two recurrent, interacting populations of neurons, one excitatory and the other inhibitory. Each neuron had 10% probability of receiving input from every other neuron in

the populations. The single neurons had Hodgkin–Huxley spike generation mechanisms, and we used conductance based synapses. Although we used the same network, the data used here had two main differences with those used in Roudi et al. (2009b). The first one is that here we use more gradient descent steps and longer Monte Carlo runs, namely, 60000 gradient descent steps and 40000 Monte Carlo steps per gradient descent step. The second one is that here we use 10000 s worth of data for estimating the mean values and correlations. This is 2.5 times larger than what we used before. Both of these improvements were made to ensure reliable estimates of the parameters, as well as the mean values and correlations, particularly for fine time bins. As can be seen in **Figure 1**, increasing either N or δt results in a decay in the quality of the IP approximation, as well as its low-rate limit. For the case of nMFT, a reasonable performance is observed only for $\delta t = 2$ ms and $N = 40$. For large population sizes and/or time bins nMFT is a bad approximation. Given the strong dependence of the quality of these simple approximations on population size and the size of the time bin, we describe below how one can extend these approximations to obtain more accurate expressions for finding the external fields and couplings of the pairwise model for large N and δt.

EXTENDING THE INDEPENDENT-PAIR APPROXIMATION

Extending the IP approximation is in principle straightforward. Instead of solving the problem of two isolated spins, we can solve the problem of three spins, i, j and k as shown in Section "Independent-triplet Approximation" in Appendix. This will lead to the IT approximation. In **Figure 1**, we show the quality of the IT approximation for finding the couplings as compared to the Boltzmann solutions. For $N = 40$, we see that the IT approximation provides an improvement over IP for different values of δt. In **Figure 1**, we also looked at the quality of the low rate limit of the IT approximation. In Section "Independent-triplet Approximation" in Appendix we show that, in the same way that the low rate limit of the IP approximation gives us the leading order terms of the low-rate expansion in Roudi et al. (2009a), the low rate limit of the IT approximation gives us the first order corrections to it. For $N = 40$, this is evident in the fact that the low-rate limit of IT outperforms the low-rate limit of IP for $\delta t \leq 15$ ms. When the population size is large, however, IT and its low rate limit outperform IP for only very fine time bins. Even for $\delta t = 4$ ms, IP and IT and their low rate limits perform very bad.

One can of course build on the idea of the IP and IT approximations and consider $n = 4, 5, \ldots$ spins. However, for any large value of n, this will be impractical and computationally expensive, for the following reason. As described in Section "Independent-triplet Approximation" in Appendix for the case of the IT approximation, there are two steps in building an independent-n-spin approximation. The first one is to express the probability of each of the 2^n possible states of a set of n spins in terms of the mean values and correlations of these spins. This requires inverting the $2^n \times 2^n$ matrix. The second step, which is only present for $n \geq 3$, is to express all correlation functions in terms of the mean values and pairwise correlations. Both of these steps become exponentially hard as n grows. Nonetheless, as shown in Section "Independent-triplet Approximation" in Appendix, even going to the triplet level can be a very useful exercise, as it offers a new way of computing the

difference between the entropy of the true model and that of the independent model (Eq. 15a in Section "Entropy Difference") as well as the difference between the entropy of the true model and that of the pairwise model (Eq. 15b in Section "Entropy Difference"). This derivation is considerably simpler than the original derivation of these equations based on the Sarmanov–Lancaster representation of the probability distribution described in Roudi et al. (2009a) as well as the derivations in Section "The Partition Function, Entropy and Moments of a Gibbs Distribution in the Limit $N\delta \to 0$" in Appendix, in which one starts by expanding the partition function of a Gibbs distribution. Furthermore, the IT approximation also yields a relation between the couplings and the mean values and pairwise correlations that coincides with leading term and the corrections found by low-rate expansion, as shown in Section "Independent-triplet Approximation" in Appendix.

EXTENDING THE NAIVE MEAN-FIELD: TAP EQUATIONS

The nMF, IP and IT approximations are good for fitting the model parameters when the typical number of spikes generated by the whole population in a time bin is small compared to 1, i.e. when $N\delta \ll 1$ (Roudi et al., 2009a,b), in which

$$\delta \equiv \frac{1}{N}\sum_i \nu_i \delta t = \frac{1}{N}\sum_i \frac{1+m_i}{2} \tag{11}$$

For large and/or high firing rate populations, however, such approximations perform poorly for inferring the model parameters. It is possible to make simple corrections to the nMF approximation such that the resulting approximation performs well even for large populations. This is the so called TAP approximation (Thouless et al., 1977). The idea dates back to Onsager, who added corrections to the nMF approximation, taking into account the effect of the magnetization of a spin i on itself via its influence on another spin j. Subsequently, it was shown that the resulting expression was exact for infinite-range spin-glass models (Thouless et al., 1977). The TAP equations are
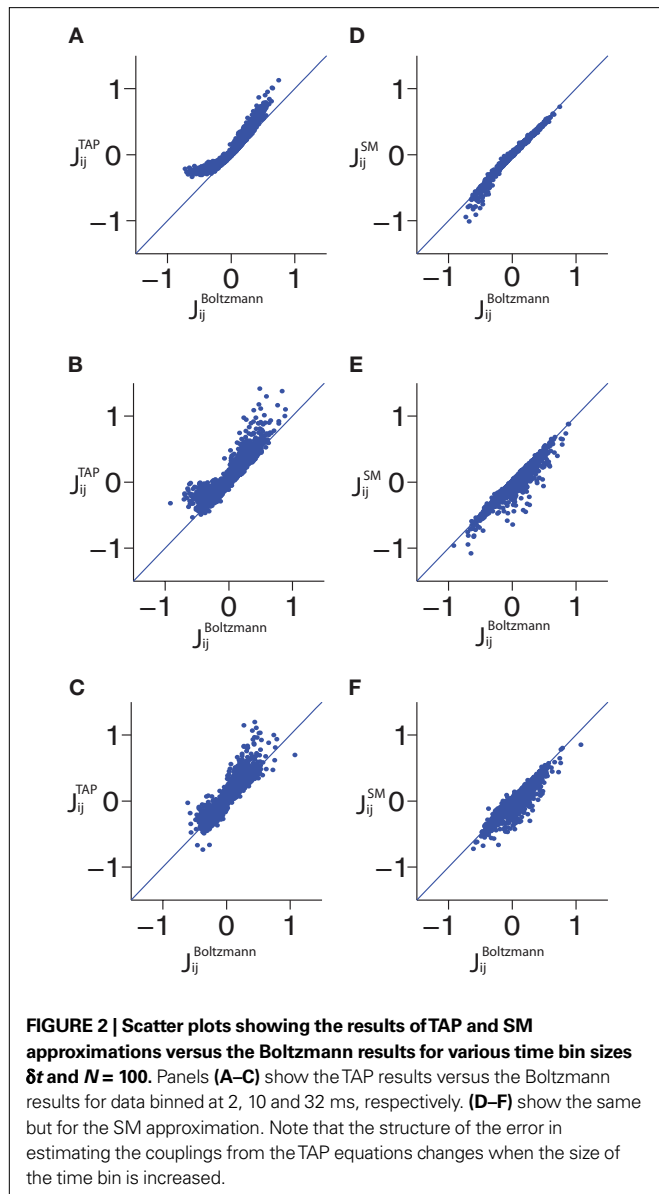
$$\tanh^{-1} m_i = h_i + \sum_{j \neq i} J_{ij} m_j - \sum_{j \neq i} J_{ij}^2 m_i \left(1 - m_j^2\right). \tag{12}$$

Differentiation with respect to m_j then gives ($i \neq j$)

$$\left(\mathbf{C}^{-1}\right)_{ij} = -J_{ij} - 2m_i m_j J_{ij}^2 \tag{13}$$

One can solve Eq. 13 for the J_{ij} and, after substituting the result in Eq. 12, solve Eq. 12 for the h_i (Kappen and Rodriguez, 1998; Tanaka, 1998).

There are several ways to derive this expression for the pairwise distribution Eq. 2. In Section "Derivation of TAP Equations from Belief Propagation" in Appendix, we show that these equations can also be derived from the celebrated Belief Propagation algorithm used in combinatorial optimization theory (Mezard and Montanari, 2009). When applied to spike trains from populations of up to 200 neurons, the inversion of TAP equations was shown to give remarkably accurate results (Roudi et al., 2009b) for fitting the pairwise model. In **Figure 2**, we show scatter plots comparing the couplings found by the TAP approximation versus the Boltzmann results for $N = 100$ and $\delta t = 2, 10, 32$ ms. The TAP approximations does well in all cases, and this is quantified in

FIGURE 2 | Scatter plots showing the results of TAP and SM approximations versus the Boltzmann results for various time bin sizes δt **and** $N = 100$**.** Panels **(A–C)** show the TAP results versus the Boltzmann results for data binned at 2, 10 and 32 ms, respectively. **(D–F)** show the same but for the SM approximation. Note that the structure of the error in estimating the couplings from the TAP equations changes when the size of the time bin is increased.

Figure 1. In **Figure 1**, we demonstrate the power of inverting the TAP equations for inferring the couplings for various time bins for both $N = 40$ and 100.

