A Statistical Physics Approach to Modelling the Joint Activity of Cortical Populations

Marius B. Mahiout

In this thesis we apply likelihood maximization methods as well as mean-field approximations to construct equilibrium and non-equilibrium Ising models for neural activity in the form of calcium imaging recordings. Before applying these methods to the neural data, we carried out tests on samples generated from the models themselves. We also assessed statistical properties of the neural recordings, as well as the performance of the inference methods under various circumstances. While the mean-field methods appear to break down for large numbers of neurons, the exact likelihood maximization procedures for these two models are reliable also for larger system sizes. Our findings indicate that both the equilibrium and non-equilibrium Ising models are capable of capturing the essential statistical features of the calcium imaging recordings, not only reproducing constraint observables but also predicting statistical properties of the neural activity not used to fit the models. There were clear differences between the models, however, with the non-equilibrium model outperforming its equilibrium counterpart on generalization.

Acknowledgments

I would like to express my deepest gratitude to Dr. Yasser Roudi for his invaluable guidance and insightful discussions, which were instrumental to forming the foundation of this project. His feedback during the editing and in preparation for the thesis defense was also greatly appreciated. I also wish to thank Dr. Ivan A. Davidovich for his valuable feedback in the editorial process, and for his assistance in developing a more efficient procedure for calculating third-order covariances.

Contents

1 Introduction

1.1 Biological neural networks

1.1.1 Population activity and cortical computations

One of the overarching problems of neuroscience is that of determining how the brain processes information. Although some findings show that glial-cells may also participate in information processing [1], the general consensus is that the *principal* computational unit of the brain is the neuron [2].

The human brain contains on the order of 10^{11} neurons which, in turn, have been classified into more than 1000 different types, based on their morphology [2]. Certain faculties, such as vision and other sensory modalities do depend crucially on very specific kinds of neurons, like the rod and cone cells of the retin [2]. It is possible that cell morphology also plays a role in cortical information processing. Nevertheless, here we follow the connectionist approach, and assume that the variety in morphology of nerve cells is not of crucial importance for understanding the general principles of brain function [3].

In this view, the complexity of behaviors observed in the animal world are thought to arise as a consequence of the organization of neurons into networks that, depending upon their connectivity, carry out specific computations [2, 3]. Indeed, the approach taken here involves the use of methods originally conceived for the study of systems constituted by a large number of identical elements whose behavior depends upon the average behaviors and interactions of these elements. Hence, much of the following depends upon the assumption that this approach can be fruitfully applied to the study of nervous systems.

1.1.2 Neurobiological preliminaries

While we will not concern ourselves with the aspects of neuronal morphology that distinguishes one kind of neuron from another, if we are to understand how neurons are organized into networks, it will be necessary to be familiar with their basic anatomy. The canonical neuron can be said to have the following three parts [2].

1. Neurons are eukaryotic cells, and as such, have a nucleus housing its genetic material. This resides in the part of the neuron referred to as the *soma* (or *cell body*), and is where most protein synthesis occurs.

In addition to the soma, two kinds of structures, collectively termed *neurites*, emanate from the soma:

2. Multiple *dendrites*, which branch out in a tree-like fashion, and are typically where a neuron receives signals from other neurons.

3. A single *axon*, which is projected onto the dendrites of other neurons, and is the route by which a neuron sends signals to other neurons.

The site of connection, where the axonal membrane of one neuron is brought into close proximity with the dendritic membrane of another, is called a *synapse*.

Like all cells, neurons maintain a negative electric potential with respect to the extracellular fluid. This arises arises as a consequence of the relative concentrations of various ions on either side of the membrane: Na^+, Ka^+, Cl^-, and Ca^{2+} are of particular importance. This negative potential is called the resting potential, and resides around -70 mV for most neurons.

In the event that the neuronal membrane reaches an elevated threshold of about -55 mV, *e.g.*, as a consequence of excitatory input form other neurons, a feedback process is initiated, which causes the potential to go all the way up to about $+40$ mV. This is referred to as an *action potential*, and neurons exhibiting action potentials are said to be active, or firing[1].

[1] Following the peak of the action potential, negative feedback mechanisms lower the membrane potential to a below-resting voltage. In this state, the neuron is said to be *hyperpolarized*, and is less susceptible to excitatory inputs, until it eventually returns to the resting potential.

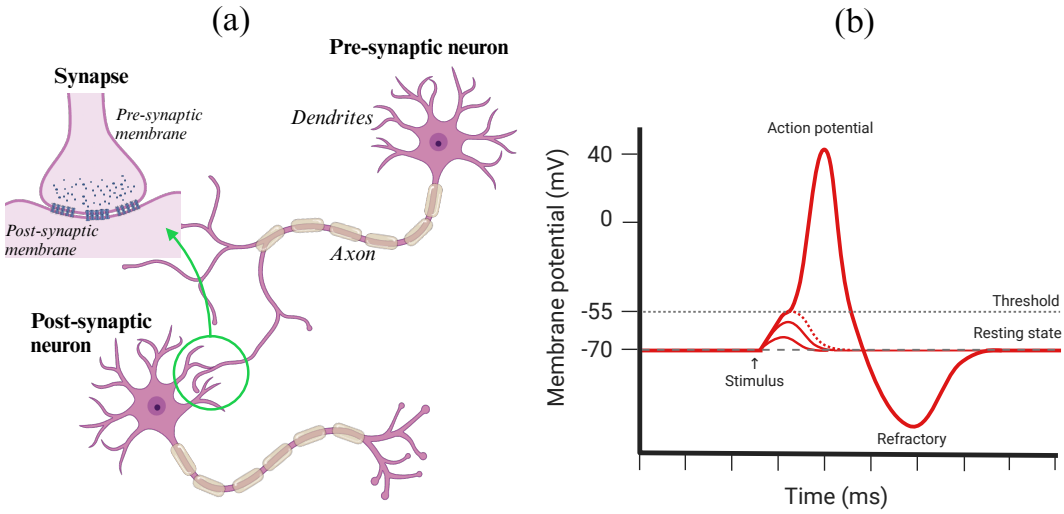


Figure 1: **(a)** Diagrammatic illustration of basic neuronal anatomy, with a synaptic connection between two neurons as described in the text. **(b)** A graph showing the general shape of an action potential. The whole process takes roughly 1-10 ms [2]. Created with BioRender.com.

It should also be noted that many neurons fire rapid bursts of multiple action potentials in sequence [2]. It is then common to say that the neuron is *active* if it is firing such bursts, and *inactive* (or *silent*) otherwise.

The action potential usually arises in the part of the soma at the base of the axon, called the *axon hillock*. From here, it propagates down the axon, until it reaches the locations at which the axon forms synapses with the dendrites of other neurons. When the action potential reaches the synapse, the electric signal is typically transduced into a chemical signal in the form of neurotransmitters, which are released from the *presynaptic membrane* (of the emitting neuron), and diffuse across the synaptic cleft to bind to receptors on the *postsynaptic membrane* (of the receiving neuron).

Neurotransmitter binding on the postsynaptic membrane can give rise to a wide variety of effects, including more complicated actions, such as regulation of gene expression, which is known to play an important role in learning [2]. On a simpler level, however, this chemical signaling lead to an increase or decrease of the voltage across the membrane of the receiving neuron, bringing it closer to, or further away from the threshold potential, referred to as *excitation* and *inhibition*, respectively. Typically, a particular neuron is only involved in one of these two kinds of signaling, and so one talks about excitatory and inhibitory neurons, and excitatory or inhibitory synapses. Finally, the efficiency with which one neuron excites or inhibits another neuron varies from synapse to synapse.

1.1.3 Idealized neurons and networks

Many models of neural networks, including the ones used here, are weighted graphs, whose nodes correspond to neurons (often called *units*, in this context), and whose edges correspond to connections (or *couplings*) between neurons. The numerical value of the weights indicate the strength of connection, and whether or not the connection is excitatory (if the weight is positive) or inhibitory (if it is negative).

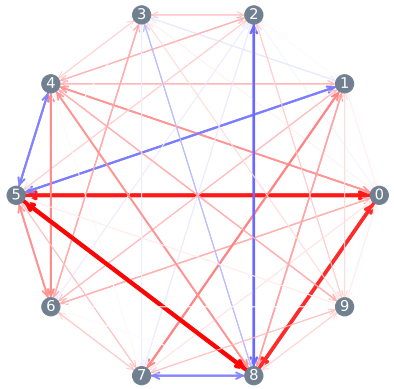


Figure 2: An illustration of an idealized neural network. Units are depicted as nodes, and their connections as arrows whose color and thickness indicates their type (red $\sim$ excitatory, blue $\sim$ inhibitory) and relative strength, respectively.

Although the models treated here will be fully connected, the biological networks of the brain are typically *sparse*, in the sense that one neuron usually connects with a comparatively small subset of the rest of the neurons in the brain. For this reason, rather than being on the order of 10^{22}, the number of connections in the human brain is only about 10^{14}. In our models, the absence of connections are accounted for by connection weights that are zero, or very close to zero.

Moreover, although neuronal signaling is directed, one of the two classes of models we'll be employing in the following imposes the restriction that the connection weights are symmetric. While this is a simplification, and a step away from biological realism, it has the advantage of ensuring the existence of an energy function which can be useful in investigating the patterns of activity seen in neural networks [4]. Further considerations regarding the relationship between biological neural networks and the idealizations used here will be discussed in Sec. 1.3.2.

1.2 Experimental set-up and pre-processing

The data analyzed in this study is in the form of approximate firing rates, inferred from 1-photon calcium imaging recordings from the secondary motor cortex (M2) and posterior parietal cortex (PPC) areas of mice that were either performing or observing a behavioral task. While a brief overview will be given here, both the imaging and the subsequent procedures used to infer the firing rates were carried out by Tombaz and her collaborators, as detailed in [5].

1.2.1 Calcium imaging

Calcium imaging is a method used to capture images which reflect the amounts of Ca^{2+} present in biological tissue. While there are various implementations of this idea, the experiment under consideration employs a technique in which viral vectors that have been engineered so as to express the calcium indicator GCaMP6m —a protein which fluoresces in the presence of Ca^{2+}— are injected into the brain regions of interest [5]. Upon injection, the viruses transfer their DNA into the cells in the region around the injection site, leading the cells to manufacture the GCaMP6m themselves. Now, Ca^{2+} is released in abundance when neurons fire action potentials, and so, images captured using a microscope indicate the relative firing rates of the imaged neurons.

1.2.2 Pre-processing

These calcium images, of course, do not provide a direct measure of the firing rates of the recorded neural population. Approximate firing rates must be inferred from the images, which is done through a process that can be understood in terms of two steps [5]:

1. First, one has to infer the calcium signal of each imaged neuron. This involves separating the calcium signal of the neurons themselves from the background activity, as well as from random fluctuations.

2. Next, once the calcium signals have been extracted, the firing rates can be inferred by deconvolving the calcium signals.

To extract the calcium signals from the images, a denoising, deconvolution, and demixing procedure called *'Constrained Non-negative Matrix Factorization for Endoscopic recordings'* (CNMF-E) [6] is used, and the firing-rates were inferred from the calcium signals using a type of deconvolution method belonging to the class of methods called *'Online Active Set methods for Spike Inference'* (OASIS) [7].

To give a brief outline of the procedure, $Y \in \mathbb{R}_+^{d\times T}$, where d is the number of pixels and T is the number of captured frames, is modelled using the matrix equation,

$$Y = AC^T + B + E,$$

where $A \in \mathbb{R}_+^d$, $C \in \mathbb{R}_+^T$, and $B,\ E \in \mathbb{R}_+^{d\times T}$ correspond to the spatial features of the recorded neurons, the calcium signal itself, the background activity signal, and random noise, respectively. CNMF-E is used to estimate A and C, by optimizing a particular set of variables, subject to a set of constraints which are constructed so as to reflect the spatiotemporal dynamics of neural calcium signals. The state-of-the-art for 2-photon imaging data is CNMF, which is based on a simpler model for the background activity B. CNMF, however, is often insufficient when dealing with noisier 1-photon captures, and so CNMF-E —based on a more accurate model for the background activity— gives better results.

Having isolated the calcium signal of the individual neurons with CNMF-E, the firing rates were inferred by modeling the calcium signal using an *autoregressive process* of order 1, $AR(1)$. More details on the pre-processing using CNMF-E and OASIS can be found in refs. [6] and [7], respectively.

1.2.3 Behavioral paradigm

As stated, the recordings were taken from mice that were either engaged in, or observing some behavior. Images were captured from a miniaturized head-mounted microscope, and observing mice were head-fixed. Moreover, in each session, the mice were placed into two contexts [5]:

1. In the first context, performing mice were tasked with turning in a circle to elicit pellet delivery, followed by reaching for and, subsequently, eating the food pellet. These sessions lasted for 15 minutes.

2. Following the pellet-reaching sessions, the performing mice were placed in an open environment for 20 minutes, where they were allowed to behave freely, and had access to a cartwheel on which they could run.

Each animal was subjected to both performing and observing sessions, and the sessions were interleaved so as to alternate between performing and observing.

1.3 Analyzing multi-unit recordings

We've been able to record from individual neurons since 1950s, when the first iterations of the micro-electrode were developed [8, 9].

Although work such as Hubel and Wiesel's investigations of the visual system demonstrate that some understanding can be gained from studies based on single-unit recordings [10], the amount by which such studies can illuminate the principles of brain functioning is inherently limited by the fundamentally coupled relationship between neurons. While some researchers have attempted to derive conclusions about the brain-basis of animal behavior from the tuning curves of individual neurons [11, 12], as argued in [13], such

conclusions may be misguided for multiple reasons, one of which is that notions such as preferred stimulus may be attributed to the experimental design, more so than to the properties of the neuron itself.

As argued in Sec. 1.1, the neurobiological processes underlying cognition and behavior are likely to arise at the population level. Consequently, an understanding of how cognition and behavior is realized by the brain requires us to study the joint activity of large neuronal populations.

Since the 1950s, various instruments and methods that allow us to record the concerted activity of non-trivially sized populations, including multi-electrode arrays and fluorescence imaging, have been developed [14, 15]. While being able to make measurements is a necessary condition for beginning to understand the brain basis of mind, it falls short of being a sufficient condition. We must, of course, also be able to extract useful information from these measurements.

Neural networks being complex systems, it seems reasonable to investigate them using a general approach that has been successful in understanding other complex systems. One such approach is that of statistical physics, which has been fruitfully applied to the study of physical systems, such as gases, that have a large number of degrees of freedom.

In the following, we will pursue an approach based on a mathematical framework that was first applied to statistical mechanics by Jaynes, and has been referred to as predictive statistical mechanics [16]. With a statistical approach to the study of neural networks, a natural first step is to determine the probability distribution over network states, as knowing this distribution is a prerequisite to other questions about the statistical features of neural activity.

Although mechanistic models of brain function might be more desirable in some respects, the successful development of many such models likely requires a deeper understanding of the basic principles underlying the functioning of the nervous system. The hope is that the kind of approach pursued here will contribute to revealing some of these principles.

1.3.1 Probability distribution over activation patterns

Specifying the probability distribution over activation patterns by directly calculating the relative frequencies of the various possible states is an experimentally intractable problem, because this would require multiple observations of every possible state.

Here, we take neurons to be either active or inactive, so for a population of N neurons, the number of possible states is 2^N, which very quickly becomes an enormous number —even for a moderately sized network. Consequently, we require a more indirect method which will allow us to determine a sensible statistical model from limited observations.

As it turns out, the *predictive statistical mechanics* of Jaynes provides precisely such a method. This method requires us to choose a set of constraints, in the form of observables that can be realistically calculated from the neural recordings. While we must rely on scientific intuition to select these constraints, once a choice has been made, there exists a natural criterion by which we can select the most appropriate distribution, among the potentially infinite number of distributions that satisfy our set of constraints.

As will be argued shortly, the natural model will be that which, subject to our constraints, maximizes the Shannon entropy. This will yield a parametrized distribution whose parameters include a set of what will be referred to as couplings, or connection weights.

It is well established that one of the basic brain processes underlying learning and memory is the modification of synaptic connections [2, 17]. Thus, since thinking and behavior is thought to arise from neural activity, we expect the connectivity of a neural network to constrain the activation patterns exhibited by the network. This can be understood in terms of a mapping:

$$\boxed{\text{network connectivity}} \longrightarrow \boxed{\text{neural activity}}$$

Extending this line of thinking, the problem of fitting a model parametrized by connection weights can be thought of as a matter of specifying the inverse mapping:

$$\boxed{\text{network connectivity}} \longleftarrow \boxed{\text{neural activity}}$$

In other words, our problem may be conceptualized as that of inferring the connection structure of a neural network by looking at its activity. While this is an attractive metaphor, it must be regarded as such, and one

must be careful to avoid being tempted by the suggestive terminology of "connections" into making stronger claims than can be reasonably justified. Hence, before we elucidate the principle of maximum entropy, we will embark on a brief detour to highlight some caveats, and elaborate on the relationship between network connectivity and the parameters of the models that will be introduced.

1.3.2 Caveats and functional connectivity

As discussed in the previous section, when faced with a model of neural activity whose parameters include something that is commonly referred to as connection strengths, it is tempting to link these to the physical synaptic connections that are found in the brain. While it is important to keep in mind that we're not making use of any direct measurements of the physical structure of the networks, such as the presence or absence of synaptic connections, it is natural to wonder if there nevertheless exists a relation between these parameters and synaptic connections, beyond the shared terminology.

As will become clear in subsequent sections, the value of the coupling parameter, J_{ij}, corresponding to two recorded neurons, i and j, is related to the covariance in the activation patterns of the two neurons. Thus, the principle of synaptic potentiation which originated with Hebb [18], often stated in the form of the well-known mnemonic "cells that fire together, wire together"[2] suggests that yes, the parameters we call couplings could indeed be indicative of real synaptic connections.

Now, while a synaptic connection between neurons i and j could cause their activation patterns to be correlated, a signal transmitted from neuron i to neuron j is not instantaneous, so we wouldn't necessarily expect to see simultaneous activation. This is easily accounted for by looking at delayed correlations instead of, or in addition to, simultaneous correlations.

A more pertinent concern, however, and one that cannot be dealt with as easily, is that of confounding variables. Finding a correlation between neurons i and j may not be due to a synaptic connection, for it could just as well indicate a common input from a set of neurons that were not recorded. As it stands, we do not have a way to distinguish between these two situations purely on the basis of correlations, and so any claims we make about the connectivity structure of a network should be understood as claims about so called functional connectivity which may, or may not, be due to synaptic connections.

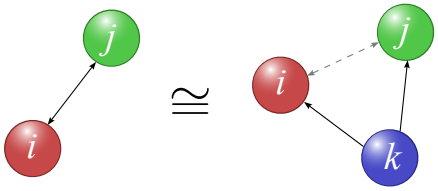


Figure 3: A confounding variable k (*e.g.,* a neuron, or population of neurons) can lead to a percieved dependency between two variables i and j.

Another issue that should be mentioned, which like the subject of functional vs. synaptic connectivity concerns the interpretations and concepts we associate with the mathematical scaffolding, is the notion of neural coding. A lot of work involving investigations of population activity is presented in the framework of neural coding, which subsumes the idea that neural activity constitutes a kind of code that has to somehow (and somewhere) be decoded in order to give rise to cognitive, perceptual or behavioral representations [19]. In a similar vein, some researchers hold that the kinds of probability distributions under investigation here are used not only by scientists in trying to understand neural activity, but are computed by the brain itself, and used to represent the uncertainty associated with actions and perception [20].

As argued in [13], this conceptual framework has certain problematic aspects, and may not be the right way to understand the relation between the mind and the brain. An alternative perspective is that neural activity, rather than being like the symbols of a code, is more appropriately understood as actions. Hence, to

[2]There is also a more recent analogue for synaptic depression: "cells that fire out of sync, lose their link".

avoid potentially misguided conceptual assumptions, we will take the more pragmatic approach, and remain neutral as to whether outcomes associated with the probability distributions we talk about correspond to "actions", "symbols", or perhaps, something entirely different.

On a less philosophical note, calcium signals are continuous. Thus, to study spike train statistics, one has to impose a discretization. More specifically, time is divided into discrete intervals (or bins) and, for a given neuron, if one or more calcium events are found to occur within the boundaries of this interval the neuron is taken to be active within the interval, and inactive otherwise. Moreover, different bin widths can give rise to different model parameters.

1.3.3 Maximum entropy modelling

We now turn to the issue of justifying the claim that once we've chosen a set of constraints, we should select the model which, subject to those constraints, maximizes the Shannon entropy.

Clearly, given multiple models, all of which agree with the empirical observations to the same extent, it is most "scientifically honest" to choose the least biased model. Moreover, since certainty, or conviction, typically accompanies bias, and *vice versa*, we will use uncertainty as a proxy for un-biasedness. The problem then reduces to showing that the information entropy (or, as we've called it until now, Shannon entropy) is the appropriate way to quantify uncertainty.

For a set $\{x\}$ of outcomes, with corresponding probabilities $\{P(x)\}$, we denote the *information* (or *Shannon*) *entropy* over the probability distribution $\{P(x)\}$ by

$$S(x) := -\sum_x P(x) \ln P(x). \tag{1}$$

Note that while we've defined the entropy in terms of the natural logarithm –corresponding to the natural unit of information, another logarithm base merely amounts to a different unit of information.

To establish that $S(x)$ can be used as a measure of uncertainty, we give a brief outline of the original argument, presented by Shannon in [21]. Firstly, as demonstrated in Appendix 2 of [21], the entropy function as defined by Eq. (1) is, up to a constant factor (again, corresponding to the unit of information), the unique function S which satisfy the following 3 conditions:

1. S is a continuous function of $P(x)$.
2. If $P(x)$ is the uniform distribution over N outcomes, S is a monotonically increasing function of N.
3. If we divide the outcomes $x = \{x_1, \ldots, x_N\}$ into two groups $y_1 = \{x_1, \ldots, x_k\}$ and $y_2 = \{x_{k+1}, \ldots, x_N\}$, with associated probabilities $P(y_1) = \sum_{i=1}^{k} P(x_i)$ and $P(y_2) = \sum_{i=k+1}^{N} P(x_i)$, respectively. Then

$$S(x) = S(\{y_1, y_2\}) + P(y_1)S(x|y_1) + P(y_2)S(x|y_2),$$

where for two sets of outcomes, $\{x\}$ and $\{y\}$, $S(x|y) := -\sum_x P(x|y) \ln P(x|y)$.

These conditions are meant to capture the essential quantitative aspects of uncertainty. The continuity condition, 1., can be understood in terms of the idea that a small change to one's beliefs about the relative likelihoods of the outcomes of x should be accompanied by a correspondingly small change in one's uncertainty about which outcome of x will occur. Condition 2. corresponds to the intuition that if all outcomes are equally likely, then more choices of outcomes should be accompanied by more uncertainty regarding which outcome will occur. Finally, condition 3. captures the idea that if we divide the original set of outcomes into two groups, then the uncertainty about which among the original set of outcomes will occur should be the sum of the uncertainty of the groups and the uncertainty associated with the outcomes within each group.

Next, it is easily seen that S is positive, is maximized by the uniform distribution, and is minimized (zero, in fact) by degenerate (or Delta) distributions, where one outcome occurs with probability 1, and all other outcomes have zero probability. This, of course, is consistent with our intuition that one should be most uncertain if every possible event is equally likely to occur, and, tautologically, least uncertain if the outcome known *apriori* [21].

In light of this, we consider ourselves justified in using the Shannon entropy as a measure of un-biasedness. This being established, Jaynes' conclusion, that the only scientifically honest choice is the model which

maximizes the entropy [22] follows. It remains to choose a set of constraints for our model. Motivation for our choice will be provided in Sec. 1.4.2. For now, however, we simply state that our requirement will be that the statistical model should have the same first and second order moments as the neural recordings, in other words: the mean firing rates of the model neurons should agree with those of the recorded neurons, and the covariance between every pair of model neurons should agree with the covariance between the corresponding pair of recorded neurons.

Denoting the state of neuron i by $\sigma_i = \pm 1$ if it is active or inactive, respectively, the network state $\vec{\sigma} = (\sigma_1, \ldots, \sigma_N)$ can be viewed as random variable whose range consists of 2^N outcomes. With this formalism, the constraints can be written as:

$$\forall i \in \{1, \ldots, N\}, \quad \langle \sigma_i \rangle_{\text{Model}} := \sum_{\vec{\sigma}} \sigma_i P(\vec{\sigma}) = \langle \sigma_i \rangle_{\text{Data}}, \tag{2}$$

and,

$$\forall i, j \in \{1, \ldots, N\}, \quad \langle \sigma_i \sigma_j \rangle_{\text{Model}} := \sum_{\vec{\sigma}} \sigma_i \sigma_j P(\vec{\sigma}) = \langle \sigma_i \sigma_j \rangle_{\text{Data}}. \tag{3}$$

We can then follow [22], and use the method of Lagrange multipliers to maximize the entropy $-\sum_{\vec{\sigma}} P(\vec{\sigma}) \ln P(\vec{\sigma})$, subject to Eqs. (2) and (3). This yields the following maximum entropy distribution:

$$P(\vec{\sigma}) = \frac{1}{Z} \exp\Big(\sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \Big), \tag{4}$$

where the normalization (or *partition function*) is given by

$$Z = Z(\vec{h}, \mathbf{J}) = \sum_{\vec{\sigma}} \exp\Big(\sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \Big). \tag{5}$$

Thus, we obtain a version of what can be recognized as the Ising model. The Ising model was originally conceived by different means, and in an entirely different context. It was later applied to neuroscience, though at first, for the purpose of describing a particular cognitive function. The next section provides some of this historical background.

1.4 The Ising model

1.4.1 A model for magnetism

In 1895, the French physicist Pierre Curie discovered that permanent magnets (ferromagnets) lose their magnetization when heated above a particular temperature[3] [23], now known as the *Curie temperature* [24].

Consistent with Stigler's law, which states that "no discovery is named after the first person to discover it", the original idea behind the Ising model was first conceived by Wilhelm Lenz in 1920, though it was worked out in greater detail by his pupil, Ernst Ising, after whom the model is named [25]. Upon Lenz' suggestion to use the model to investigate the phase transitions exhibited by ferromagnets at the Curie temperature, Ising solved it in the case of a 1-dimensional lattice. Unfortunately, however, he found that the model exhibits no phase transitions in this case, and assumed this would generalize to higher dimensional lattices [25].

This, however, was later shown by Peierls to be incorrect [26, 25], and indeed, around two decades later, in 1944, Lars Onsager managed to calculate the partition function for the Ising model for a 2-dimensional lattice, and demonstrated that it does exhibit the desired phase-transition [27, 25].

[3] The value of the Curie temperature depends on the material constituting the magnet.

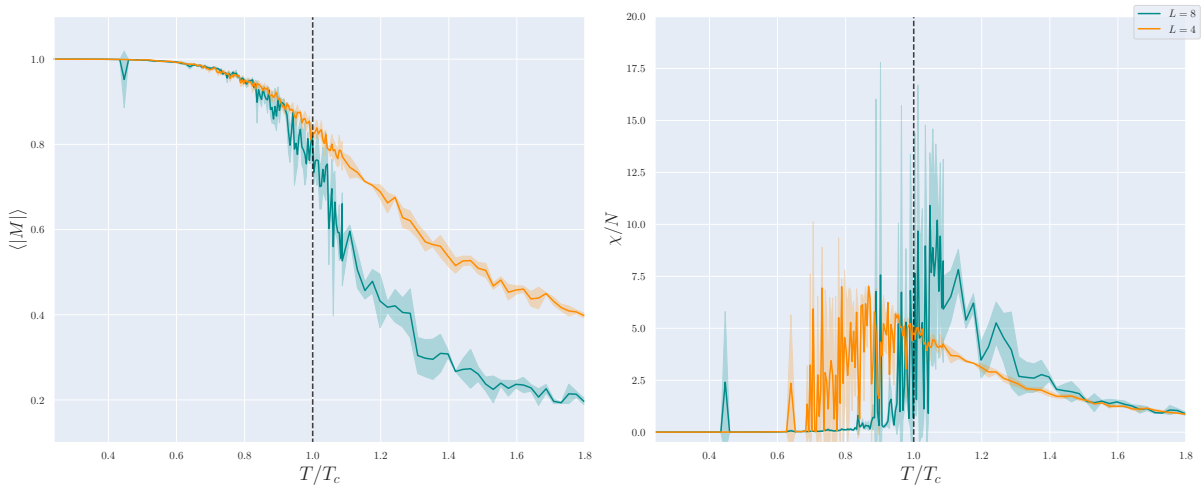


Figure 4: Phase transition in the 2-D Ising model. Mean absolute magnetization $\langle |M| \rangle$, and magnetic susceptibility (per spin) χ/N vs. temperature in units where the Curie (or *critical*) temperature equals 1, T/T_c. The absolute magnetization is known to fall abruptly to zero, whereas the susceptibility exhibits a peak at the critical temperature. In each case, we've plotted curves for lattices with side-lengths L set to 4 (cyan) and 8 (orange), with periodic boundary conditions. The model parameters were $J = 1/T$ and $h = 0$. The observables were calculated on the basis of 3 runs of $15,000$ simulations for $T > 2.3$, and 1 million simulations at lower temperatures.

It is possible to obtain this model also in the physical context, by maximizing the entropy using the method of Lagrange multipliers, like we did in the preceding section, with the exception that here, the covariance constraint, Eq. (3), is imposed only for neighboring lattice sites. Jaynes' maximum entropy approach to statistical physics, however, was proposed only in 1957 [22], long after the Ising model was first invented. The Ising model was first constructed by setting up the *Hamiltonian*,

$$\mathcal{H}(\vec{\sigma}) = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j - h \sum_i \sigma_i, \tag{6}$$

where lattice sites correspond to atoms whose spin can be either up or down, rather than neurons that are active or inactive. The notation $\langle i, j \rangle$ indicates that the sum is restricted to neighboring lattice sites i and j. J corresponds to interactions between the atomic magnets, and h corresponds to an external magnetic field.