SESSAK–MONASSON APPROXIMATION

Most recently, Sessak and Monasson (2009) developed a perturbative expansion expressing the fields and couplings of the Ising distribution as a series expansion in the pairwise correlation functions C_{ij}. Some of the terms of the expression they found for the couplings could be summed up. It was noted in Roudi et al. (2009b), that one can think of the resulting expression as a combination of the nMF approximation and the IP approximation. The SM result can be written as

$$J_{ij}^{SM} = -\left(C^{-1}\right)_{ij} + J_{ij}^{IP} - \frac{C_{ij}}{\left(1 - m_i^2\right)\left(1 - m_j^2\right) - \left(C_{ij}\right)^2}, \tag{14}$$

The reason why the last terms should be subtracted is discussed below.

Let us consider two neurons connected to each other. In the IP approximation we calculate the fields and couplings for this pair exactly within the assumption that they do not affect the rest of the network and vice versa. If we were to find the coupling between this pair of neurons using the nMF approximation Eq. 7, the result would just be the last term in Eq. 14. The reason why we should subtract it is now clear: the first term in Eq. 14 includes an nMF solution to the pair problem. We subtract this part and replace it by the exact solution of the pair problem. **Figure 1** shows that this result is very robust to changing δt, although for $N = 100$, we can note a small decay in R^2 with δt. The good performance of the SM approximation can be also seen in the scatter plots shown for $N = 100$ in **Figure 2**. These observations support the SM approximation as a very powerful way of inferring the functional connections. Following the observation made in Roudi et al. (2009a), in **Figure 1**, we also show how a simple averaging of the best approximate methods, i.e. TAP inversion and SM can provide a very accurate approximation to the couplings across different time bin and population sizes.

ASSESSMENT OF MODEL QUALITY

The experimental result that the binary pairwise models provide very good models for the statistics of spike trains is very intriguing. However, the message they carry about the architecture and function of the nervous system is not clear. This is largely due to the fact that, as reviewed in Section "Review of Experimental Results," the experimental studies were conducted on populations of small number of neurons ($N \sim 10$) and their implications on the real sized system are not trivial. Is it the case that observing a very good pairwise model on a subsystem of a large system constrains the structure and function of the real sized network? Does it mean that there is something unique about the role of pairwise interactions in the real sized system? Answering this question depends to a large extent on answering *the extrapolation problem*: to what degree the experimentally reported success of pairwise models holds for the real sized system? In what follows, we discuss some theoretical results that bear on this question.

ENTROPY DIFFERENCE

The extrapolation problem was first addressed in Roudi et al. (2009a) by analysing the dependence of the misfit measure Δ (defined in Eq. 5) on N. The authors considered an arbitrary true distribution and computed the KL divergence between this distribution and an independent-neuron model as well as between it and the pairwise model. This was done using a perturbative expansion in $N\delta \ll 1$, where δ is defined in Eq. 11. The results were the following equations:

$$D_{KL}(p \| p_{ind}) = S_{ind} - S_{true} = g_{ind} N(N-1)\delta^2 + \mathcal{O}(N\delta^3) \tag{15a}$$

$$D_{KL}(p \| p_{pair}) = S_{pair} - S_{true}$$
$$= g_{pair} N(N-1)(N-2)\delta^3 + \mathcal{O}((N\delta^4) \tag{15b}$$

where g_{ind} and g_{pair} are constants that do not depend on N or δt and are defined in Eqs. 36a and 38a. Using these expressions for the KL divergences yields

$$\Delta = \frac{g_{\text{pair}}}{g_{\text{ind}}}(N-2)\delta. \tag{16}$$

Equations 15 and 16 show that for small $N\delta$, Δ will be very close to 0, independent of the structure of the true distribution. In other words, in this regime, a very good pairwise-model fit is a generic property and does not tell us anything new about the underlying structure of the true probability distribution. It is important to note that the perturbative expansion is always valid if δt is small enough. That is, simply by choosing a sufficiently small time bin, we can push Δ as close to 0 as we want.
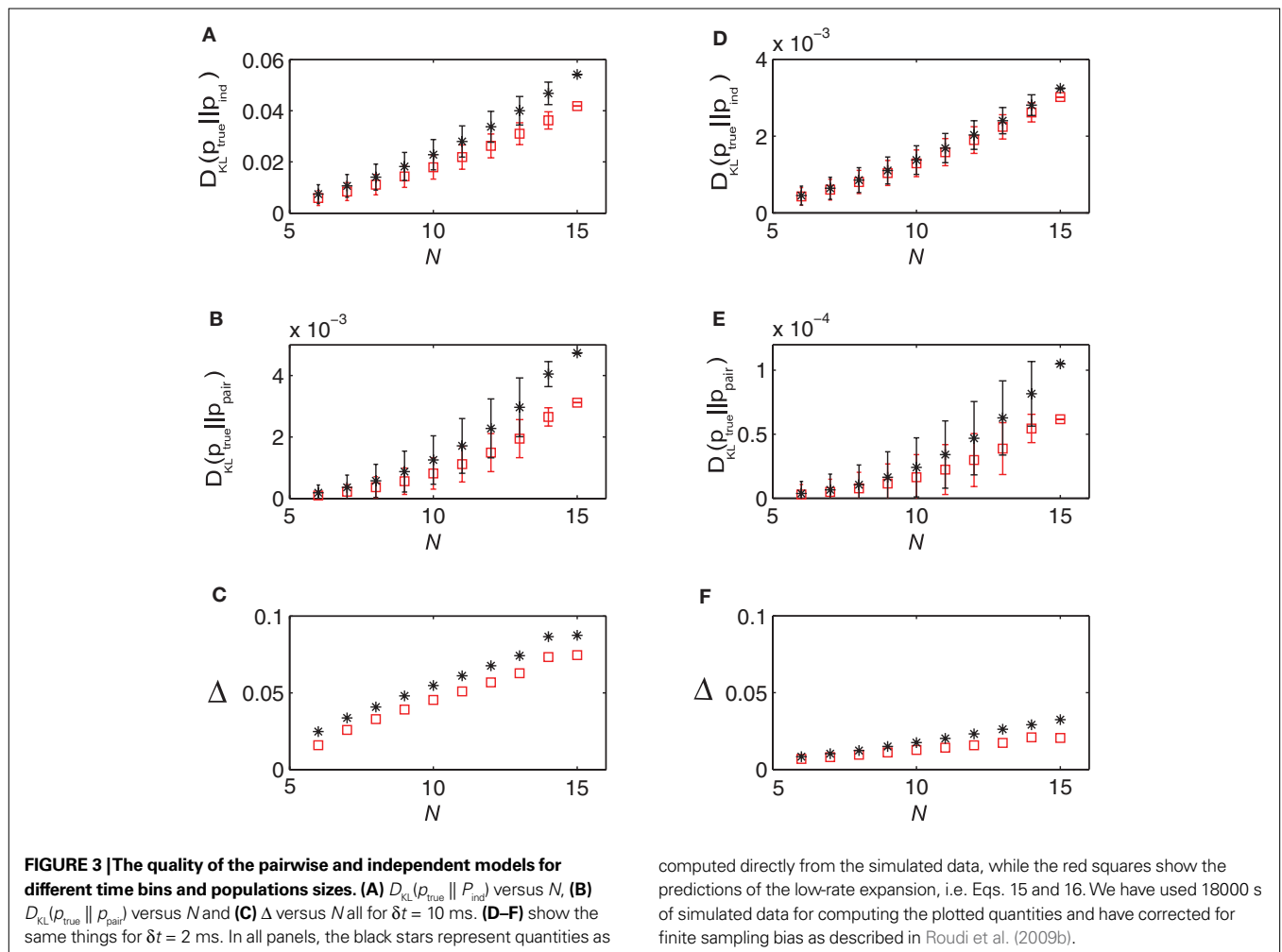
In Roudi et al. (2009a), g_{pair} and g_{ind} are related to the parameters of the pairwise model up to corrections of $\mathcal{O}(N\delta)$. In Section "The Partition Function, Entropy and Moments of a Gibbs Distribution in the Limit $N\delta \to 0$" in Appendix, we present a different derivation by expanding the partition function of a true Gibbs probability distribution around the partition function of a distribution without couplings. In the following subsection, we also use the results of this derivation to compare the probability of synchronous spikes under the model and the true distributions. Furthermore, in Section "Independent-triplet Approximation" in Appendix, we extend the idea beyond the IP approximation and approximate the entropy of a given distribution as a sum over the entropies of triplets of

isolated neurons. We show that this approach leads to Eq. 15 in a substantially simpler way than those reported in Roudi et al. (2009a) and Section "The Partition Function, Entropy and Moments of a Gibbs Distribution in the Limit $N\delta \to 0$" in Appendix.

In **Figure 3**, we show how $D_{\text{KL}}(p \parallel p_{\text{ind}})$, $D_{\text{KL}}(p \parallel p_{\text{pair}})$ and Δ vary with N and δt for data generated from a simulated network. We have also plotted the predictions of the low-rate expansion Eqs. 15 and 16. As shown in these figures, the low-rate expansion nicely predicts the behaviour of the measurements from the simulations particularly for small N and δt. For $\delta t = 10$ ms, we have $\delta_{10} = 0.076$ and for $\delta t = 2$ ms, we have $\delta_2 = 0.019$ (note that this gives $\delta_{10}/\delta_2 = 3.95$, a ratio that would have been equal to 5 if the bins were independent). Both the results from the low-rate expansion and those found directly from the simulations show that using finer time bins decrease Δ for fixed N.