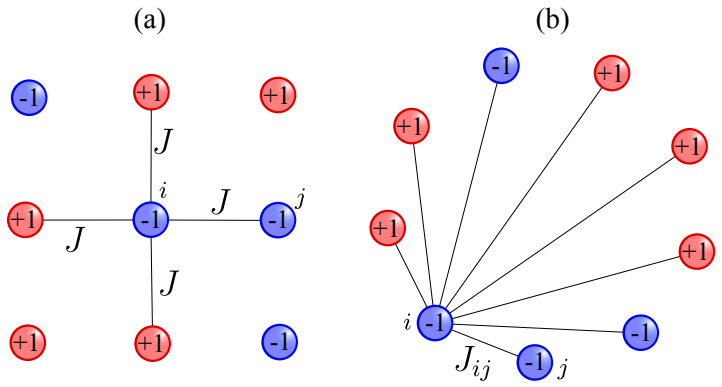


Figure 5: An illustration of the difference between the two versions of the Ising model discussed so far. **(a)** The classical 2-D lattice Ising model, where a unit i connects only to its neighbors (the ones above, below, and to the left and right of it). **(b)** The fully-connected version of the model that we'll employ, where a unit i has a connection with every other unit j. As mentioned previously, the fully connected models are not necessarily *fully connected* in the sense that all units influence one another directly, because some units might have a coupling J_{ij} with zero value, which represents the absence of a connection.

Similarly, the Hamiltonian corresponding to our version of the model is given by

$$\mathcal{H}(\vec{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i. \tag{7}$$

While we've already addressed the interpretation of the J_{ij} in Eqs. (4) and (7), we note in passing that the h_i may be conceptualized as a firing threshold, or an external input. Now, the distribution over spin configurations for a magnetic system is obtained by substituting the Hamiltonian, Eq. (6), into the *Boltzmann distribution*,

$$P(\vec{\sigma}) = \frac{1}{Z}\exp\big[-\beta\mathcal{H}(\vec{\sigma})\big], \tag{8}$$

$$Z(\beta) = \sum_{\vec{\sigma}}\exp\big[-\beta\mathcal{H}(\vec{\sigma})\big], \tag{9}$$

where β is inversely proportional to the temperature. Our model can be retrieved in the same way, by substituting the Hamiltonian of Eq. (7) into the Boltzmann distribution, except we have chosen to subsume the temperature by the couplings and fields, and so β would be set to 1 in this case.

Note also that, while the models considered in the neuroscientific context do not restrict which units can be connected based on a notion of lattice proximity, Eq. (4) can be understood as describing the Ising model for an infinite-dimensional lattice, where all lattice sites are neighbors.

Before moving on to the applications of Ising models in neuroscience, we take a moment to further emphasize the advantages gained from the origins of the Ising model in physics. As will be explored further later, a wide range of powerful techniques have been developed to facilitate the study of systems such as the Ising model, including a family of methods referred to as mean field theory. The basic idea of mean field theory is to approximate the solution to a many-body problem by replacing it with a corresponding one-body problem which, generally, is much easier to solve. In the case of the Ising model, this takes the form of replacing all of the interactions between pairs of spins with a single *effective interaction*.

$$\sigma_i \overset{\forall j\neq i}{\longleftrightarrow} \sigma_j \quad \longrightarrow \quad \sigma_i \longleftrightarrow \boxed{\text{rest of system}}$$

In the physical context, the methods of mean field theory, as well as other techniques, are generally used to solve problems in which the parameters of the model are known, and the goal is to calculate partition functions, or statistical observables. Collectively referred to as forward statistical problems [28].

By contrast, we will primarily be concerned with the so called *inverse problem*, in which the observables are known, and the goal is to infer the parameters $\{\vec{h}, \mathbf{J}\}$ of the model. Fortunately, it turns out that many of the techniques developed for the purpose of solving forward problems can be re-purposed for the inverse problem.

1.4.2 A model for neural networks

Applying Bayes' theorem to Eq. (4), the probability that a neuron $i \in \{1, \ldots, N\}$ assumes a particular value $\sigma_i = \pm 1$, provided that the rest of the network has a given configuration, $\{\sigma_j\}_{j \neq i}$, is easily seen to be:

$$P(\sigma_i = \pm 1 | \{\sigma_j\}) = \frac{1}{1 + \exp\left[\pm 2(h_i + \sum_{j \neq i} J_{ij}\sigma_j)\right]}. \tag{10}$$

While we'll postpone a detailed description of this until later, we note that this conditional probability can be used to simulate a network whose states follow the distribution of Eq. (4) by iteratively selecting a random neuron i, and changing its state (either stochastically or deterministically) to active or inactive, according to which of $P(\sigma_i = +1|\{\sigma_j\})$ or $P(\sigma_i = -1|\{\sigma_j\})$ has the greatest value.

Alternatively, one could update every neuron in each iteration of the simulation. In either case, progressively changing the state of the system in this way can be understood as movement of the system through *state space*, whose points are all the 2^N possible configurations of the system.

This idea allows us to understand the first wave of interest in the application of Ising models in neuroscience, which originated with a model for associative memory proposed by John Hopfield in 1982 [29]. The *Hopfield model* can be thought of as a deterministic version of the Ising model whose movement in state space is simulated as described above [4, 29]. Hopfield's key idea was that any set of states, $\{\vec{\sigma}^\mu\}$ can be made locally stable in state space by choosing the weights according to the so called *Hebb rule* [4]:

$$J_{ij} = \frac{1}{N} \sum_\mu \sigma_i^\mu \sigma_j^\mu. \tag{11}$$

The states that are made stable in this way are now effectively stored in the system: if the simulation is initialized in a state that is sufficiently close to one of these stored states, its state space trajectory will converge to the stored state [29].

This is reminiscent of associative memory, in that the perception of a stimulus which is similar to one that has been experienced in the past can lead to the retrieval of the past experience. The Hopfield model, however, is restricted to the phenomenon of retrieval, and does not include a prescription for how a network can adapt to carry out some desired task. In other words, it does not address the problem of learning.

The Hopfield model was subsequently generalized by Hinton and Sejnowski in 1983 to include stochastic updating by introducing the notion of temperature (reinforcing the connection with the Ising model), as well as so called *hidden units*, inventing what is known as the *Boltzmann machine* [30], which reduces to the Hopfield model in the case of only visible units and zero temperature [4]. They also developed an algorithm by which the Boltzmann machine could be made to learn [30, 31].

While not free of disadvantages, this Boltzmann learning algorithm turns out to be a very accurate way to solve one of our remaining problems. More specifically, although the method of Lagrange multipliers is sufficient to give us the form of distribution to model the neural recordings, *cf.*, Eq. (4), we have yet to address the problem of specifying its parameters, $\{\vec{h}, \mathbf{J}\}$. Boltzmann learning is one of the methods we'll employ for this objective.

The Hopfield model and Boltzmann machine certainly established a connection between the Ising model and cognitive science, and have inspired the development of more sophisticated models for cortical function [3], as well as a range of machine learning algorithms [32, 33]. These developments, however, are primarily concerned with problems which are orthogonal to that of modelling neural recordings.

Interest in the use of Ising models in the context with which we are concerned here can be attributed to a study carried out by Schniedman and his collaborators in 2006, in which they used the principle of maximum entropy to arrive at the Ising model as the most appropriate model of recordings from retinal ganglion cells of the larval tiger salamander and the guinea pig [34].

This leads us to our choice of constraints. Although it should now be clear that one should use the model which maximises the information entropy, we have yet to justify our choice to use the firing rates of each neuron and the covariance between every pair of neuron to constrain the model. Firstly, although Jaynes' method allows for constraints based on agreement with the sample average of a wide range of functions of the activation patters, $\langle f(\vec{\sigma})\rangle$ [22], simpler functions lead to simpler (and more easily interpretable) models. Thus, we will restrict our attention to k^{th} order moments, i.e., sample averages of products over the state of k neurons, such as $\langle\sigma_1 \dots \sigma_k\rangle$, where $k \leq N$.

Denoting the maximum entropy distribution constrained by all moments of order $\leq k$ $(k = 1, 2, ..., N)$ by $P_k(\vec{\sigma})$, one could ask the following question: what is the lowest order k needed to get a model which accurately describes the data? Just as simpler constraint functions make for a simpler model, so does a smaller number of constraints. Therefore, if it turns out that the independent model, $P_1(\vec{\sigma})$, constrained only by the mean firing rates is sufficient, then this would be the most desirable choice.

In [34], although the independent model proved to be insufficient, the Ising model, $P_2(\vec{\sigma})$, showed good agreement with the data. While we won't reproduce their argument here, the key observation was that the the inclusion of second order moments accounted for 90% of the entropy difference between the independent model $P_1(\vec{\sigma})$ and the "full-description" model $P_N(\vec{\sigma})$.

1.5 The kinetic Ising model

The maximum entropy model discussed so far, Eq. (4), is static, and depends only on correlations between the simultaneous activation of the recorded neurons. Writing the recorded activation pattern in time bin t as,

$$\vec{\sigma}(t) = \big[\sigma_1(t), \dots, \sigma_N(t)\big]^T,$$

these 'instantaneous correlations' can be written as

$$C_{ij} = \langle\sigma_i(t)\sigma_j(t)\rangle_t.$$

Neural activity, on the other hand, involves complex temporal dynamics, and any causal interactions that might occur between neurons would not emerge as instantaneous co-activation. Instead, such interactions would lead to a dependence of the network state at time $t+\Delta t$ on the state at an earlier time t. Incorporating this into our model requires looking not only at correlations *within* a network state, but also at correlations *across* states. As mentioned previoiusly, this can be captured by so called *delayed correlations* [28, 35]:

$$D_{ij} := \langle\sigma_i(t + \Delta t)\sigma_j(t)\rangle_t. \tag{12}$$

Recall the conditional probability, Eq. (10), introduced in the preceding section. It can be written as

$$P(\sigma_i|\{\sigma_j\}) = \frac{\exp\left[\sigma_i H_i(\{\sigma_j\})\right]}{2\cosh H_i(\{\sigma_j\})}, \tag{13}$$

where we've introduced the *effective field*,

$$H_i(\{\sigma_j\}) := h_i + \sum_{j\neq i} J_{ij}\sigma_j.$$

Naively, we could get a dynamical model by re-interpreting Eq. (13) as the conditional probability of σ_i at time $t + \Delta t$ given the network state $\{\sigma_j\}$ at an earlier time t:

$$P(\sigma_i|\{\sigma_j\})$$

$$\downarrow$$

$$P(\sigma_i(t + \Delta t)|\vec{\sigma}(t)) := \frac{\exp\left[\sigma_i(t + \Delta t)H_i(t)\right]}{2\cosh H_i(t)}, \tag{14}$$

where

$$H_i(t) := h_i + \sum_{j \neq i} J_{ij}\sigma_j(t). \tag{15}$$

While this is how we use the conditional probability in practice, when sampling from the equilibrium model, given by Eq. (4), it would be desirable to have a stronger justification than this argument by re-interpretation.

While there are numerous ways to derive this dynamical Ising model, a simple, yet more rigorous approach is to apply the principle of maximum entropy to the conditional distribution $P(\vec{\sigma}(t + \Delta t)|\vec{\sigma}(t))$, *i.e.*, the measure over states, $\vec{\sigma}(t + \Delta t)$, given the state at some earlier time, $\vec{\sigma}(t)$. Doing this as before, using the method of Lagrange multipliers, and including the constraint that the model agrees also with the delayed correlations, D_{ij}, as defined by Eq. (12), one obtains the *kinetic* (or *non-equilibrium*) Ising model [28, 36]:

$$P(\vec{\sigma}(t + \Delta t)|\vec{\sigma}(t)) = \frac{\exp\left[\sum_i \sigma_i(t + \Delta t)H_i(t)\right]}{\prod_i 2\cosh H_i(t)}. \tag{16}$$

Moreover, by marginalizing, it is easily seen that this is consistent with Eq. (14):

$$P(\sigma_i(t + \Delta t)|\vec{\sigma}(t)) = \sum_{j \neq i} P(\vec{\sigma}(t + \Delta t)|\vec{\sigma}(t)) = \frac{\exp\left[\sigma_i(t + \Delta t)H_i(t)\right]}{2\cosh H_i(t)}.$$

Next, we highlight some of the choices we make in endowing the Ising model with temporal dynamics. Firstly, this particular embodiment of the kinetic Ising model is often referred to as the *synchronous* kinetic Ising model, because all units are updated simultaneously during simulation. An alternative approach is to update only one unit at a time – like we'll be doing when sampling from the static Ising model, or to let the updating times for each unit be independent random variables, giving rise to a doubly stochastic process [37], which corresponds to what is called the *asynchronous* kinetic Ising model [28, 35].

Many of inference techniques, including mean field methods and Boltzmann learning, mentioned in the previous section, have been adapted for the kinetic Ising model. These take much simpler forms for the synchronous version of the kinetic Ising model than for the asynchronous one [28, 35]. On the other hand, it has been argued that the synchronous model lacks in biological realism, as compared to certain asynchronous realizations of the model, as the brain does not have an internal metronome that decides when neurons can and cannot fire action potentials [28, 35].

While the static Ising model is restricted to symmetric connections, the kinetic Ising model, is not. Additionally, while we will not include this in our models, the kinetic Ising models can also accommodate self connections, $J_{ii} \neq 0$.

Just as the selection of bin widths involves choice, so does the selection of which time delay, Δt, for the kinetic model. For simplicity, we typically set the time delay between states equal to the width of a single bin, *i.e.*, $\Delta t = 1$. Now, it should be noted that if the couplings symmetrical, Eqs. (14) and (13) differ only in notation, and so it should come as no surprise that the stationary distribution of the kinetic Ising model with asynchronous updates is the Boltzmann distribution, given by Eq. (4). This is no longer the case when the couplings are asymmetrical, or when using synchronous updating, even though these systems may still have a steady state.

More details about the methodology for the kinetic Ising models used here are provided in later sections.

1.6 Motivation and structure

We believe that studying the statistical properties of population activity may contribute to a better understanding of the overarching principles that govern the brain or, on a broader level, nervous systems in general. It follows then that finding statistical models which capture the essential features of such population activity might be a very fruitful endeavor, and we have found two candidate models that possess favorable properties: the equilibrium and non-equilibrium Ising models

As discussed, one can think of the constraints that were selected to yield these models as our hypotheses regarding what the *essential features* of the population activity are. Yet, the true test of the models' suitability for modelling neural activity is the extent to which they can be used to make predictions, and

so we will consider the models successful if they prove to be capable of not only matching the constraint observables of the recording data, but also to predict other statistical observables which were not explicitly used as constraints for the models.