PROBABILITY OF SIMULTANEOUS SPIKES

As discussed in Section "Review of Experimental Results," in addition to entropic measures such as KL divergence and Δ, which in a sense ask how well the model approximates the experimental probabilities of all possible spike patterns, we can restrict our quality measure to a subset of possible spike patterns. We can, for instance, ask how well the pairwise model approximates the probability of M

FIGURE 3 | The quality of the pairwise and independent models for different time bins and populations sizes. (A) $D_{\text{KL}}(p_{\text{true}} \parallel P_{\text{ind}})$ versus N, **(B)** $D_{\text{KL}}(p_{\text{true}} \parallel p_{\text{pair}})$ versus N and **(C)** Δ versus N all for $\delta t = 10$ ms. **(D–F)** show the same things for $\delta t = 2$ ms. In all panels, the black stars represent quantities as computed directly from the simulated data, while the red squares show the predictions of the low-rate expansion, i.e. Eqs. 15 and 16. We have used 18000 s of simulated data for computing the plotted quantities and have corrected for finite sampling bias as described in Roudi et al. (2009b).

simultaneous spikes. Similar to the case of Δ, before getting too impressed about the power of pairwise models in approximating the true distributions, we should find out what we expect in the case of an arbitrary, or random, true probability distribution.

Suppose now that we have a distribution over a set of variables of the form of Eqs. 27a and 27b. For this distribution, the probability that a set of M neurons, $I = \{i_1, i_2, \ldots, i_M\}$ out of the whole population of N fire in a time bin while the rest do not is

$$\log p_I = \sum_{i \in I} \mathcal{H}_i + \sum_{i<j \in I} \mathcal{J}_{ij} + \sum_{i<j<k \in I} \mathcal{K}_{ijk} + \cdots - \log Z. \qquad (17)$$

Averaging over all possible I we get

$$q(M,N) \equiv \binom{N}{M}^{-1} \sum_I \log p_I = M\overline{\mathcal{H}} + \binom{M}{2}\overline{\mathcal{J}} + \binom{M}{3}\overline{\mathcal{K}} + \cdots - \log Z. \qquad (18)$$

where $\overline{\mathcal{H}}, \overline{\mathcal{J}}$ and $\overline{\mathcal{K}}$ are the mean values of the fields and pairwise, third-order etc coupling. In Section "The Partition Function, Entropy and Moments of a Gibbs Distribution in the Limit $N\delta \rightarrow 0$" in Appendix, we show that, to leading order in $N\delta$, the external fields and pairwise couplings of fitted models (independent or pairwise) match those of the true model (see Eqs. 32a and 32b). Using this, we see that

$$q_{\text{true}}(M,N) - q_{\text{ind}}(M,N) \sim \left(\overline{\mathcal{J}}N\right)(M/N)^2$$
$$+ \left(\overline{\mathcal{K}}N^2\right)(M/N)^3 + \mathcal{O}\left((M/N)^4\right) \qquad (19a)$$

$$q_{\text{true}}(M,N) - q_{\text{pair}}(M,N) \sim \left(\overline{\mathcal{K}}N^2\right)(M/N)^3 + \mathcal{O}\left((M/N)^4\right). \qquad (19b)$$

To have well defined behaviour in large N limit, one should have $\overline{\mathcal{J}}N \sim 1$ and $\overline{\mathcal{K}}N^2 \sim 1$. Eq. 19 show that, for $M/N \ll 1$, both the independent and pairwise models are close to the true distribution. For $M/N \sim \mathcal{O}(1)$ (of course still $M/N < 1$), the difference between the model and true probabilities of observing M synchronous spike increases. For all ranges of M, the difference is larger for the independent model. These predictions are consistent with the experimental results found in the retina (Tkacik et al., 2006).

In the above calculation, we compute the difference between the true and model values of q, i.e. the *mean of the log probability of M synchronous spikes*. However, one can also calculate the *log mean probability of M synchronous spikes*, i.e.

$$w(M,N) \equiv -\log\binom{N}{M}$$
$$+ \log \sum_I \exp\left[\sum_{i \in I} \mathcal{H}_i + \sum_{i<j \in I} \mathcal{J}_{ij} + \sum_{i<j<k \in I} \mathcal{K}_{ijk} + \cdots\right] - \log Z. \qquad (20)$$

Calculating w is in theory very hard, because the second term above involves calculating the averages of exponential functions of variables. However, if the population is homogenous enough, such that the average couplings and fields from one sample of M neurons to the next does not change much, we can approximate the average of the exponential of a variable with the exponential of the average it. Doing this, will again lead to Eq. 19. The difference

between the two measure w and q will most likely appear for small M where the averages of the fields and couplings depends on the sample of M neurons more strongly than when M is large.

EXTENSIONS OF THE BINARY PAIRWISE MODEL

In the previous sections, we described various approximate methods for fitting a pairwise model of the type of Eq. 2. We also studied how good a model it will be for spike trains, using analytical calculations and computer simulations. As we describe below, there are two issues with a model of the type of Eq. 2 that lead to new directions for extending the pairwise models studied here.

The first issue is the use of binary variables as a representation of the states of the system. For fine time bins and neural spike trains, the binary representation serves its purpose very well. However, in many other systems, a binary representation will be a naive simplification. Examples of such systems are modular models of the cortex in which the state of each cortical module is described by a variable taking a number of states usually much larger than 2. In Section "Extension to Non-binary Variables" we briefly describe a simple non-binary model useful for modelling the statistics of such systems.

The second issue is that by using Eq. 2 in cortical networks, one is essentially approximating the statistics of a highly non-equilibrium system with asymmetric physical interactions, e.g. a balanced cortical network, by an equilibrium distribution with symmetric interactions. This manifests itself in a lack of a simple relationship between the functional connectives to real physical connections. In our simulations we observed that there was no obvious relation between the synaptic connectivity and the inferred functional connections. Second, as we showed here and in our previous work, for large populations, the model quality decays. Although one can avoid this decay by decreasing δt as N grows, eventually one will get into the regime of very fine δt, where the assumption of independent bins used to build the model does not hold any more and one should start including the state transitions in the spike patterns (Roudi et al., 2009a). In fact, Tang et al. (2008) showed that even in the cases that the pairwise distribution of Eq. 2 is a good model for predicting the distribution of spike patterns, it will not be a good one for predicting the transition probabilities between them. These observations encourage one to go beyond an equilibrium distribution with symmetric weights. In the second extension, described in Section "Extension to Dynamics and Asymmetric Interactions," we propose one such model, although a detailed study of the properties of such model is beyond the scope of this paper.

EXTENSION TO NON-BINARY VARIABLES

The binary representation is probably a good one for spike trains binned into fine time bins. However, for larger time bins where there is a considerable probability of observing more than one spike in a bin, as well as for a number of other systems, the binary representation may only serve as a naive simplification and going to non-binary representations is warranted. Example of such systems include the protein chains and modular cortical models. In probabilistic models of protein chains, each site is represented by a non-binary variable that takes one of its possible q states depending on the amino acid that sits on that site. In a number of models for the operations of cortical networks, one considers a network of interconnected modules, the state of each of which is represented by

a non-binary variable (Kropff and Treves, 2005; Russo et al., 2008). Each state of one such variable corresponds to, e.g. one of the many memory states stored in the corresponding module.

For a set of non-binary variables $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$, $\sigma_i = 1, \ldots, q$, one can simply write down a maximum entropy pairwise Gibbs distribution as

$$p(\sigma) = \frac{1}{Z} \exp\left[\sum_i \sum_\alpha \mathcal{H}_i^\alpha u_{\alpha\sigma_i} + \sum_{i<j} \sum_{\alpha,\beta} \mathcal{J}_{ij}^{\alpha\beta} u_{\alpha\sigma_i} u_{\beta\sigma_j} \right], \quad (21a)$$

$$u_{\alpha\sigma_i} = \delta_{\alpha\sigma_i} - \delta_{1\sigma_i}, \quad (21b)$$

where α and β go from $1, \ldots, q$ and index the q possible states of each variable. For $q = 2$, the above distribution reduces to the binary case with Boolean variables, and when one forces $\mathcal{J}_{ij}^{\alpha\beta} = 0$ for $\alpha \neq \beta$ one recovers the q-state Potts model. Similar to the binary case, here also one is given the experimentally observed values of $\langle u_{\alpha\sigma_i} \rangle_{\text{data}}$ and $\langle u_{\alpha\sigma_i} u_{\beta\sigma_j} \rangle_{\text{data}}$ and wants to infer the fields and couplings that are consistent with them.

The approximate methods described in this paper can be adopted, with some effort, to the case of non-binary variable as well. In particular, it is easy to derive the difference between the entropy of the true distribution, the pairwise model and the independent model in the low-rate limit of the non-binary model. Assuming that $\langle u_{\alpha\sigma_i} \rangle_{\text{data}} = \mathcal{O}(\varepsilon)$ for $\alpha \neq 1$, the low-rate regime in the case of non-binary variables is characterised by $N(q-1)\varepsilon \ll 1$. In this regime, similar to the binary case, $S_{\text{pair}} - S_{\text{true}} \propto [N(q-1)\varepsilon]^3$ and $S_{\text{ind}} - S_{\text{true}} \propto [N(q-1)\varepsilon]^2$ and $\Delta = N(q-1)\varepsilon$, and consequently the pairwise model performs very well in the low-rate regime.

As we described before, the low-rate regime is where most experimental studies on binary pairwise models were performed, and the result of the low-rate expansion of the entropies explains the reported success of binary pairwise models in those studies. On the other hand, the low-rate regime of the non-binary variable may be of little use. This is because the systems to which the non-binary representation should be applied are unlikely to fall into the low-rate regime. For instance, in the case of modular memory networks the low-rate regime would be the case in which the network spends a significantly larger time in its ground state (no memory retrieved) compared to the time it spends operating and retrieving memory. A more likely scenario is the one where all the states (memory or no-memory) have approximately similar probabilities of occurrence over a period of time. How useful pairwise models are in describing the statistics of such non-binary systems away from the trivial regime of the low-rate expansion is not known. Studying the quality of non-pairwise models in these cases and developing efficient ways to fit such models, in particular based on extensions of the powerful approximations such TAP and SM to non-binary variables will be the focus of future work. In particular, it is important to note that writing the SM approximation for the non-binary case will be a straightforward task in light of the relation we described between the SM, nMFT and IP approximations in Section "Sessak–Monasson Approximation."

EXTENSION TO DYNAMICS AND ASYMMETRIC INTERACTIONS

The Glauber model (Glauber, 1963) is the simplest dynamical model that has a stationary distribution equal to the Ising model distribution (Eq. 2). It is defined by a simple stochastic dynamics

in which at each timestep $\delta t = \tau_0/N$ one spin is chosen randomly and updated, taking the value $+1$ (i.e. the neuron spikes) with probability

$$p_+ = \frac{1}{1 + \exp\left[-2\left(h_i + \sum_j J_{ij} s_j \right) \right]}. \quad (22)$$

Although the interactions J_{ij} in the static Ising model are symmetric [any antisymmetric piece would cancel in computing the distribution (Eq. 2)], a Glauber model with asymmetric J_{ij} is perfectly possible (Crisanti and Sompolinsky, 1988; Ginzburg and Sompolinsky, 1994).

This kinetic Ising model is closely related to another class of recently-studied network models called generalized linear models (GLMs, Okatan et al., 2005; Truccolo et al., 2005; Pillow et al., 2008). In a GLM, neurons receive a net input from other neurons of the linear form,

$$\sum_j \int_0^\infty d\tau J_{ij}(\tau) s_j(t - \tau) \quad (23)$$

and spike with a probability per unit time equal to a function f of this input. Maximum-likelihood techniques have been developed for solving the inverse problem for GLMs (finding the linear kernels that give the stimulus input and that from the other neurons in the network, given spike train data) (Okatan et al., 2005; Truccolo et al., 2005). They have been applied successfully to analysing spatiotemporal correlations in populations of retinal ganglion cells (Pillow et al., 2008).

The Glauber model looks superficially like a GLM with instantaneous interactions and $f(x)$ equal to a logistic sigmoid function $1/[1 + \exp(-2x)]$. But there is a difference in the dynamics: spins are not spikes. In the Glauber model, a spin/neuron retains its value $(+1 \text{ or } -1)$ until it is chosen again for updating. Since the updating is random, this persistence time is exponentially distributed with mean τ_0. Thus a Glauber-model "spike" has a variable width, and the autocorrelation function of a free spin exhibits exponential decay with a time constant of τ_0. In the GLM, in contrast, a spike is really a spike and the autocorrelation function (for constant input) is a delta-function at $t = 0$.

The time constant characterizing the kernel $J_{ij}(\tau)$ in a GLM has a similar effect to τ_0 in the Glauber model, but a GLM with an exponential kernel is not exactly equivalent to the Glauber model. In the GLM, the effect of a presynaptic spike is spread out and delayed in time by the kernel, but once it is felt, the postsynaptic firing rate changes immediately. In the Glauber model, the presynaptic "spike" is felt instantaneously and without delay, but the firing state of the neuron takes on the order of τ_0 to change in response.

GLMs grew out of a class of single-neuron models called LNP models. The name LNP comes from the fact that there is a linear (L) filtering of the inputs, the result of which is fed to a nonlinear (N) function that specifies an instantaneous Poisson (P) firing rate. In the earlier studies, the focus was on the sensory periphery, where the input was an externally specified "stimulus". An aim of this modelling effort was to improve on classical linear receptive field models. Thus, in the usual formulation of a GLM network, one writes the

total input as a sum of two terms, one linear in the stimulus as in the LNP model and the other linear in the spike trains of the other neurons. Of course, one can trivially add a "stimulus" term in a Glauber model, so this differences is not an essential one.

Similarly, interactions with temporal kernels $J_{ij}(\tau)$ can be included straightforwardly in a Glauber model. Such a model is equivalent to a GLM in the limit $\tau_0 \to 0$. (One has to multiply the kernels by $1/\tau$ while taking the limit, because the integrated strength of a "Glauber spike" is proportional to τ_0, while that of an ordinary spike in a GLM is 1.)

One can derive a learning algorithm for a Glauber model, given its history, in a standard way, by maximizing the likelihood of the history. The update rule, which is exact in the same way that Boltzmann learning is for the symmetric model, is

$$\delta J_{ij} = \eta \left\langle \left[s_i(t_i + \varepsilon) - \tanh\left(b_i + \sum_k J_{ik}\delta s_k(t_i)\right) \right] \delta s_j(t_i) \right\rangle, \quad (24)$$

where $\delta s_j(t) = s_j(t) - \langle s_j \rangle$, the average is over the times t_i at which unit i is updated, and $b_i = \tanh^{-1}\langle s_i \rangle = h_i + \Sigma_j J_{ij}\langle s_j \rangle$.

One can also get a simple and potentially useful approximate algorithm which requires no iteration by expanding the tanh in Eq. 24 to first order in the J_{ij} (i.e. to first order around the independent-neuron model). Then at convergence ($\delta J_{ij} = 0$) we have

$$\left\langle \delta s_i(t_i + \varepsilon)\delta s_j(t_i) \right\rangle = \left(1 - \langle s_i \rangle^2\right)\sum_k J_{ik}\left\langle \delta s_k(t_i)\delta s_j(t_i) \right\rangle, \quad (25)$$

which is a simple linear matrix equation that can be solved for the J_{ij}.

DISCUSSION

The brain synthesizes higher-level concepts from multiple inputs received, and in many fields of research scientists are interested in inferring as simple a description as possible, given potentially vast amounts of data. Such processes of learning take a set of average values, correlations, or any other pattern in the data, and from these arrive at another representation, which is useful for speed or accuracy of prediction, for compressed storage of the data, as a grounds for decision making, or in any other aspect which improves the functionality or competitiveness of the brain or the researcher. Models and algorithms for learning in these contexts have been studied, in great detail, in neuroscience, image processing, and many other fields, for quite some time; consider, e.g. the introduction of Boltzmann machines more than a quarter of a century ago (Hinton and Sejnowski, 1983), and the now more than 10-year-old monograph on learning in graphical models (Jordan, 1998).

As one could have expected, there is a trade-off between the complexity of the model to be learned and the efficiency of the learning algorithms. Boltzmann machines are, in principle, able to learn very complex models, and provably so, but convergence is then quite slow; modern applications of those methods centre on various restricted Boltzmann machine models, which can be learned faster (Hinton, 2007).

When we consider very large systems and/or problems where convergence time of the learning is a serious concern, then we are primarily interested fast learning processes. Often such fast processes come as approximation grounded on assumption on

the model parameters, and the central issue is then obviously the accuracy of the learning outcomes based on those assumptions. In this paper we have revisited some classical and some more recent approximations of this kind that take their inspiration from statistical physics. The two simplest algorithms considered are *nMF* and *IP approximation*. As we have shown for neural data, both of these perform poorly except for small systems (small N), or for systems which have little variability (small value of $N\delta$), where δ here can be thought of as a proxy for the deviation from a uniform state. On the other hand, one generalization of each method, the TAP approximation for the nMF and the recent SM approximation for the IP approximation, perform much better. In the case of neural data analysed here, the parameter δ depends on the size of the time bin, δt. We showed that the TAP and SM approximations perform well for a large range of δt. In addition to studying the effect of δt on the performance of these approximations, we also studied the quality of pairwise models using a variety of methods. Using a few alternative analytical calculations, we derived mathematical expressions relating the quality of pairwise models to the size of the population and the lower order statistics of the spike trains. This analysis shows that for finer time bins the pairwise distribution becomes a better approximation to the true distribution of spike trains (Roudi et al., 2009a). We confirmed this analytical result on synthetic data.