This, however, relies on having inferred models which agree with the constraint observables in the first place. Thus, our efforts will be directed not only to investigating the properties of the models themselves, but also towards assessing the performance of the inference methods used to specify the model parameters. Here, we would like to understand the circumstances under which these inference methods can expected to be reliable, as well as those for which they break down. More specifically, we will look at how system size and temperature[4] influence their performance.

In the next section, we introduce these inference algorithms, starting with the Boltzmann learning procedure, followed by the mean-field approximations. We will also introduce algorithms to generate samples from given models, and briefly mention some methods to calculate observables from such samples, or from the neural recordings. We then turn to the investigations described above, where we will start by applying these algorithms to models whose parameters we have selected ourselves, and finally, to the calcium imaging data.

[4]Temperature here does not refer to the physical temperature of the recorded neurons. Instead, it should be concieved of as a mathematical variable that controls the coupling (and external field) strengths.

2 Simulation and inference methods

2.1 Inference algorithms

Thus far, we've presented the form of distributions that will be used, see Eqs. (4) and (16). While this lays the foundation for constructing a model, we still require the parameters $\{\vec{h}, \mathbf{J}\}$ to be fitted to the neural recordings. This is one of the central practical problems we need to solve and so, in the next couple of sections, we introduce the methods that will be employed for this objective.

2.1.1 Exact likelihood maximization

The first kind of inference method we will introduce is the Boltzmann learning algorithm. As explained in Sec. 1.4.2, Boltzmann learning was developed as a learning algorithm for Hinton and Sejnowski's Boltzmann machine, which generally consists of both hidden and visible units [30]. We will use a special case of this algorithm, as the models we are using here are restricted to include visible units only.

In the present context, Boltzmann learning is used to determine the parameters $\{\vec{h}, \mathbf{J}\}$ which maximize the (log-) likelihood function [28]:

$$\begin{aligned}\mathcal{L}_{EQ}(\vec{h}, \mathbf{J}) &:= \frac{1}{M} \ln P(\text{Data}|\{\vec{h}, \mathbf{J}\}) \\ &= \sum_{i<j} J_{ij} \langle \sigma_i \sigma_j \rangle_{\text{Data}} + \sum_i \langle \sigma_i \rangle_{\text{Data}} - \ln Z(\vec{h}, \mathbf{J}),\end{aligned} \tag{17}$$

where M is the number of samples, and $P(\text{Data}|\{\vec{h}, \mathbf{J}\})$ is the probability of seeing the neural data under the Boltzmann distribution, Eq. (4), given a particular set of model parameters $\{\vec{h}, \mathbf{J}\}$.

While, intuitively, it might seem more appropriate to maximize the probability $P(\{\vec{h}, \mathbf{J}\}|\text{Data})$, note that if we assume the model parameters $\{\vec{h}, \mathbf{J}\}$ to be uniformly distributed, Bayes' theorem yields:

$$P(\{\vec{h}, \mathbf{J}\}|\text{Data}) = \frac{P(\text{Data}|\{\vec{h}, \mathbf{J}\}) P(\{\vec{h}, \mathbf{J}\})}{P(\text{Data})}.$$

Thus, since $P(\text{Data})$ is independent of the model, the parameters, $\{\vec{h}, \mathbf{J}\}$, that maximize $P(\{\vec{h}, \mathbf{J}\}|\text{Data})$ will be the same as those that maximize $P(\text{Data}|\{\vec{h}, \mathbf{J}\})$. Moreover, since the logarithm is monotonically increasing, it will also be the same as the parameters that maximize the function of Eq. (17). For more details regarding the advantages of the maximum likelihood estimator, see sections 2.2.1 – 2.2.3 of [28]. For the kinetic Ising model, the (log-) likelihood function is given by [28, 35]:

$$\begin{aligned}\mathcal{L}_{NEQ}(\vec{h}, \mathbf{J}) &= \frac{1}{M} \sum_{t=1}^{M-1} \ln P(\vec{\sigma}(t+1)|\vec{\sigma}(t)) \\ &= \frac{1}{M} \sum_{t=1}^{M-1} \sum_i \Big[\sigma_i(t+1) H_i(t) - \ln \big(2 \cosh H_i(t) \big) \Big]\end{aligned} \tag{18}$$

Now, the procedure we will use to maximize the likelihood is a gradient ascent algorithm which, regardless of whether we're maximizing $\mathcal{L}_{EQ}$ or $\mathcal{L}_{NEQ}$, has the same global structure, given by Alg. (1).

Algorithm 1

procedure MAXIMUM LIKELIHOOD($\vec{h}_0, \mathbf{J}_0, \eta, m, \text{args}$)
  $N \leftarrow \text{length}(\vec{h})$
  fields $\leftarrow [\vec{h}_0]$
  couplings $\leftarrow [\mathbf{J}_0]$
  for $i = 1, \ldots, m$ **do**
    $\vec{h}_{\text{current}} \leftarrow \text{fields}[i-1]$
    $\mathbf{J}_{\text{current}} \leftarrow \text{couplings}[i-1]$
    $\Delta\vec{h}, \Delta\mathbf{J} \leftarrow \text{parameters-change}(\vec{h}_{\text{current}}, \mathbf{J}_{\text{current}}, \text{args})$
    $\vec{h}_{\text{updated}} \leftarrow \vec{h}_{\text{current}} + \eta\Delta\vec{h}$
    $\mathbf{J}_{\text{updated}} \leftarrow \mathbf{J}_{\text{current}} + \eta\Delta\mathbf{J}$
    fields.insert($\vec{h}_{\text{updated}}$)
    couplings.insert($\mathbf{J}_{\text{updated}}$)
  end for
  $\vec{h} \leftarrow \text{fields}[m]$
  $\mathbf{J} \leftarrow \text{couplings}[m]$
  return $\vec{h}, \mathbf{J}$
end procedure

The basic idea is to start from some initial estimate (or guess) for the parameters, and iteratively adjust them by an amount which is proportional to the appropriate gradients of the likelihood function:

$$\Delta h_i^{EQ/NEQ} = \frac{\partial \mathcal{L}_{EQ/NEQ}}{\partial h_i}(\vec{h}, \mathbf{J}),$$
$$\Delta J_{ij}^{EQ/NEQ} = \frac{\partial \mathcal{L}_{EQ/NEQ}}{\partial J_{ij}}(\vec{h}, \mathbf{J}).$$

There are some key differences, however, beyond this general structure. Differentiating the equilibrium likelihood function, Eq. (17), yields:

$$\begin{aligned} \frac{\partial \mathcal{L}_{EQ}}{\partial h_i}(\vec{h}, \mathbf{J}) &= \langle \sigma_i \rangle_{\text{Data}} - \langle \sigma_i \rangle_{\text{Model}}, \\ \frac{\partial \mathcal{L}_{EQ}}{\partial J_{ij}}(\vec{h}, \mathbf{J}) &= \langle \sigma_i \sigma_j \rangle_{\text{Data}} - \langle \sigma_i \sigma_j \rangle_{\text{Model}}. \end{aligned} \tag{19}$$

The model averages, $\langle \sigma_i \rangle_{\text{Model}}$ and $\langle \sigma_i \sigma_j \rangle_{\text{Model}}$ of Eq. (19) can be calculated directly by differentiating the partition function $Z(\vec{h}, \mathbf{J})$, given by Eq. (5). Calculating the partition function however, is infeasible for systems of non-trivial sizes. Instead, the averages are typically approximated using numerical sampling procedures, which will be introduced later.

While sampling is much faster than direct calculation, as a rule of thumb, each run requires one to generate a number of samples on the order of the number M of time-bins in the data [35], not to mention the throwaway samples used to let the model equilibrate. Because sampling can be time-consuming, and since Alg. (1) may require a large number of iterations to converge to the correct parameters, Boltzmann learning can be very slow, especially for larger system sizes.

On the other hand, Boltzmann learning is exact in the sense that, provided that one uses a sufficient number of sampling steps and a sufficient number of learning steps, the algorithm is guaranteed to converge (in probability) to the right parameters [38].

Next, for the non-equilibrium model, differentiating the likelihood, given by Eq. (18), we get:

$$\begin{aligned} \frac{\partial \mathcal{L}_{NEQ}}{\partial h_i}(\vec{h}, \mathbf{J}) &= \langle \sigma_i(t+1) \rangle_{\text{Data}} - \underbrace{\langle \tanh H_i(t) \rangle}_{\text{Data \& Model}}, \\ \frac{\partial \mathcal{L}_{NEQ}}{\partial J_{ij}}(\vec{h}, \mathbf{J}) &= \langle \sigma_i(t+1)\sigma_j(t) \rangle_{\text{Data}} - \underbrace{\langle \tanh H_i(t)\sigma_j(t) \rangle}_{\text{Data \& Model}}. \end{aligned} \tag{20}$$

While we still need to calculate the averages $\langle \tanh H_i(t) \rangle$ and $\langle \tanh H_i(t)\sigma_j(t) \rangle$ *de novo* in each iteration, these can be evaluated in MN^2 computational steps, and so the non-equilibrium likelihood maximization procedure is generally faster than the equilibrium procedure [28].

2.1.2 Mean-field methods

We now turn to some of the approximate methods that have been developed for the Ising model. Among these, the ones that will be used most extensively here are the so called naive mean field (nMF) and Thoules, Anderson, and Palmer (TAP) approximations, which belong to the mean field family of methods. While these do not have the convergence guarantees as does the likelihood maximization algorithms introduced in the previous section, they are generally much faster. Moreover, they can be used to infer an initial estimate for Boltzmann learning.

We first discuss the mean field methods for the static Ising model. As explained in Sec. 1.4.1, the basic idea behind mean field theory is to approximate the solution to a many-body problem by replacing it with a corresponding single-body problem. To make this more precise, note first that the mean firing rates can be written

$$\begin{aligned} m_i = \langle \sigma_i \rangle &= \sum_{\{\sigma_j\}} \sigma_i P(\{\sigma_j\}) = \\ &= \frac{\exp\left(h_i + \sum_{j\neq i} J_{ij}\sigma_j\right) - \exp\left(-h_i - \sum_{j\neq i} J_{ij}\sigma_j\right)}{\exp\left(h_i + \sum_{j\neq i} J_{ij}\sigma_j\right) + \exp\left(-h_i - \sum_{j\neq i} J_{ij}\sigma_j\right)} \\ &= \tanh\Big(h_i + \sum_{j\neq i} J_{ij}\sigma_j\Big). \end{aligned}$$

Now, the intuitive notion of replacing the interactions between every pair of units with a single 'effective interaction' for each unit that was alluded to in Sec. 1.4.1 takes the concrete form of substituting the means m_j for the states σ_j in the equation above. Doing this yields what is known as the self-consistent equations for the magnetizations:

$$m_i = \tanh\Big(h_i + \sum_{j\neq i} J_{ij}\sigma_j\Big) \longrightarrow m_i = \tanh\Big(h_i^{\mathrm{nMF}} + \sum_{j\neq i} J_{ij}^{\mathrm{nMF}} m_j\Big) \tag{21}$$

Inverting Eq. (21) yields the nMF approximation for the fields:

$$h_i^{\mathrm{nMF}} = \operatorname{arctanh}\, m_i - \sum_{j\neq i} J_{ij}^{\mathrm{nMF}} m_j \tag{22}$$

To get the corresponding approximation for the couplings, we note that the connected correlations can be obtained by differentiating the means with respect to the fields:

$$m_j = \sum_{\{\sigma_k\}} \sigma_j P(\{\sigma_k\}) = \frac{1}{Z}\sum_{\{\sigma_k\}} \sigma_j \exp\big[-\mathcal{H}(\{\sigma_k\})\big]$$

$$\Downarrow$$

$$\begin{aligned} \frac{\partial m_j}{\partial h_i} &= \frac{1}{Z}\sum_{\{\sigma_k\}} \sigma_j\sigma_i \exp\big[-\mathcal{H}(\{\sigma_k\})\big] - m_i m_j \\ &= \langle \sigma_i\sigma_j \rangle - m_i m_j = C_{ij} \end{aligned} \tag{23}$$

The nMF approximation for the couplings can now be obtained by applying the inverse function theorem to Eq. (23):

$$J_{ij}^{\mathrm{nMF}} = -\frac{\partial h_i^{\mathrm{nMF}}}{\partial m_j} = -(C^{-1})_{ij} \quad (j \neq i) \tag{24}$$

There exists a more accurate mean field method, however, called the Thoules, Anderson and Palmer (or TAP) approximation. It can be derived by adjusting Eq. (21) using what is known as the *Onsager correction term*, which yields the corrected mean field equaitions:

$$m_i = \tanh\left[h_i^{\mathrm{TAP}} + \sum_{j\neq i} J_{ij}^{\mathrm{TAP}} m_j - \underbrace{m_i \sum_{j\neq i} (J_{ij}^{\mathrm{TAP}})^2 (1-m_j^2)}_{\text{Onsager's correction term}}\right] \tag{25}$$

The basic idea behind Onsager's modification is to remove the contribution of the 'central unit', σ_i to the effective field, as this should not contribute to the field 'felt' by σ_i (itself). The TAP approximation is derived in much the same way as we derived the nMF approximation. Like before, the fields are obtained by inverting the appropriate mean field equations (Eq. (25), in this case):

$$h_i^{\mathrm{TAP}} = \operatorname{arctanh} m_i - \sum_{j\neq i} J_{ij}^{\mathrm{TAP}} + m_i \sum_{j\neq i} (J_{ij}^{TAP})^2 (1-m_j^2) \tag{26}$$

Also as before, the couplings are derived by differentiating the fields with respect to the means:

$$-(C^{-1})_{ij} = \frac{\partial h_i^{\mathrm{TAP}}}{\partial m_j} = J_{ij}^{\mathrm{TAP}} - 2(J^{\mathrm{TAP}})^2 m_i m_j \quad (i \neq j) \tag{27}$$

These quadratic equations, in turn, must be solved for J_{ij}.

The equilibrium mean field methods are especially accurate for networks that are fully connected and whose connections are drawn from a distribution with a variance proportional to $1/N$ [28]. Before turning to the corresponding methods for the kinetic Ising model, we point out that there are multiple ways to derive mean field theory. Another method involves starting from a factorized distribution, deriving its Helmholtz free energy, and carrying out a Legendre transfrom to get the Gibbs free energy. The nMF approximation for the fields and couplings can now be derived by differentiating the Gibbs potential with respect to the conjugate variables of the parameters [28]. As demonstrated by Plefka, this approach can be extended to yield the TAP (and even higher order) approximations by carrying out a Taylor expansion of the Gibbs free energy in the weights around this factorized model [39].

This approach, however, requires the existence of thermodynamic potentials from which the parameters can be derived. Models such as the kinetic Ising model that allow asymmetrical connections, however, do not necessarily satisfy this requirement, as their stationary distribution is unknown [40]. There exists another framework, however, that can accommodate the kinetic Ising model. It is based on an information geometric argument in which the distribution to be approximated is viewed as a point in a manifold of probability distributions. The nMF and TAP approximations can then be understood as a projection onto a sub-manifold of factorized distributions [40]. This framework yields exactly the same mean field equations as for the equilibrium model, Eqs. (21) and (25).

Now, as shown in [36], the nMF approximation for the kinetic Ising model can be derived from Eqs. (21) and (20), and are given by:

$$\vec{h}^{\mathrm{nMF}} = \operatorname{arctanh} \vec{m} - \mathbf{J}^{\mathrm{nMF}} \vec{m}, \tag{28}$$

$$\mathbf{J}^{\mathrm{nMF}} = (\mathbf{A}^{\mathrm{nMF}})^{-1} \mathbf{D} \mathbf{C}^{-1}, \tag{29}$$

where $\mathbf{D}$ is the matrix of delayed correlations, given by Eq. (12). We've also committed a slight abuse of notation, and arctanh( ) should be understood as applied element-wise to the means. Moreover,

$$\mathbf{A}^{\mathrm{nMF}} = \begin{bmatrix} (1-m_1^2) & 0 & \dots & 0 \\ 0 & (1-m_2^2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (1-m_N^2) \end{bmatrix}.$$

Similarly, Eqs. (25) and (20) can be used to derive the TAP approximation [36]:

$$\vec{h}^{\mathrm{TAP}} = \operatorname{arctanh} \vec{m} - \mathbf{J}^{\mathrm{TAP}} \vec{m} + \vec{\Delta}^{\mathrm{Ons}}, \tag{30}$$

$$\mathbf{J}^{\mathrm{TAP}} = (\mathbf{A}^{\mathrm{TAP}})^{-1} \mathbf{D} \mathbf{C}^{-1}, \tag{31}$$

where $\vec{\Delta}^{\mathrm{Ons}}$ is the vector consisting of the Onsager correction terms (*cf.* Eq. (25)):

$$\Delta_i^{\mathrm{Ons}} := m_i \sum_{j \neq i} (J_{ij}^{\mathrm{TAP}})^2 (1 - m_j^2).$$

The matrix $\mathbf{A}^{\mathrm{TAP}}$ is now given by:

$$\mathbf{A}^{\mathrm{TAP}} = \begin{bmatrix} (1-m_1^2)(1-F_1) & 0 & \dots & 0 \\ 0 & (1-m_2^2)(1-F_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (1-m_N^2)(1-F_N) \end{bmatrix},$$

where the F_i are the smallest (real) roots of the equation [36]:

$$F_i(1-F_i)^2 = (1-m_i^2) \sum_j (J_{ij}^{\mathrm{TAP}})^2 (1-m_j^2).$$

It should be noted that the roots, F_i, cannot exceed $1/3$. Thus, the non-equilibrium TAP approximation algorithm is limited to the weak coupling (or high temperature) regime [36].

Regarding how these mean field approximations compare to the exact likelihood maximization algorithms introduced in the previous section: the mean field approximation algorithms involve only elementary matrix operations such as products and inversions, the latter of which requires $\mathcal{O}(N^3)$ operations, where N is the number of units in the network. For the equilibrium model, the number of operations involved in exact computation of the partition function or its derivatives is grows exponentially with system size, though as discussed, we can speed it up to some extent using numerical sampling [28]. The speed up in going from exact likelihood maximization to approximate methods, albeit significant, is less dramatic for the non-equilibrium model, because it is already polynomial in M and N [28].

2.2 Methods for the forward problem

Although our principal aim is to construct models, which is an inverse problem, we will also need to solve a series of forward problems. Indeed, as discussed in Sec. 2.1.1, one of our methods for solving the inverse problem – the Boltzmann learning algorithm for the equilibrium model, relies on sampling and calculating observables from samples. Moreover, we'll also have to sample the models and calculate observables whenever we make a comparison between distributions, or between a distribution and the data it supposedly models.

2.2.1 Sampling and simulation procedures

As mentioned in Sec. 1.4.2, we make use of two sampling procedures. The former to sample the equilibrium model, and the latter to sample (or simulate) the non-equilibrium model.

Just like for the Boltzmann learning procedure, the two sampling procedures have a shared overall structure: we first generate a certain number of samples to be discarded. The purpose of this is to get closer to equilibrium and decrease the dependency on the initial state. Then, starting from the last state generated in this first run, we generate the samples that are kept. In each iteration the previously generated network state is updated according to a procedure which we've referred to as update-state( ), whose contents depend on whether we're sampling from the equilibrium or non-equilibrium distribution, and the new network state is stored. *cf.*, Alg. (2).

Algorithm 2

procedure GENERATE SAMPLES($\vec{h}, \mathbf{J}, M, M_{\text{burn}}$)
  $N \leftarrow \text{length}(\vec{h})$ ▷ there is exactly one field per unit
  $\vec{s}_0 \leftarrow \text{randomFromSet}(\{1, -1\}, N)$ ▷ random vector of N elements ± 1
  ▷ we first generate M_{burn} samples, each of which will be discarded
  (except for the last one), so as to ensure we've reached equilibrium
  $\mathbf{S}_{\text{burn}} \leftarrow [\vec{s}_0]$ ▷ container for burner samples
  for $i = 1, ..., M_{\text{burn}}$ **do**
    $\vec{s}_{current} \leftarrow \mathbf{S}_{\text{burn}}[i-1]$
    ▷ the update-state( ) procedure depends on whether we're sampling
    the equilibrium model, or simulating the non-equilibrium model
    $\vec{s}_{\text{updated}} \leftarrow \text{update-state}(\vec{s}_{\text{current}}, \vec{h}, \mathbf{J})$
    $\mathbf{S}_{\text{burn}}.\text{insert}(\vec{s}_{\text{updated}})$
  end for
  ▷ the procedure is now repeated, starting from the last burner sample
  $\mathbf{S} \leftarrow [\mathbf{S}_{\text{burn}}[M_{\text{burn}}]]$
  for $i = 1, ..., M$ **do**
    $\vec{s}_{\text{current}} \leftarrow \mathbf{S}[i-1]$
    $\vec{s}_{\text{updated}} \leftarrow \text{update-state}(\vec{s}_{\text{current}}, \vec{h}, \mathbf{J})$
    $\mathbf{S}.\text{insert}(\vec{s}_{\text{updated}})$
  end for
  return S ▷ output the M samples generated
end procedure

Thus, the difference between the two sampling procedures lies in how we move from one state to the next. For the equilibrium model, update-state( ) selects a random index $k \in \{1, \ldots, N\}$. The state of the corresponding unit, σ_k, is 'flipped' if either

$$\Delta E_k := 2\sigma_k \Big(\sum_{i \neq k} J_{ik}\sigma_i + h_k \Big) < 0,$$

or, if the transition probability,

$$P_{\text{flip}}(k, \vec{\sigma}_{\text{current}}) = e^{-\Delta E_k}, \tag{32}$$

is greater than or equal to a random number, u, sampled from the uniform$(0,1)$ distribution[5].

For the non-equilibrium model, we use parallel updating, which entails updating the whole population at every step, rather than only selecting a single unit for updating, as we do in the equilibrium sampling procedure. As before, each unit is updated by comparing transition probabilities with a sample from the uniform$(0,1)$ distribution. In this case, we use the probability of a given unit being active,

$$P_+(\sigma_i(t)) = \frac{1}{1+\exp\left[-2H_i(t)\right]}. \tag{33}$$

2.2.2 Calculating observables

In the following we denote a sample consisting of the states of neurons $i = 1, ..., N$ in the time bins $t = 1, ..., M$, by

$$\mathbf{S} := \left[s_i(t)\right]_{i,t} = \begin{bmatrix} s_1(1) & \dots & s_N(1) \\ \vdots & \ddots & \vdots \\ s_1(M) & \dots & s_N(M) \end{bmatrix}$$

We first consider the constraining observables. For the equilibrium model these are the mean firing rates,

$$\vec{m} = \left[\langle s_i \rangle\right]_i = \langle \mathbf{S} \rangle_t,$$

and the pair-wise covariances,

$$\chi = \left[\langle s_i s_j \rangle\right]_{i,j} = \frac{1}{M}\mathbf{S}^T\mathbf{S}.$$

Alternatively, we may use the so-called connected correlations

$$\mathbf{C} = \left[\langle (s_i - m_i)(s_j - m_j) \rangle\right]_{i,j} = \frac{1}{M}\mathbf{S}^T\mathbf{S} - \vec{m}\vec{m}^T.$$

The non-equilibrium model is also constrained by the so-called delayed correlations:

$$\mathbf{D} = \left[\langle s_i(t+1)s_i(t) \rangle_t\right]_{i,j} = \frac{1}{M-1}\mathbf{S}_{tail}^T\mathbf{S}_{head} - \vec{m}_{tail}\vec{m}_{head}^T,$$

where $\mathbf{S}_{tail}$ and $\mathbf{S}_{head}$ are defined as $\mathbf{S}$ with the first and last rows removed, respectively.

Finally, we will also make use of certain non-constraining observables. In particular, we will approximate the distribution over the number of co-activated units using the relative frequency of k units being active, for $k = 0, ..., N$. Additionally, we will calculate 3^{rd} order correlations, which is given by,

$$\mathbf{C}^{(3)} := \left[\langle s_i s_j s_k \rangle\right]_{i,j,k},$$

and can be calculated according to Alg. (3).

[5]Note that since $u < 1$, the state will necessarily be changed if $\Delta E_k \leq 0$, while it may or may not be changed otherwise. This is why we deal with this case seperately, so as to avoid the subsequent computations.

Algorithm 3

procedure CALCULATE THIRD-ORDER COVARIANCES($\mathbf{S}$)
  $N \leftarrow \text{ncols}(\mathbf{S})$ ▷ ncols( ) yields the number of columns of a matrix
  $\mathbf{C}^{(3)} \leftarrow$ empty array of shape (N, N, N)
  $\mathbf{S}_{\text{extd}} \leftarrow \text{extend}(\mathbf{S}, N)$ ▷ Make N copies along a new axis
  $\mathbf{S}^T_{\text{extd}} \leftarrow \text{transpose}(\mathbf{S}_{\text{extd}}, (1, 3, 2))$
  $\mathbf{P}^2 \leftarrow \text{Hadamard}(\mathbf{S}_{\text{extd}}, \mathbf{S}^T_{\text{extd}})$
  for $k = 1, \ldots, N$ **do**
    $\mathbf{S}^k_{\text{extd}} \leftarrow \text{extend}(\text{extend}(\mathbf{S}, N), N)$
    $\mathbf{P}^3 \leftarrow \text{Hadamard}(\mathbf{P}^2, \mathbf{S}^k_{\text{extd}})$
    $\mathbf{C}^{(3)}_k \leftarrow \langle \mathbf{P}^3 \rangle_t$
  end for
  return
end procedure

3 Testing the methods

To test the methods discussed in the previous section, we start by applying them to the models they are supposed to fit. Since the behavior of the methods have already been extensively investigated in this context by others, we can use the results gained here to confirm the correctness of their implementation.

Our approach involves selecting a particular configuration for the fields, h_i, and the couplings J_{ij}, and generateing samples using the appropriate version of Alg. (2) for the static and kinetic Ising models. By assumption, the couplings of the equilibrium model must be symmetrical $J_{ij} = J_{ji}$, and the diagonal elements J_{ii} must all be zero. While we will keep the latter condition also for the kinetic Ising model (even though this is not strictly necessary), we relax the symmetry condition.

Once we have generated the samples, we can use the inference methods introduced in Secs.2.1.1 and 2.1.2 to infer their parameters, which can now be directly compared with the true parameters of the models.

3.1 Testing the Monte-Carlo sampler

The steps in the approach taken here all depend on Monte-Carlo simulation in one way or another. More specifically,

- whenever we compare the statistics of the neural data to a fitted model, we need to sample from the model[6],
- in the case of the equilibrium model, we sample the model at every iteration of the Boltzmann learning procedure.

In light of this, it is crucial that we assess the accuracy and correct implementation of the Monte-Carlo sampler, and that we understand the factors contributing to errors in the statistics of the samples. Due to the similarity of the sampling algorithms for the two models, we restrict our attention the equilibrium model here. In certain cases, analytic expressions for the first and second moments of the equilibrium model can be derived. We have done this for the case of *independent-pair* (or *gas-of-dimer*) couplings[7]. In this case, we can compare the moments of the samples to the analytic expressions, and thereby assess the accuracy of the MCMC sampler, as well as its sources of error.

Our results demonstrate that the MCMC sampler yields accurate estimates of the moments of the model, and so is a suitable alternative to calculating moments directly. Nevertheless, we observe that the error in the estimated moments is influenced by the number of units in the network, the sample size, and the coupling strengths, see Fig. 6 (and accompanying table 1).

[6] When fitting models to data generated from another, known model, the simulation step occurs at least twice – once for the known model, and once for the fitted model.

[7] See Sec.A.1 for a derivation.

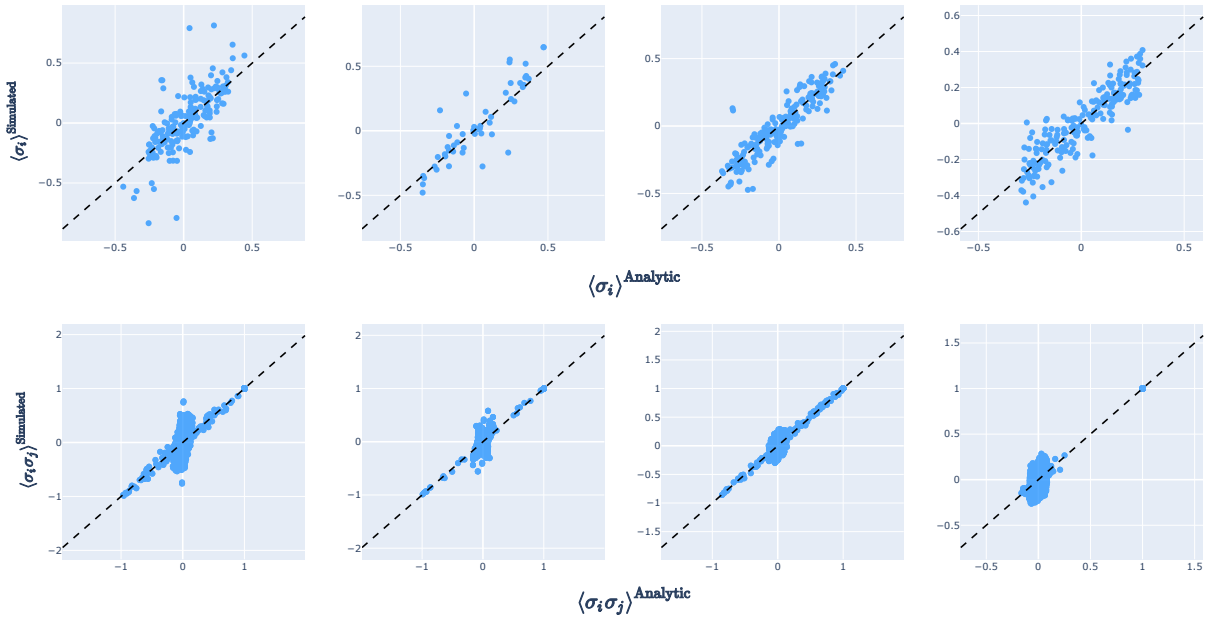


Figure 6: $M_{\text{burn}} = 1000$. Original: $N = 200$, $\beta = 10$, $M = 30,000$; Less units: $N = 50$; More samples: $M = 100,000$; Lower β: $\beta = 1$

	RMSE(m_i)	RMSE(χ_{ij})	$\frac{\text{RMSE}(m_i)+\text{RMSE}(\chi_{ij})}{2}$
Original	0.162	0.096	0.129
Less units	0.140	0.090	0.115
More samples	0.093	0.056	0.074
Lower β	0.082	0.063	0.073

Table 1: RMSE values corresponding to the subplots in Fig. 6

A more thorough sensitivity analysis reveals that the relationships of the RMSEs with system size, N, and sample size, M, both appear to follow power laws with exponets of roughly $1/2$ and $-1/2$, respectively. Additionally, the N and β sensitivities both exhibit heteroscedasticity, with a volatility that decays as the system size is increased, and grows as β is increased. Regarding the β sensitivity, in particular, we see that the growth in the maximum values are larger than that of the min values, showing an overall tendency towards higher errors as the couplings get stronger.