Our focus in this paper was on approximate inference methods that offer closed form equations relating model parameters to the measured statistics of the model. Another class include methods that rely on iterative algorithms. This class of approximate methods include the recent improvement of the Boltzmann learning in Broderick et al. (2007) and the "susceptibility propagation" (Mora, 2007; Mezard and Mora, 2008). These methods are generally slower than the closed form approximation, and a systematic comparison between their performance against what we have studied here is lacking.

To conclude, we have here framed the presentation in terms of inferring representations of neural data and assessing the goodness of the model. Although our focus was on neural data analysis, it is important to note that similar problems appear in many other fields of modern biology, for instance, e.g. in network reconstruction of gene regulatory networks, in genomic assembly in metagenomics projects, and in many other problems. Given ==the present explosion in sequencing technologies it is conceivable that the more novel applications will soon appear outside neuroscience.==

APPENDIX

THE PARTITION FUNCTION, ENTROPY AND MOMENTS OF A GIBBS DISTRIBUTION IN THE LIMIT $N\delta \to 0$

Suppose we have a true distribution of the following form

$$p(\mathbf{s}) = \frac{1}{Z}\exp\left[\sum_i h_i s_i + \sum_{i<j} J_{ij}s_i s_j + \sum_{i<j<k} K_{ijk}s_i s_j s_k + \cdots\right]. \quad (26)$$

In this Appendix, we first find the relation between the external fields and pairwise and third order couplings. As we show below, this allows us to compute the probability of synchronous spikes. We also compute the entropy of this distribution, in the small spike probability regime to derive the expression in Eqs. 15. This task

can be accomplished much more easily if we rewrite the distribution of Eq. 26 in terms of zero-one variables r_i, instead of the spin variables s_j, i.e.

$$p(\mathbf{r}) = \frac{1}{Z} \exp\left[\beta\left(\sum_i \mathcal{H}_i r_i + \sum_{i<j} \mathcal{J}_{ij} r_i r_j + \sum_{i<j<k} \mathcal{K}_{ijk} r_i r_j r_k + \cdots \right) \right] \quad (27a)$$

$$r_i = \frac{(1+s_i)}{2}, \quad (27b)$$

where the auxiliary inverse temperature β is introduced because it will allow us to compute the entropy, as will become clear later.

The crucial step in computing the relation between the moments, parameters and entropy of a distribution is computing its partition function. To compute the partition function of Eq. 27a, Z, we note that

$$\Gamma_p \equiv \frac{Z}{Z_0} - 1 = \left\langle \exp\left[\sum_{i<j} \beta \mathcal{J}_{ij} r_i r_j + \sum_{i<j<k} \beta \mathcal{K}_{ijk} r_i r_j r_k + \cdots \right] \right\rangle_0 - 1$$

$$= \sum_{n=1}^{\infty} \frac{1}{n!} \left\langle \left[\sum_{i<j} \beta \mathcal{J}_{ij} r_i r_j + \sum_{i<j<k} \beta \mathcal{K}_{ijk} r_i r_j r_k + \cdots \right]^n \right\rangle_0, \quad (28)$$

where $\langle\rangle_0$ indicates averaging with respect to the distribution

$$p_0(r) = \frac{1}{Z_0} \exp\left[\sum_i \mathcal{H}_i r_i \right]. \quad (29)$$

Note that p_0 is *not* the independent model for p in Eqs. 27a and 27b but only the part of this distribution that includes the fields. In fact, as we show in the following (Eq. 32a) the fields of the independent model to p only match $\mathcal{H}_i$ to $\mathcal{O}(N\delta)$ corrections, i.e. $\mathcal{H}_i = \mathcal{H}_i^{\mathrm{ind}} + \mathcal{O}(N\delta)$. When dealing with corrections to the fields and couplings, this note will be important.

Since $\langle r_i^n \rangle_0 = \langle r_i \rangle_0 \equiv \delta_i$, a term with l distinct indices in the expansion of the term inside the average in Eq. 28 is of the order of δ^l. Therefore, to $\mathcal{O}((N\delta)^3)$, we have

$$\Gamma = \sum_{i<j} \sum_n \frac{\beta^n}{n!} J_{ij}^n \delta_i \delta_j$$

$$+ \sum_{i<j<k} \sum_{n_1,n_2} \frac{\beta^{n_1+n_2}}{n_1! n_2!} \left[J_{ij}^{n_1} J_{ik}^{n_2} + J_{ij}^{n_1} J_{jk}^{n_2} + J_{ik}^{n_1} J_{jk}^{n_2} \right] \delta_i \delta_j \delta_k$$

$$+ \sum_{i<j<k} \sum_{n_1,n_2,n_3} \frac{\beta^{n_1+n_2+n_3}}{n_1! n_2! n_3!} J_{ij}^{n_1} J_{ik}^{n_2} J_{jk}^{n_3} \delta_i \delta_j \delta_k$$

$$+ \sum_{i<j<k} \sum_n \frac{\beta^n}{n!} K_{ijk}^n \delta_i \delta_j \delta_k$$

$$+ \sum_{i<j<k} \sum_{n_1,n_2} \frac{\beta^{n_1+n_2}}{n_1! n_2!} K_{ijk}^{n_1} \left[J_{ij}^{n_2} + J_{ik}^{n_2} + J_{jk}^{n_2} \right] \delta_i \delta_j \delta_k$$

$$+ \sum_{i<j<k} \sum_{n_1,n_2,n_3} \frac{\beta^{n_1+n_2+n_3}}{n_1! n_2! n_3!} K_{ijk}^{n_1} \left[J_{ij}^{n_2} J_{ik}^{n_3} + J_{ij}^{n_2} J_{jk}^{n_3} + J_{ik}^{n_2} J_{jk}^{n_3} \right] \delta_i \delta_j \delta_k$$

$$+ \sum_{i<j<k} \sum_{n_1,n_2,n_3,n_4} \frac{\beta^{n_1+n_2+n_3+n_4}}{n_1! n_2! n_3! n_4!} K_{ijk}^{n_1} J_{ij}^{n_2} J_{ik}^{n_3} J_{jk}^{n_4} \delta_i \delta_j \delta_k + \mathcal{O}((N\delta)^4), \quad (30)$$

where the sums over n, n_1, n_2, n_3 and n_4 run from 1 to infinity. Performing these sums yields

$$\Gamma = \sum_{i<j} \phi_{ij} \delta_i \delta_j + \sum_{i<j<k} [\phi_{ij}\phi_{ik} + \phi_{ik}\phi_{jk} + \phi_{ij}\phi_{jk} + \phi_{ij}\phi_{ik}\phi_{jk}] \delta_i \delta_j \delta_k$$

$$+ \sum_{i<j<k} \phi_{ijk}(1+\phi_{ij})(1+\phi_{ik})(1+\phi_{jk}) \delta_i \delta_j \delta_k + \mathcal{O}((N\delta)^4) \quad (31)$$

where $\phi_{ij} = \exp(\beta \mathcal{J}_{ij}) - 1$ and $\phi_{ijk} = \exp(\beta \mathcal{K}_{ijk}) - 1$. From Eq. 31, we can immediately compute the relation between the mean values, pairwise and three-point correlation functions and the parameters of the distribution. For $\beta = 1$, we have

$$\tilde{m}_i \equiv \langle r_i \rangle = \frac{\partial \log Z}{\partial \mathcal{H}_i} = \delta_i + \frac{\partial \log(1+\Gamma)}{\partial \mathcal{H}_i} = \delta_i \left[1 + \sum_{j\neq i} \phi_{ij}\delta_j + \mathcal{O}((N\delta)^2) \right] \quad (32a)$$

$$\tilde{C}_{ij} \equiv \langle r_i r_j \rangle = \frac{\partial \log Z}{\partial \mathcal{J}_{ij}} = \frac{\partial \log(1+\Gamma)}{\partial \mathcal{J}_{ij}}$$

$$= \exp(\mathcal{J}_{ij})\delta_i \delta_j \left[1 + \sum_{k\neq i,j} \left(\phi_{ik} + \phi_{jk} + \phi_{ik}\phi_{jk} \right)\delta_k + \mathcal{O}[(N\delta)^2] \right] \quad (32b)$$

$$\tilde{C}_{ijk} \equiv \langle r_i r_j r_k \rangle = \frac{\partial \log Z}{\partial \mathcal{K}_{ijk}} = \frac{\partial \log(1+\Gamma)}{\partial \mathcal{K}_{ijk}}$$

$$= \exp(\mathcal{K}_{ijk})\exp(\mathcal{J}_{ij} + \mathcal{J}_{ik} + \mathcal{J}_{jk})\delta_i \delta_j \delta_k + \mathcal{O}(N\delta^4). \quad (32c)$$

The relations between mean values and pairwise correlations and the external fields and pairwise couplings in Eqs. 32a and 32b to their leading orders were reported previously in Roudi et al. (2009a), using a slightly different approach. However, the corrections and three-neuron correlations were not computed there.

An interesting result of this calculation is a relation between the three-neuron correlations for the pairwise distribution, i.e. when $\mathcal{K}_{ijk} = 0$, and the lower moments

$$\tilde{C}_{ijk}^{\mathrm{pair}} \equiv \langle r_i r_j r_k \rangle_{\mathrm{pair}} = \frac{\tilde{C}_{ij}\tilde{C}_{ik}\tilde{C}_{jk}}{\tilde{m}_i \tilde{m}_j \tilde{m}_k} + \mathcal{O}(N\delta^4). \quad (33)$$

The fact that, to leading order in $N\delta$, the external fields and couplings are determined by mean values and pairwise correlations allows us to compute the leading-order probabilities of synchronous spikes reported as we did in Section "Probability of Simultaneous Spikes."

We can now use Eq. 31 to find the entropies of the distribution in Eq. 27a

$$S = \log(Z) - \frac{\partial \log Z}{\partial \beta} = \log(Z_0) + \log(1+\Gamma) - \sum_i \mathcal{H}_i \langle r_i \rangle$$

$$- \sum_{i<j} \mathcal{J}_{ij} \langle r_i r_j \rangle - \sum_{i<j<k} \mathcal{K}_{ijk} \langle r_i r_j r_k \rangle. \quad (34)$$

From Eqs. 32a, 32b and 32c for an independent fit to the distribution Eq. 27a

$$S_{\mathrm{ind}} - S = -\Gamma + \sum_{i<j} \mathcal{J}_{ij} \langle r_i r_j \rangle + \sum_{i<j<k} \mathcal{K}_{ijk} \langle r_i r_j r_k \rangle + \mathcal{O}((N\delta)^4)$$

$$= -\sum_{i<j} \phi_{ij} \tilde{m}_i \tilde{m}_j + \sum_{i<j} \mathcal{J}_{ij} \langle r_i r_j \rangle + \mathcal{O}((N\delta))^3. \quad (35)$$

Using the fact that $\tilde{m}_i = \delta_i + \mathcal{O}(N\delta^2)$ from Eq. 32a, and $\phi_{ij} = \exp(\mathcal{J}_{ij}) - 1 = \tilde{C}_{ij}/(\tilde{m}_i \tilde{m}_j) - 1 + \mathcal{O}(N\delta)$ from Eq. 32b, we get Eq. 15a with g_{ind} defined as

$$g_{\text{ind}} \equiv \frac{1}{N(N-1)} \sum_{i<j} \left[\left(1+\rho_{ij}\right)\log\left(1+\rho_{ij}\right) - \rho_{ij} \right] \frac{\tilde{m}_i}{\delta}\frac{\tilde{m}_j}{\delta} \tag{36a}$$

$$\rho_{ij} = \frac{\tilde{C}_{ij}}{\tilde{m}_i \tilde{m}_j} - 1. \tag{36b}$$

For a pairwise model to the distribution in Eq. 27a, we have $\mathcal{H}_i^{\text{pair}} = \mathcal{H}_i + \mathcal{O}[(N\delta)^2]$ and $\mathcal{J}_{ij}^{\text{pair}} = \mathcal{J}_{ij} + \mathcal{O}[(N\delta)^2]$ and thus

$$S_{\text{pair}} - S = -\sum_{i<j<k} \phi_{ijk}\left(1+\phi_{ij}\right)\left(1+\phi_{ij}\right)\delta_i\delta_j\delta_j + \sum_{i<j<k}\mathcal{K}_{ijk}\left\langle r_i r_j r_k\right\rangle. \tag{37}$$

Using Eqs. 32c and 33, this will lead to Eq. 15a with g_{pair} defined as

$$g_{\text{pair}} \equiv \frac{1}{N(N-1)(N-2)} \sum_{i<j} \left[\left(1+\rho_{ijk}\right)\left[\log\left(1+\rho_{ij}\right) - \log\left(1+\rho_{ijk}^{\text{pair}}\right)\right] \right.$$
$$\left. - \left[\rho_{ijk} - \rho_{ijk}^{\text{pair}}\right] \right] \frac{\tilde{m}_i}{\delta}\frac{\tilde{m}_j}{\delta}\frac{\tilde{m}_j}{\delta} \tag{38a}$$

$$\rho_{ijk} = \frac{\tilde{C}_{ijk}}{\tilde{m}_i \tilde{m}_j \tilde{m}_k} - 1, \quad \rho_{ijk}^{\text{pair}} = \frac{\tilde{C}_{ijk}^{\text{pair}}}{\tilde{m}_i \tilde{m}_j \tilde{m}_k} - 1. \tag{38b}$$

THE INDEPENDENT-PAIR APPROXIMATION FOR THE EXTERNAL FIELDS IN THE LIMIT $\delta \to 0$

Replacing $\delta_i = \exp(\mathcal{H}_i)/[1 + \exp(\mathcal{H}_i)]$ in Eq. 32a and solving to find $\mathcal{H}_i$ we get

$$\mathcal{H}_i = -\log\left[\frac{1-\left\langle r_i\right\rangle}{\left\langle r_i\right\rangle}\right] - \sum_{j\neq i}\phi_{ij}\delta_j + \mathcal{O}[(N\delta)^2]$$

$$= -\log\left[\frac{1-\left\langle r_i\right\rangle}{\left\langle r_i\right\rangle}\right] - \sum_{j\neq i}\frac{\left\langle r_i r_j\right\rangle - \left\langle r_i\right\rangle\left\langle r_j\right\rangle}{\left\langle r_i\right\rangle} + \mathcal{O}[(N\delta)^2], \tag{39}$$

where in the second line we have used the fact that $\phi_{ij} = \frac{\left\langle r_i r_j\right\rangle}{\left\langle r_i\right\rangle\left\langle r_j\right\rangle} - 1 + \mathcal{O}(N\delta)$ from Eq. 32b and that $\left\langle r_i\right\rangle = \delta_i + \mathcal{O}(N\delta)$. Changing the variables from $r_i = 0,1$ to the original spin variables $s_i = \pm 1$, we have

$$h_i = \frac{\mathcal{H}_i}{2} + \sum_{j\neq i}\frac{\mathcal{J}_{ij}}{4} = \frac{1}{2}\log\left[\frac{1+m_i}{1-m_i}\right] - \sum_{j\neq i}\frac{C_{ij}}{4\left(1+m_i\right)} + \sum_{j\neq i}J_{ij}. \tag{40}$$

For the independent-pair (IP) approximation, on the other hand, we have

$$h_i^j = J_{ij} + \frac{1}{2}\log\left[\frac{1+m_i}{1-m_i}\right] - \frac{C_{ij}}{4\left(1+m_i\right)}. \tag{41}$$

Summing h_i^j over j and subtracting the over counted terms from single spin contributions gives the same expression as Eq. 40. Therefore, the IP approximation to the external fields get the leading term and the first order corrections of the low-rate approximation correctly.

INDEPENDENT-TRIPLET APPROXIMATION

Considering the following Gibbs distribution over three Boolean variables r_i, r_j and r_k,

$$p_{ijk}(r_i, r_j, r_k) = \frac{1}{Z_{ijk}}\exp\left[\mathcal{H}_i r_i + \mathcal{H}_j r_j + \mathcal{H}_k r_k + \mathcal{J}_{ij}r_i r_j + \mathcal{J}_{ik}r_i r_k + \mathcal{J}_{jk}r_j r_k\right], \tag{42}$$

we can use the definition of the mean values and correlations and write

$$\begin{pmatrix} \tilde{m}_i \\ \tilde{m}_j \\ \tilde{m}_k \\ \tilde{C}_{ij} \\ \tilde{C}_{ik} \\ \tilde{C}_{jk} \\ \tilde{C}_{ijk} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} p_{ijk}(0,0,0) \\ p_{ijk}(0,0,1) \\ p_{ijk}(0,1,0) \\ p_{ijk}(0,1,1) \\ p_{ijk}(1,0,0) \\ p_{ijk}(1,0,1) \\ p_{ijk}(1,1,0) \\ p_{ijk}(1,1,1) \end{pmatrix}. \tag{43}$$

Inverting the matrix of coefficients we can express the probabilities in terms of the mean values, pairwise and third order correlations. Using the result in

$$\mathcal{H}_i^{jk} = \log\left[\frac{p_{ijk}(1,0,0)}{p_{ijk}(0,0,0)}\right] \tag{44a}$$

$$\mathcal{J}_{ij}^{k} = \log\left[\frac{p_{ijk}(0,0,0)\,p_{ijk}(1,1,0)}{p_{ijk}(1,0,0)\,p_{ijk}(0,1,0)}\right] \tag{44b}$$

and similar equations for the other fields and couplings, we can also express these parameters in terms of mean values, pairwise and third order correlations. Before using Eqs. 48 and 49 as approximations for the parameters of a pairwise model, we need to perform two more steps.