As for the interpretation of these observations, the decay in the errors with increasing sample size is merely an instance of the well known *law of large numbers*: the more counts we have of each state, the more accurately our averages will reflect the expectations of the underlying process. The rate of convergence to the true expectations, of course, depend on the size of the state space: the more possible outcomes we have to deal with, the more observations are needed to get accurate statistics.

This brings us to the observation that the estimation errors grow as the system size is increased (whilst the sample size is held fixed): this happens because the size of the state space has a cardinality of 2^N, and so grows exponentially in size with the addition of more units. Next, the decay in the *variablitiy* of the RMSEs with increasing system size is also an instance of the law of large numbers: The RMSE itself can be viewed as realizations of a random variable whose standard deviation depends on the number of data points used in the estimator, which are N and $N(N-1)/2$ for the magnetizations and pair-wise covariances, respectively. This also explains why the volatility in the RMSEs of the χ_{ij}s decay more quickly than those of the m_is.

Finally, turning to the sensitivity of the RMSEs to the inverse-temperature, β, which controls the coupling strengths $J_{ij} \sim \mathcal{N}(0, \beta/\sqrt{N})$. The SK model has a "normal" phase, and a *spin-glass* phase that can occur at higher values of β (i.e., lower temperature, in the physical context). This spin-glass phase is characterized

by a state-space trajectory that gets stuck in particular configurations, which would result in the sampler not exploring the state-space efficiently, thereby giving us an effectively smaller number of samples, thereby resulting in lower-accuracy estimates [41].

Indeed, looking at samples generated from the IP[8] SK model with different values of β (see Fig.8) reveals that, whilst everything looks normal at $\beta \leq 1$, the state space is completely frozen at $\beta = 100$, and mostly static (albeit with a handfull of spin flips) at $\beta = 10$, illustrating the spin-glass behavior exhibited by the model in the low temperature regime.

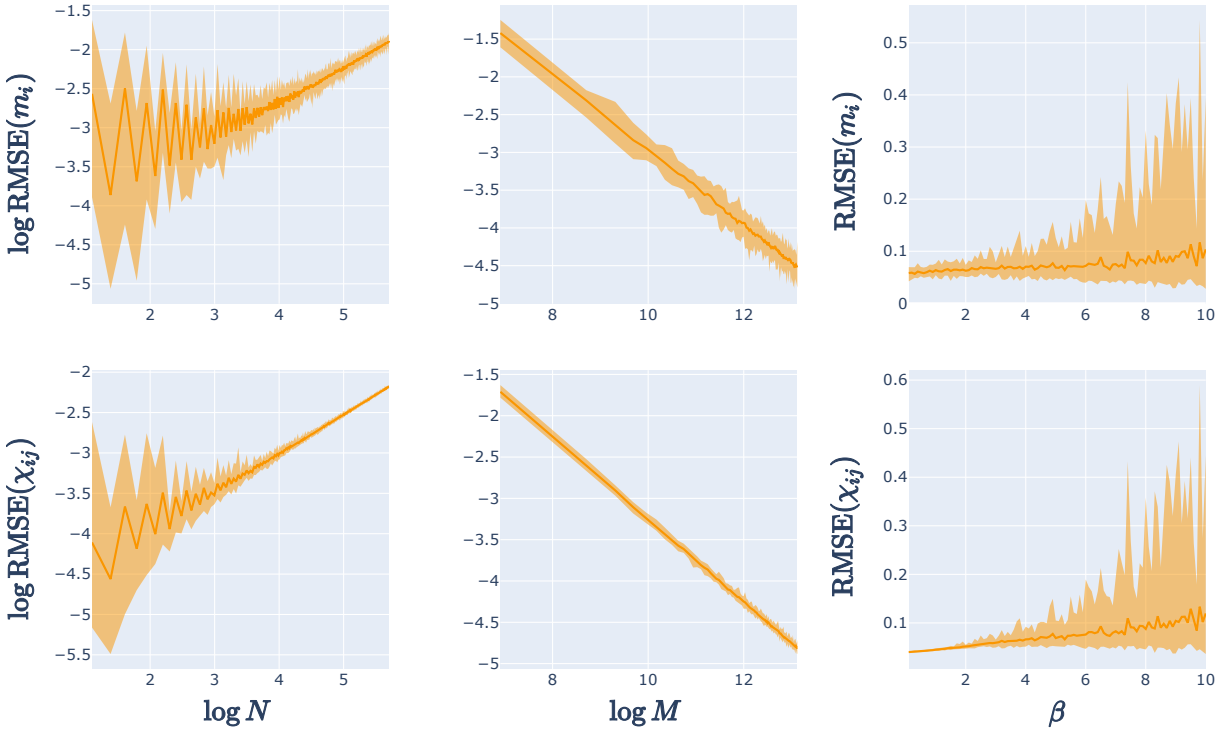


Figure 7: Average RMSE values over 30 runs. For column 1, the N ranges from $3 - 300$, and for columns 2 and 3, $N = 50$; For columns 1 and 3, $\beta = 1.3$, and for column 2, β ranges from $0.1 - 10$; For columns 1 and 2, $M = 15,000$, whereas for column 3, M ranges from $1000 - 500,000$

[8]Independent-pair

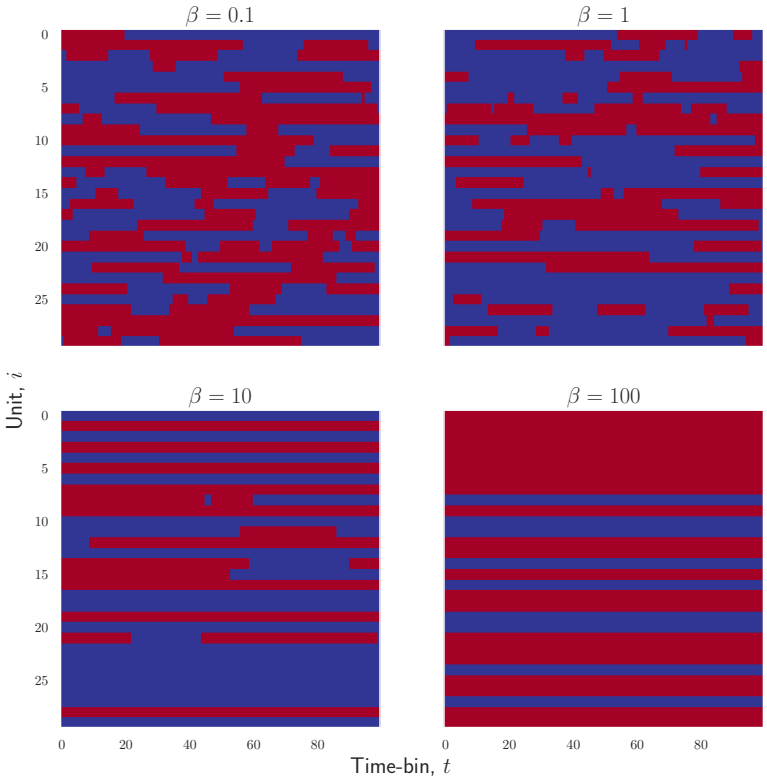


Figure 8: Frozen states at large β. Examples from simulations of 30 units differing in the inverse temperature β. With large couplings, the system exhibits critical slowing down, characteristic of the spin-glass phase.

3.2 Inference on the models themselves

3.2.1 The equilibrium model

The first- and second-order moments of the spin data are sufficient statistics for specifiying the (equilibrium) Ising model. Achieving a close correspondence between these, therefore, results in a closer correspondence between the estimated and true parameters[9]. This is exemplified in Fig. 9a, wherein we see a close connection between the accuracy of parameter estimates and observable predictions.

Regarding the relative performance of the different inverse methods used, our findings are similar to those of previous studies [28, 42]: the quality of the inference using these methods generally follows the ordering, from best to worst, of ML, TAP, and nMF.

Indeed, we also observe the same tendency of the nMF magnetizations to be reflected with respect to the 45° line as seen in Fig. 4 of [28]. To understand why this happens in the case of nMF, we recall that $\mathbf{J}^{\mathrm{nMF}}$ depends entirely on the inverse of the covariance matrix, which can become ill-conditioned, and therefore very sensitive to numerical errors under inversion, whenever the couplings of the underlying system are sufficiently strong, especially in the case of a fully-connected system, as used here. Sign errors in the estimated couplings, in turn, can lead to sign errors in the effective fields which, finally, can result in sign errors in the magnetizations which, as the reader will recall, are give by the hyperbolic tangent of the effective fields.

The fact that ML achieves the best performance among these four methods is unsurprising, as there exist guarantees that it asymptotically converges to the global maximum of the likelihood, provided that the underlying model is an Ising model [28].

	RMSE(h_i)	RMSE(J_{ij})	RMSE(m_i)	RMSE(χ_{ij})
nMF	0.795	0.084	1.138	0.523
TAP	0.499	0.081	0.267	0.206
ML	0.298	0.075	0.078	0.070

Table 2: RMSE values corresponding to the scatters in Fig. 9a. Interestingly, although the inference results follow the usual order of nMF < TAP < ML, in terms of goodness-of-fit.

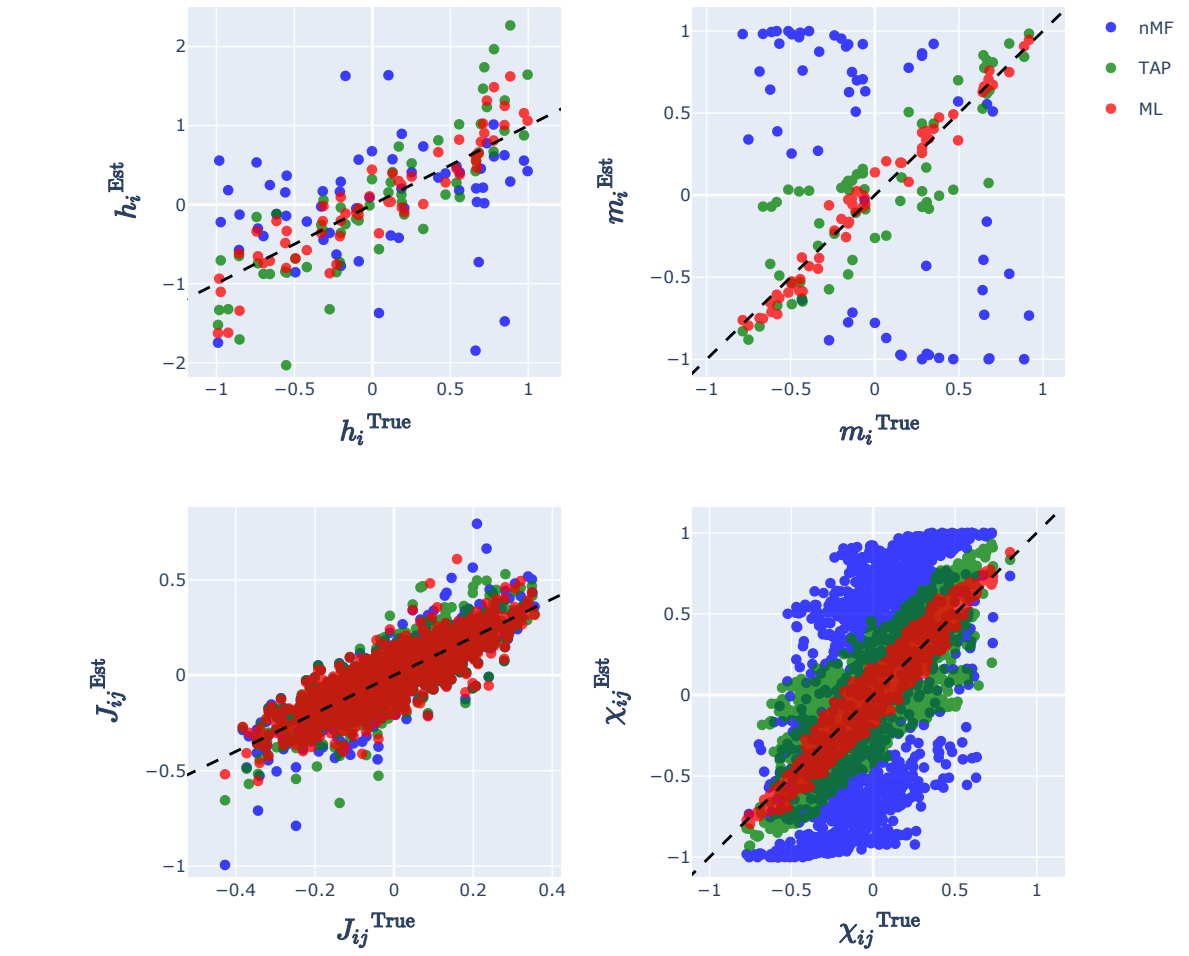
We also investigated the sensitivity of the parameters to β and N in the case that the couplings have a graph configuration with long loops[10]. Similar to what was found in [28], the reconstruction error as function of β exhibits a U-shape, with the ML procedure[11] the recondstruction error of the mean-field methods blows up earlier than that of ML, demonstrating the robustness of the latter in the lower-temperature/strong-couplings regime. See Fig. 9b.

As seen in Fig. 9c, the reconstruction error is also seen to grow linearly with the number of units (c.f., [42]), and, again, the ML method is seen to be much less volatile overall and exhibits a significantly lower reconstruction error at a given system size.

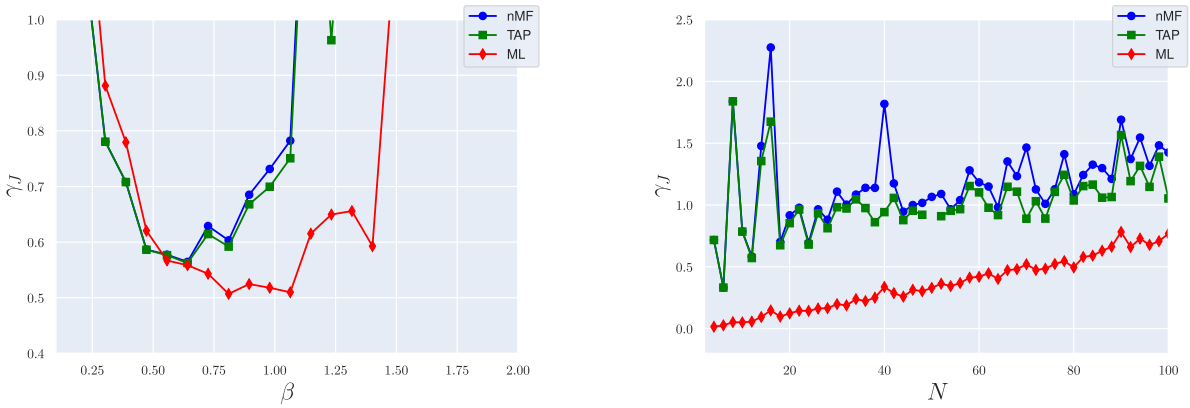
[9] What is meant by *true* depends on the context: here, we have a known Ising model that the spin data is sampled from, and so there really are true parameters. When analyzing cortical recordings, one assumes that the underlying biological neural network can be represented by an Ising model whose parameters are not known, but nevertheless exist.

[10] As in [28], we used a random graph with fixed degree $z = 3$ for this purpose.

[11] [28] actually used Pseudolikelihood maximization instead of ML, but hypothesizes that one would obtain similar results with ML, which our findings confirm.



(a)



(b) (c)

Figure 9: Reconstruction of the equilibrium Ising model with the naive mean-field (nMF), Thoules, Anderson and Palmer (TAP), and maximum likelihood (ML) methods. The scatters in (a) depict comparisons of the true parameters with those inferred using the three inverse methods, as well as comparisons of the observables resulting from another round of MCMC simulation. Here, the true model consited of $N = 60$ units, with fields sampled from a uniform distribution over the interval $[-\beta, \beta]$, and couplings from the $\mathcal{N}(0, \beta/\sqrt{N})$, with $\beta = 1$. In (b) and (c), we show the sensitivity of the reconstruction error γ_J to the inverse temperature, β, and the system size, N, respectively.

3.2.2 The non-equilibrium model

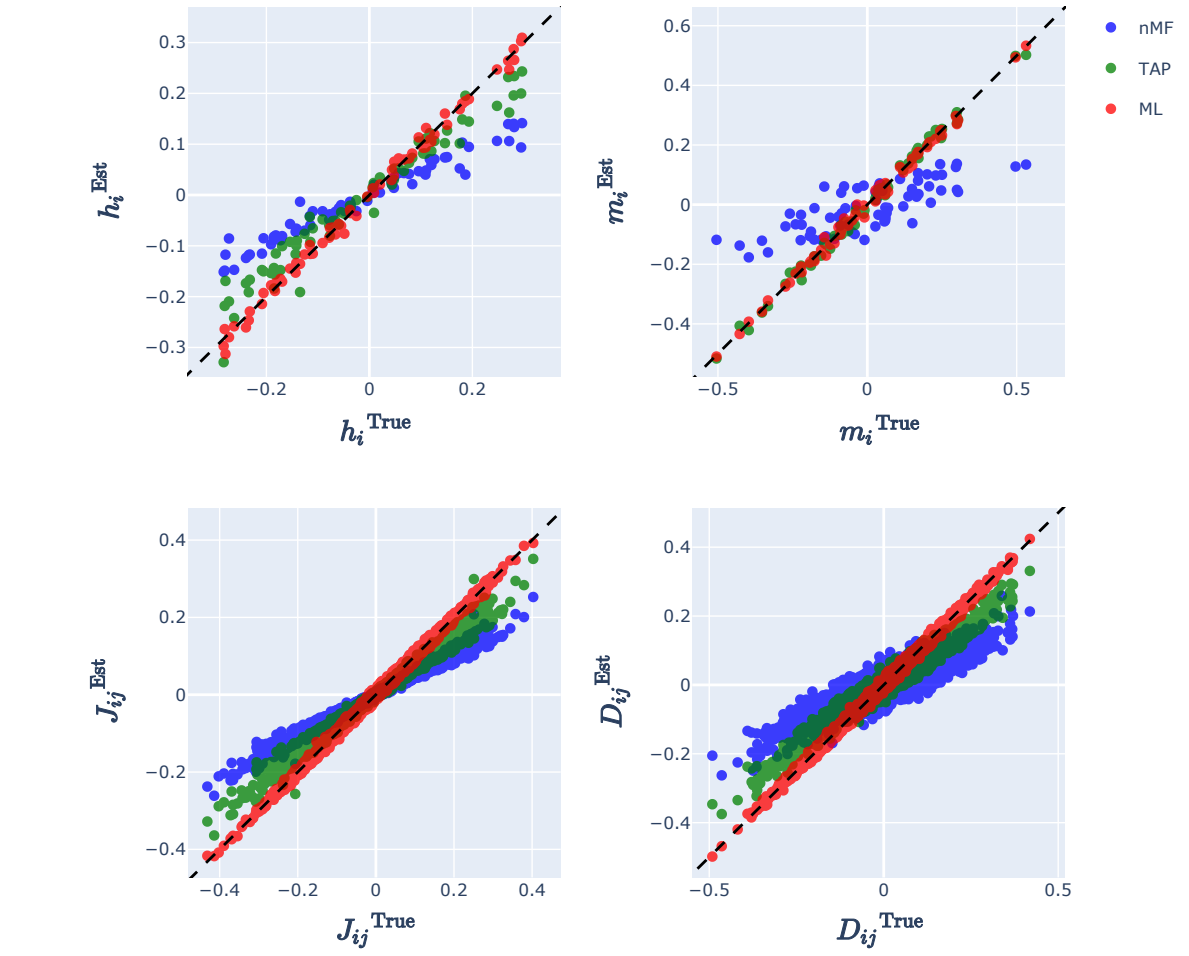
For the NEQ model, although we are still interested in the magnetizations[12], the likelihood function is now maximized by minimizing the discrepancy between the average time-delayed covariances and the corresponding model-predicted delayed covariances, rather than the covariances in the instantaneous activation of the units. Consequently, just as for the EQ model, small deviations in these observables corresponds to small deviations in the parameters. See Fig. 10a and Table 3.

	RMSE(h_i)	RMSE(J_{ij})	RMSE(m_i)	RMSE($D_{ij}(1)$)
nMF	0.090	0.056	0.150	0.081
TAP	0.044	0.032	0.015	0.045
ML	0.011	0.007	0.015	0.008

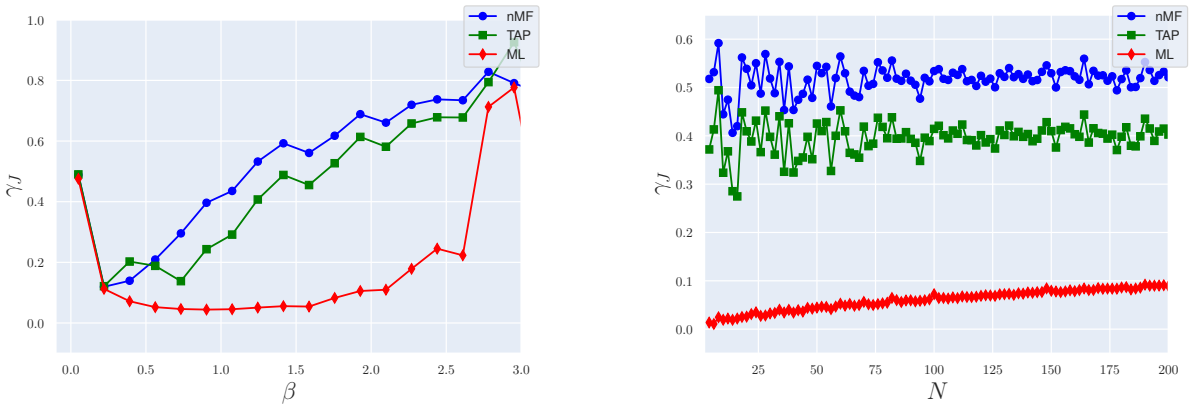
Table 3: RMSE values corresponding to the scatters in Fig. 10a

The relative performance of the three inference methods for the NEQ model follow the same ordering as the corresponding EQ methods, which is consistent with our expectations on the basis of theory and previous work [28, 38]. That said, the NEQ inference appears to be more successful than the EQ inference in multiple ways. Given a similar system size and temperature, the NEQ inference methods acheive smaller deviations across both model parameters and observables used to fit the model. We also see that the NEQ methods are more robust to increases in system size and inverse temperature. See Figs. 10b and 10c.

[12]This is a slight abuse of terminology considering the time-dependence in the NEQ model. The quantities of interest are really the activation in a time bin t, conditional on the state of the network in the previous time-bin $t-1$. See A.2 for a derivation of the model expectations.



(a)



(b) (c)

Figure 10: Reconstruction of the non-equilibrium Ising model using the corresponding versions of the naive mean-field (nMF), Thoules, Anderson and Palmer (TAP), and maximum likelihood (ML) methods. As in the equilibrium case, the scatters in the upper panels, (a), depict comparisons of the true parameters with those inferred using the three inverse methods, as well as comparisons of the observables resulting from another round of MCMC simulation. Here, the system size was $N = 70$, with fields sampled from a uniform distribution over the interval $[-0.3\beta, 0.3\beta]$, and couplings from the $\mathcal{N}(0, \beta/\sqrt{N})$, with $\beta = 1$. In (b) and (c) we look at the reconstruction error as a function of inverse temperature β and system size N.

There are multiple reasons as to why one obtains better parameter estimates for the NEQ model than for EQ model. A straight-forward one is that, in the equilibrium case, in order to side-step the intractability of the partition function, we estimate the $\langle \sigma_i \rangle^{\text{Model}}, \langle \sigma_i \sigma_j \rangle^{\text{Model}}$, using MCMC samples. Thus, Boltzmann learning involves an additional simulation error contribution that is not present in the NEQ maximum likelihood procedure.

On the topic of sampling, an advantage of Glauber dynamics with parallel updates is that one explores the configuration space at a far more rapid pace than in the EQ case, where we flip at most one spin at a time. The EQ sampling procedure therefore generates states with a high autocorrelation as compared to those resulting from the NEQ sampler. The result is that given the same number of simulation steps, the *effective* sample size resulting from a run of the NEQ sampler is greater than that resulting from the EQ sampler. Parallel dynamics may also prevent the occurance of a glassy phase and the accompanying critical slowing-down, as exhibited by the EQ model at large β.

Thus, when inferring the parameters of the NEQ model, the only contribution of simulation error to the total reconstruction error is from sampling the known model[13] and even this contribution is expected to be smaller than the corresponding simulation error contribution for the EQ model.

Another important consideration is that, in contrast with the EQ model, which treats subsequent states as independent snapshots of the state space, the NEQ gradient ascent procedure uses temporal information in adjusting the parameters. This can be seen by noting that both the data and model averages in Eq. 19 are just sums over states of a given unit or the product of the simultaneous states of different units, both of which are invariant to permutations of the time bins. The NEQ gradients on the other hand, involves the states of units in subsequent time-bins, and so are not invariant to perumations of the time bins.

Indeed, the NEQ gradients can be interpreted in terms of the average prediction error in predicting the state of a unit i in bin $t+1$ on the basis of the state of the network, $\sigma(\vec{t})$, in the previous time bin (*c.f.* Eq. 20).

[13] Strictly speaking, when comparing the observables of the model to those of the original model, we also carry out an additional simulation run. However, only one run of sampling is involved in reconstruction of the parameters for the NEQ model.

4 Application to the cortical recordings

Having investigated some of the properties of the models and the inference methods introduced in Sec. 2 in the ideal case, and confirmed that they behave in accordance with what we know from the literature, we now turn to their application to the neural data.

As mentioned in the introduction, the data set that forms the basis for what follows consists of both perfomance and observation sessions, and recordings were taken from are M2 in one mouse, and PPC in another. In the interest of limiting our scope, we chose to focus on one of the eight recording sessions. We consider the performing sessions to be of greatest interest for our purposes because they are more likely to reflect neural activity under normal circumstances than ones in which the mice are head confined. Another consideration is the average firing rate of the recorded population. Theoretical work on the pairwise binary model indicates that higher average firing rates[14] may be associated with neural recordings of greater value, in the sense that smaller sub-populations can be used to make predictions about the statistics of larger populations [43]. We therefore selected the performing session associated with the greatest average firing rate (0.572 spikes/sec), which turned out to be one of the M2 recordings.

Obviously, the true parameters of the underlying models are not known in this case, however, we can still compare the statistics of the recordings with the corresponding observables calculated on the basis of simulations generated from the inferred models.

When analyzing the neural data, we have to make a series of decisions regarding the hyperparameters of the models and the associated inference methods. Some, such as the learning rate η, and the number of simulations to use are not of scientific interest in and of themselves, and were determined through trial and error. Others, such as the number of neurons, as well as quantities relating to the temporal dynamics of the recorded activity, including bin widths and the time-lag used in the non-equilibirum model merit further investigation.

To facilitate these choices, a key consideration is how the neural data itself is influenced by these variables, which is what we investigate subsequently. Finally, we turn to the results from applying the inference methods to the neural data.

4.1 Temporal dynamics

There are two knobs we can turn when studying the temporal dynamics of the neural data: the bin-width Δt and the time-lag δt. Of course, we are limited by the scanning speed determined by the experimental set-up, which for our data is 20 Hz. This results in bin durations of $\Delta t = 50$ ms. As mentioned in the introduction, action potentials can last for 1-10 ms, and so a limitation[15] is that certain spiking dynamics may not be well captured in our data.