The first step is a familiar one that we noted when dealing with the external fields in the IP approximation, namely the fact that $\mathcal{H}_i^{jk}$ depends on j and k in addition to i and that $\mathcal{K}_{ij}^k$ depends on k. We can use the same logic that we used in building an IP approximation to the external fields, and build an approximation to the external fields that does not depend on j and k, and an approximation to the couplings that does not depend on k. For example, for the case of the pairwise couplings we get

$$\mathcal{J}_{ij}^{\text{IT}} = \mathcal{J}_{ij}^{\text{IP}} + \sum_{k\neq i,j}\log\left[\frac{\left(\frac{1-\tilde{C}_{ij}}{\tilde{C}_{ij}}\right)\left[\frac{1-\left(\tilde{m}_k-\tilde{C}_{ik}-\tilde{C}_{jk}+\tilde{C}_{ijk}\right)}{1-\left(\tilde{m}_i-\tilde{m}_j-\tilde{C}_{ij}\right)}\right]}{\left[\frac{1-\left(\tilde{C}_{ik}-\tilde{C}_{ijk}\right)}{\tilde{m}_i-\tilde{C}_{ij}}\right]\left[\frac{1-\left(\tilde{C}_{jk}-\tilde{C}_{ijk}\right)}{\tilde{m}_j-\tilde{C}_{ij}}\right]}\right] \tag{45}$$

The second step has to do with the fact that Eqs. 44a and 44b (as well as their transformed version after performing the first step, e.g. Eq. 45) depend on the third order correlations in addition to the pairwise correlations and the mean values. Hence to derive an expression that relates model parameters to pairwise correlations and mean values we should first find the third order correlation in

terms of them. Note that this step is not present in the IP approximation. To express the third order correlations in terms of the lower order statistics we take advantage of the following equation

$$p_{ijk}(0,0,1)p_{ijk}(0,1,0)p_{ijk}(1,0,0)p_{ijk}(1,1,1)$$
$$= p_{ijk}(0,0,0)p_{ijk}(0,1,1)p_{ijk}(1,0,1)p_{ijk}(1,1,0). \quad (46)$$

Writing the probabilities in terms of the moments, this equation can be solved to find the third order correlations in terms of the mean values and pairwise correlations. The equations have two imaginary solutions for $\tilde{C}_{ijk}$, which are unphysical, and one real solution, which is the correct solution to be considered. The resulting expression for $\tilde{C}_{ijk}$ in terms of the mean values and pairwise correlations is complicated. However, in the limit of $\tilde{m}_i, \tilde{m}_j, \tilde{m}_k \to 0$, it can be shown to have the same form as the one reported in Eq. 33. We noted in the text that in this limit, the IP approximations to the couplings will give the same result as the leading order term of Eq. 32b for the couplings. With the independent-triplet (IT) approximation, we can go one step further, and as can be shown by doing a small amount of algebra, we can recover $\mathcal{O}(N\delta)$ corrections to the couplings in that we calculated in Eq. 32b.

As mentioned in Section "Extending the Independent-pair Approximation," one can continue the above process to build approximations based on quadruples of spins and so on. However, this soon becomes difficult in practice for the reason that solving equations of the type Eq. 46 to find the higher moments in terms of the mean values and pairwise correlations will be as difficult as the original problem of finding the external fields and couplings of the original N body problem in terms of the mean values and correlations. Nevertheless, this simple triplet expansion offers an alternative derivation of Eq. 15, simpler than the derivations in Roudi et al. (2009a) and Section "The Partition Function, Entropy and Moments of a Gibbs Distribution in the Limit $N\delta \to 0$" in Appendix. Here we show this for Eq. 15b, as deriving Eq. 15a will be similar but less involved. To derive Eq. 15b using the triplet expansion, we approximate the entropy of the whole system of N neurons as a sum of the entropies of all triplets denoted by S_{ijk}. We then expand the resulting expression keeping terms of up to $\mathcal{O}(\delta^3)$ noting that $\tilde{C}_{ijk} \sim \mathcal{O}(\delta^3)$, and $\tilde{C}_{ij} \sim \mathcal{O}(\delta^2)$. The result takes the form of

$$S = \sum_{i<j<k} S_{ijk} = -\sum_{i<j<k} \sum_{s_i,s_j,s_k} p_{ijk}(s_i,s_j,s_k) \log\left[p_{ijk}(s_i,s_j,s_k)\right]$$

$$= -\sum_{i<j<k} Q\left(\tilde{C}_{ijk}\right) - \tilde{C}_{ijk}$$

$$- \sum_{i<j<k} \tilde{C}_{ijk} \left[-\log\left(\tilde{C}_{ij}\right) - \log\left(\tilde{C}_{jk}\right) - \log\left(\tilde{C}_{ik}\right) \right.$$
$$\left. + \log\left(\tilde{m}_i\right) + \log\left(\tilde{m}_j\right) + \log\left(\tilde{m}_k\right) \right]$$

$$- \sum_{i<j<k} Q\left(\tilde{m}_i - \tilde{C}_{ij} - \tilde{C}_{jk}\right) + Q\left(\tilde{m}_i - \tilde{C}_{ij} - \tilde{C}_{jk}\right) + Q\left(\tilde{m}_i - \tilde{C}_{ij} - \tilde{C}_{jk}\right)$$

$$- \sum_{i<j<k} Q\left(\tilde{C}_{ij}\right) + Q\left(\tilde{C}_{ik}\right) + Q\left(\tilde{C}_{jk}\right)$$

$$- \sum_{i<j<k} Q\left(1 - \tilde{m}_i - \tilde{m}_j - \tilde{m}_k + \tilde{C}_{ij} + \tilde{C}_{ik} + \tilde{C}_{jk}\right) \quad (47)$$

where $Q(x) = x\log(x)$. For a pairwise model, the IT approximation to the entropy in the limit $\tilde{m} \to 0$ has the same form, except that $\tilde{C}_{ijk}$

of the true model should be replaced by $\tilde{C}_{ijk}^{\text{pair}} = (\tilde{C}_{ij}\tilde{C}_{ik}\tilde{C}_{jk})(\tilde{m}_i\tilde{m}_j\tilde{m}_k)^{-1}$ (see Eq. 33), i.e.

$$S_{\text{pair}} = -\tilde{C}_{ijk}^{\text{pair}}$$

$$- \sum_{i<j<k} Q\left(\tilde{m}_i - \tilde{C}_{ij} - \tilde{C}_{jk}\right) + Q\left(\tilde{m}_i - \tilde{C}_{ij} - \tilde{C}_{jk}\right)$$

$$+ Q\left(\tilde{m}_i - \tilde{C}_{ij} - \tilde{C}_{jk}\right)$$

$$- \sum_{i<j<k} Q\left(\tilde{C}_{ij}\right) + Q\left(\tilde{C}_{ik}\right) + Q\left(\tilde{C}_{jk}\right)$$

$$- \sum_{i<j<k} Q\left(1 - \tilde{m}_i - \tilde{m}_j - \tilde{m}_k + \tilde{C}_{ij} + \tilde{C}_{ik} + \tilde{C}_{jk}\right) \quad (48)$$

Using Eqs. 32c, 47 and 48 yields Eq. 15b.

DERIVATION OF TAP EQUATIONS FROM BELIEF PROPAGATION

In this appendix we derive the TAP equations (Eq. 12) starting from the Belief Propagation update rules. Let us begin with the result to be established. The TAP equations are a set of nonlinear equations for the magnetizations m_i, which we will write:

$$\tanh^{-1} m_i = h_i + \sum_{j\in\partial i}\left[\varepsilon J_{ij}m_j - \varepsilon^2 J_{ij}^2 m_i\left(1 - m_j^2\right)\right] + \mathcal{O}(\varepsilon^3). \quad (49)$$

Here h_i is the external field acting on spin i, εJ_{ij} are the pairwise couplings, and the notation $J \in \partial_i$ means that the sum is over neurons j connected with neuron i. As we also did in the text in Eq. 12, the above equation is generally quoted with ε set to one and without the error term of cubic order in ε.

Starting from the pairwise distribution (Eq. 2), we define the following distribution, with the auxiliary variable ε that we set to 1 in the end of our calculation

$$p^\varepsilon(\mathbf{s}) = \frac{1}{Z}\exp\left[\sum_i h_i s_i + \varepsilon \sum_{i<j} J_{ij}s_i s_j\right] \quad (50)$$

and the exact marginal distribution over spin s_i is defined by

$$p_i^\varepsilon(s_i) = \sum_{\mathbf{s}\backslash s_i} p^\varepsilon(\mathbf{s}) \quad (51)$$

where the sum goes over all spins except s_i. The exact magnetization of spin s_i is then

$$m_i = \frac{\sum_{\mathbf{s}} s_i \exp\left(\sum_i h_i s_i + \varepsilon \sum_{i<j} J_{ij}s_i s_j\right)}{\sum_{\mathbf{s}} \exp\left(\sum_i h_i s_i + \varepsilon \sum_{i<j} J_{ij}s_i s_j\right)}. \quad (52)$$

Belief Propagation is a family of methods for approximately computing the marginal distributions of probability distributions (Kschischang et al., 2001; Yedidia et al., 2003; Mezard and Montanari, 2009). Since the model defined by Eq. 50 contains only pairwise interactions, it is convenient to adopt the pairwise Markovian random field formalism of Yedidia, Freeman and Weiss (Yedidia et al., 2003). Note that the important recent contribution by Mezard and Mora uses a more general formalism, which may prove to be more convenient in the perspective of extending an "inverse BP" learning algorithm beyond pairwise models (Mora, 2007).

Belief Propagation applied to the model in Eq. 50 in the Yedidia–Weiss–Freeman formalism is built on probability distributions, called BP messages, associated with every directed link in the graph. If $i \rightarrow j$ is such a link, starting at i and ending at j, then the BP message $\eta_{i \rightarrow j}(s_j)$ is a probability distribution on the variable s_j associated to node where the link ends. For Ising spins one can use the parametrization

$$\eta_{i \rightarrow j}(s_j) = \frac{1 + s_j m_{i \rightarrow j}}{2}, \tag{53}$$

where $m_{i \rightarrow j}$ is a real number called the cavity magnetization. BP is characterized by two equations, the *Belief Propagation update equations*, and the *Belief Propagation output equations*. The BP update equations are used0 iteratively to find a fixed point, which is an extremum of the Bethe free energy. At the fixed point, the BP update equations form a (large) set of compatibility conditions for the η's which, for the model Eq. 50, read

$$\eta_{j \rightarrow i}(s_i) = \frac{1}{\Omega_{j \rightarrow i}} \sum_{s_j} e^{h_j s_j + \varepsilon J_{ij} s_i s_j} \prod_{k \in \partial j \setminus i} \eta_{k \rightarrow j}(s_j). \tag{54}$$

The BP output equations determine the marginal probability distributions from the η's and read

$$p_i^{\varepsilon}(s_i) = \frac{1}{\Omega_i} e^{h_i s_i} \prod_{j \in \partial i} \eta_{j \rightarrow i}(s_i), \tag{55}$$

where $\Omega_{j \rightarrow i}$ and Ω_i in Eqs. 54 and 55 are normalizations. To lighten the notation, we will not distinguish between the exact marginals, as in Eq. 51, and the approximate marginals from BP, as in Eq. 55.

We now write the BP update and BP output equations using the cavity magnetizations, $m_{i \rightarrow j}$, from Eq. 53. The notation simplifies if one define an ancillary quantity

$$q_{i \rightarrow j} = \frac{1}{2} \log \frac{1 + m_{i \rightarrow j}}{1 - m_{i \rightarrow j}} = \tanh^{-1} m_{i \rightarrow j} \tag{56}$$

in terms of which the BP update equation can be written as

$$m_{j \rightarrow i} = \tanh\left(\varepsilon J_{ij}\right) \tanh\left(h_j + \sum_{k \in \partial j \setminus i} q_{k \rightarrow j}\right) \tag{57}$$

and the BP output equation as

$$m_i = \tanh\left(h_i + \sum_{j \in i} q_{j \rightarrow i}\right). \tag{58}$$

The task is now to expand the right hand side of Eq. 58 in ε and compare with the TAP equations, Eqs. 49. To do this, we first note that

$$q_{j \rightarrow i} = \varepsilon J_{ij} \tanh\left(h_j + \sum_{k \in \partial j \setminus i} q_{k \rightarrow j}\right) + \mathcal{O}\left(\varepsilon^3\right) = \varepsilon J_{ij} m_j + \mathcal{O}\left(\varepsilon^3\right) \tag{59}$$

This follows from expanding Eq. 57 in ε, using the result in Eq. 56, and finally, expanding the logarithm. We then rewrite the BP output Eq. 58 as

$$\tanh^{-1} m_i = h_i + \varepsilon \sum_{j \in i} J_{ij} \tanh\left(h_j - q_{i \rightarrow j} + \sum_{k \in \partial j} q_{k \rightarrow j}\right) + \mathcal{O}\left(\varepsilon^3\right). \tag{60}$$

Since $m_j = \tanh(h_j + \Sigma_{k \in \partial j} q_{k \rightarrow j})$ (no expansion in ε) we want to separate $q_{i \rightarrow j}$ and $h_j + \Sigma_{k \in \partial j} q_{k \rightarrow j}$ in the arguments of the tanh'es in Eq. 60, and according to Eq. 59, $q_{i \rightarrow j}$ is of order ε. This means that we can write, to order ε,

$$\tanh\left(h_j - q_{i \rightarrow j} + \sum_{k \in \partial j} q_{k \rightarrow j}\right) = m_j - \varepsilon J_{ij} m_i \left(1 - m_j^2\right). \tag{61}$$

Introducing this into Eq. 60, we finally have

$$\tanh^{-1} m_i = h_i + \sum_{j \in i} \left[\varepsilon J_{ij} m_j - \varepsilon^2 J_{ij}^2 m_i \left(1 - m_j^2\right)\right] + \mathcal{O}\left(\varepsilon^3\right) \tag{62}$$

which was to be proved.

REFERENCES

Bethge, M., and Berens, P. (2008). Near-maximum entropy models for binary neural representations of natural images. In Advances in Neural Information Processing Systems, Vol. 20, J. Platt, D. Koller, Y. Singer and S Roweis, eds (Cambridge, MA, MIT Press), pp. 97–104.

Broderick, T., Dudik, M., Tkacik, G., Schapire, R. E., and Bialek, W. (2007). Faster solutions of the inverse pairwise ising problem. arXiv:07122437v2 [q-bioQM].

Crisanti, A., and Sompolinsky, H. (1988). Dynamics of spin systems with randomly asymmetric bonds: ising spins and Glauber dynamics. Phys. Rev. A 37, 4865–4874.

Ginzburg, I., and Sompolinsky, H. (1994). Theory of correlations in stochastic neural networks. Phys. Rev. E 50, 3171–3191.

Glauber, R. J. (1963). Time-dependent statistics of the ising model. J. Math. Phys. 4, 294–307.

Hinton, G. E. (2007). Boltzmann machine. Scholarpedia 2, 1668.

Hinton, G. E., and Sejnowski, T. J. (1983). Analyzing cooperative computation. In Proceedings of the 5th Annual Congress of the Cognitive Science Society, Rochester, NY.

Jordan, M. I. (1998). Learning in Graphical Models. Norwell, MA, Kluwer Academic Publishers.

Kappen, H. J., and Rodriguez, F. B. (1998). Efficient learning in Boltzmann machines using linear response theory. Neural Comput. 10, 1137–1156.

Kropff, E., and Treves, A. (2005). The storage capacity of potts models for semantic memory retrieval. J. Stat. Mech. Theory Exp. 2, P08010.

Kschischang, F. R., Frey, B. J., and Loeliger, H. (2001). Factor graphs and the sum-product algorithm. IEEE Trans. Inf. Theory 47, 498–519.

Lancaster, H. (1958). The structure of bivariate distributions. Ann. Math. Stat. 29, 719–736.

Mezard, M., and Montanari, A. (2009). Information, Physics, and Computation. Oxford, Oxford University Press.

Mezard, M., and Mora, T. (2008). Constraint satisfaction problems and neural networks: a statistical physics perspective. arXiv:08033061v1 [q-bioNC].

Mora, T. (2007). G'eométrie et inférence dans l'optimisation et en théorie de l'information. Ph.D. thesis, Université Paris Sud - Paris XI. http://tel.archives-ouvertes.fr/tel-00175221/en/

Nicolelis, M. A. L. (ed.) (2007). Methods for Neural Ensemble Recordings. New York, CRC Press.

Okatan, M., Wilson, M. A., and Brown, E. N. (2005). Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. Neural Comput. 17, 1927–1961.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 45, 995–999.

Roudi, Y., Nirenberg, S., and Latham, P. E. (2009a). Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. PLoS Comput. Biol. 5: e1000380. doi:10.1371/journal. pcbi.1000380.

Roudi, Y., Tyrcha, J., and Hertz, J. (2009b). The ising model for neural data: model quality and approximate methods for extracting functional connectivity. Phys. Rev. E. 79, 051915.

Russo, E., Namboodiri, V. M. K., Treves, A., and Kropff, E. (2008). Free association

transitions in models of cortical latching dynamics. *New J. Phys.* 10, 015008.

Sarmanov, O. (1963). Maximum correlation coefficient (nonsymmetric case). In: Selected Translations in Mathematical Statistics and Probability, Vol. 4 (Providence, RI, American Mathematical Society), pp. 271–275.

Schneidman, E., Berry, M., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440, 1007–1012.

Sessak, V., and Monasson, R. (2009). Small-correlation expansions for the inverse ising problem. *J. Phys. A* 42, 055001.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., and Chichilnisky, E. J. (2009). Spatial organization of large-scale concerted

activity in the primate retina. *J. Neurosci.* 29, 5022–5031.

Shlens, J., Field, G. D., Gauthier, J. L., Greschner, M., Sher, A., Litke, A. M., and Chichilnisky, E. J. (2006). The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* 26, 8254–8266.

Tanaka, T. (1998). Mean-field theory of Boltzmann machine learning. *Phys. Rev. E* 58, 2302–2310.

Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., Prieto, A., Petrusca, D., Grivich, M. I., Sher, A., Hottowy, P., Dabrowski, W., Litke, A. M., and Beggs, J. M. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.* 28, 505–518.

Thouless, D. J., Anderson, P. W., and Palmer, R. G. (1977). Solution of 'solvable model of a spin glass'. *Philos. Mag.* 35, 593–601.

Tkacik, G., Schneidman, E., Berry, M. J. II, and Bialek, W. (2006). Ising models for networks of real neurons. arXiv: q-bio/0611072v1 [q-bioNC].

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* 93, 1074–1089.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2003). Understanding belief propagation and its generalizations. In Exploring Artificial Intelligence in the New Millennium, G. Lakemeyer and B. Nebel, eds (San Francisco, CA, Morgan Kaufmann Publishers).

Yu, S., Huang, D., Singer, W., and Nikolic, D. (2008). A small world of neuronal synchrony. *Cereb. Cortex* 18, 2891–2901.