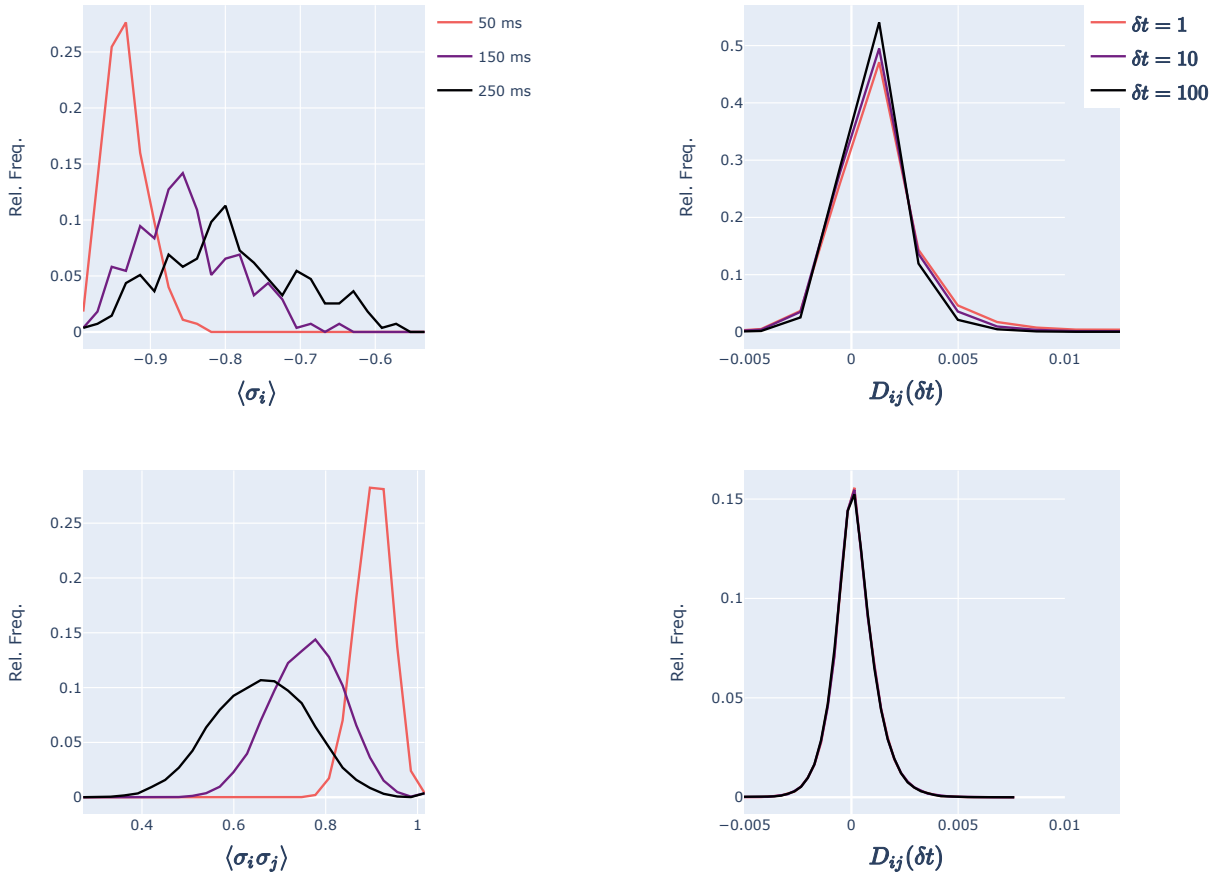
While there is nothing we can do, on the analysis side to increase the temporal resolution, something we can do is to pool the activity of a neuron across bins, counting the activity its activity in an extended bin as active if the neuron was active in any of the original bins contained therein. This may impart certain advantages, such as reducing noise and highlighting sustained activity. On the other hand, besides the loss in temporal precision, there is the risk of falsely increasing the correlations between certain neurons: as we extend the bin size, neurons that do not have any synaptic or other causal interactions, yet happen to have overlapping activity within the longer time intervals will be correlated. This can give a false impression of the synchrony.

For simplicity, we choose to use a time-lag of $\delta t = 1$ for the kinetic Ising model. This also seems a reasonable choice given the temporal resolution of our data relative to the time scale of neuronal spike trains. Nevertheless, we are interested in how the resulting delayed correlations are influenced by this parameter, and consider the effect of increasing it.

To explore the temporal dynamics in the M2 recordings, we considered how the distributions of the observables were influenced by the hyperparameters discussed above. The results of this analysis are depicted in Fig. 11.

[14] Given the same bin width and number of neurons

[15] This is really a limitation of calcium imaging itself, much higher temporal resolution would require a different approach, such as direct electrophysiological recordings.



(a) (b)

Figure 11: Curves depicting the relative frequencies of the means $\langle \sigma_i \rangle$, covariances $\langle \sigma_i \rangle$, and delayed covariaces $D_{ij}(\delta t)$, in different circumstances. In (a) we consider how increasing the bin-widths from the original 50 ms influences the means and covariances, whereas (b) depicts how the distribution of the delayed covariances changes as we increase the *time-lag*, δt, from 1 to 10, and to 100 time bins in the original neural time series (top), as well as in the shuffled time series (bottom), where we have permuted the time-bins.

We start by considering the bin-widths. See Table 4 for hypothesis test results comparing samples differing in this regard.

In Fig 11a we see that when we increase the bin widths, the means exhibit less positive skewness and appear to go from a positive kurtosis (leptokurtic) towards a more moderate, or negative one (seemingly mesokurtic). To understand why this happens, suppose for the sake of simplicity that the probability that a neuron fires within a bin is constant accross bins. Then as the bins are widened, naturally the probability of firing within a given bin would increase proportionately.

The covariances, instead, show a less negative skewness as the bin widths increase. Like the means, when looking at the kurtosis of the covariances, they appear to go from being leptokurtic towards having a mesokurtic distribution. The fact that the covariances seems to decrease as we widen the bins indicates that the neuronal activity co-varies at small time-scales.

	50 & 150 ms	**50 & 250 ms**	**150 & 250 ms**
Stats.	4.05×10^{-2}	7.09×10^{-2}	3.04×10^{-2}
p-values	0%	0%	0%

Table 4: Test statistics and p-values for 2-sample K-S tests used to compare the samples, *i.e.*, neural time-series, differing in that the time-bins of the original sample have been widened from 50 ms to 150 and 250 ms. In each case, the p-values where lower than the numerical precision on the authors machine.

Next, we consider the influence of time-lag on the distribution of the delayed covariances. See Fig. 11b. Here, we have two samples: the original recordings, and ones for which we've randomly permuted (or shuffled) the time-bins, thereby breaking up any temporal structure that may be exhibited by the neural activity. In the shuffled data, the delayed covariances are sharply peaked around zero, and increasing the time-lag does not significantly alter the distribution, *c.f.* Table 5[16]. The delayed covariances calculated from the original data, on the other hand, exhibit a peak close to, albeit not exactly at zero. These also become more sharply peaked and shorter tailed as we increase the time-lag, suggesting that there are temporal associations between some of the recorded neurons, that are mostly short ranged (*i.e.,* on the order of 1 bin-width).

	Sample	1 & 10	1 & 100	10 & 100
Stats.	**Original**	4.08×10^{-2}	8.13×10^{-2}	4.20×10^{-2}
	Shuffled	4.40×10^{-3}	6.88×10^{-3}	7.50×10^{-3}
P-values	**Original**	5.27^{-55}	1.10^{-217}	2.59^{-58}
	Shuffled	4.55×10^{-1}	5.57×10^{-2}	2.84×10^{-2}

Table 5: Table showing test statistics and p-values from 2-sample K-S tests used to compare the distributions of the delayed covariances $D_{ij}(\delta t)$, with different time-lags δt, for both the original neural time series, and those in which the time-bins have been shuffled randomly. Although none of the p-values exceed the commonly used significance level of 5%, there is a clear difference between the shuffled and non-shuffled data, with p-values being many orders of magnitude smaller for the original time series. This indicates that there are temporal dynamics in the original data that are lost when the recorded activity is treated as *i.i.d.* samples.

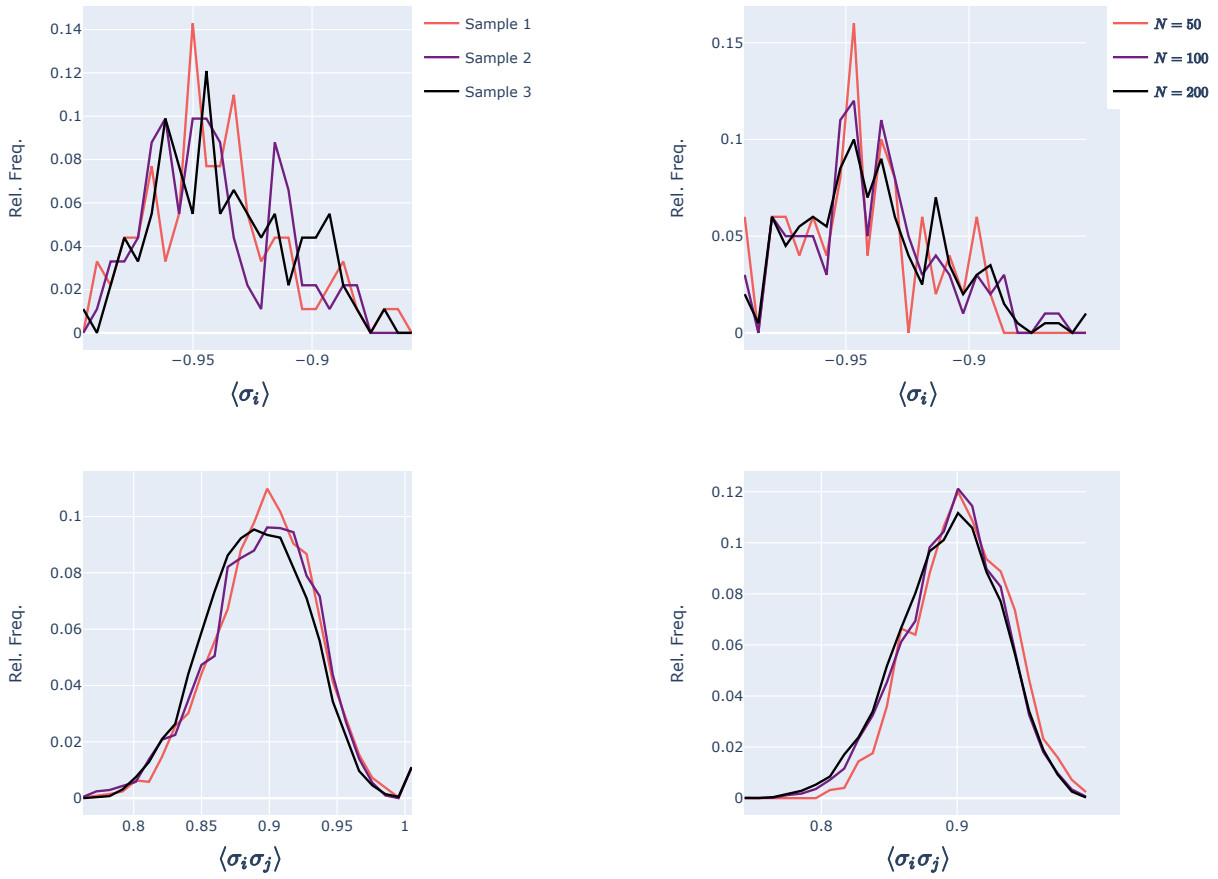
4.2 Network size

In the previous section, one of the things we looked at was how the inference methods perform as the system size, N, grows larger. There are two broad reasons as to why we are interested in system, or network size. One is that both the inference and simulation methods are sensitive to this hyperparameter, and so it is of interest purely from a methodological point of view. The other reason, as disussed in the introduction, is

[16] 2-sample K-S tests assume *i.i.d.* data points within each sample. It is therefore not interesting to directly compare the original and shuffled recording data itself. This is why we compare the distributions of the delayed covariances instead.

that neurons cooperate in large populations to process information. Thus, to get closer to a complete picture of how this information processing works, we need to include a sufficient number of neurons in our models.

We now extend this line of inquiry, and consider how the number of neurons, as well as which specific neurons are used influence the observables of the resulting neural data. See Fig. 12.



(a) (b)

Figure 12: Curves depicting the relative frequencies of the means $\langle \sigma_i \rangle$ and covariances $\langle \sigma_i \rangle$ in different circumstances. In (a) we look at how the means and covariances are distributed when calculated on the basis of different non-overlapping sub-populations of 91 neurons from the recorded polulation, consisting of 275 neurons. (b) shows how the distributions on the same observables are influenced by including more and more units from the recording data, *i.e.*, the $N = 200$ sample includes the units in the $N = 100$ sample, which in turn includes the units in the $N = 50$ sample.

Overall, as seen if Figs. 12a and 12b, the distributions of the observables are quite similar across distinct sub-populations of the same size, as well as across samples with differing sizes, respectively.

Table 6 shows that not only do the relative frequencies of the observables from the different sub-populations appear similar, but K-S tests also indicate that the neural time-series themselves are generated from the same distribution. This reveals a certain degree of homogeneity in the neural activity, and may indicate that the sub-populations are not organized into specialized functional groups at this level.

Although the overall shape of the distributions are also similar across the sub-populations differing in number of neurons, there appears to be a tendency for the relative frequencies of the observables to become smoother as more neurons are included.

This is likely due to the smaller sub-population being insufficient to capture the statistical properties of

	Samples 1 & 2	**Samples 1 & 3**	**Samples 2 & 3**
Stats.	5.47×10^{-4}	1.62×10^{-3}	1.07×10^{-3}
p-values	99.7%	11.5%	55.8%

Table 6: Test statistics and p-values for 2-sample K-S tests used to compare the samples, *i.e.*, neural time-series of (non-overlapping) sub-populations of the of the neurons that were recorded from In the complete time-series data. Here, the p-values are all quite high, and exceed the conventional significance level of 5% by a factor of two even in the "worst case", indicating that we do not have grounds to conclude that the samples are distributed differently.

the network that subsumes the recorded neurons as a whole[17], i.e., variability in the statistics of individual neurons leads to the need for a larger sample to to accurately capture the broader population's overall behavior.

As confirmed by the K-S test results in Table 7, this change as we go from a smaller to a larger population is seen to be a gradual one: a significant difference in the neural time series is found only when we compare the 50 neuron sample to the 200 neuron one, whilst intermediate samples are not sufficiently dissimilar to be flagged by the K-S test. In particular, the transition from 100-200 neurons results in a P-value of 96.4%, indicating that a population of in the neighborhood of 100 neruons is sufficiently large to capture the statistics of the overall network. If this is true, then this is good news, as larger populations are harder to analyzex.

	50 & 100	**50 & 200**	**100 & 200**
Stats.	1.83×10^{-3}	2.39×10^{-3}	5.59×10^{-4}
p-values	13.8%	0.843%	96.4%

Table 7: Same as Table 6, except here, instead of being non-overlapping sub-populations, the time series come from populations differing in size, and are such that the smaller populations of neurons form sub-populations of the larger ones. We see that adjacent samples are not significantly different from one another, but the null-hypothesis that the samples come from the same distribution is rejected when comparing the $N = 50$ sample to the $N = 200$ one.

4.3 Inference on the Neural Data

Finally, we turn to the inference results. We were able to acheive a somewhat good fit for both models, even on relatively large networks, consisting of $N = 270$ units on the upper end. That being said, as seen in Fig. 13, there is a clear difference in the errors of the two models: as we saw previously in the simulation context, the NEQ model exhibits a closer fit compared with the EQ model. This is also reflected in the corresponding numerical results. See Table 8.

[17] We're making the assumption that the recorded neurons are part of a larger network, however, we have not investigated their synaptic connectivity directly here — though it would have been useful to consider.

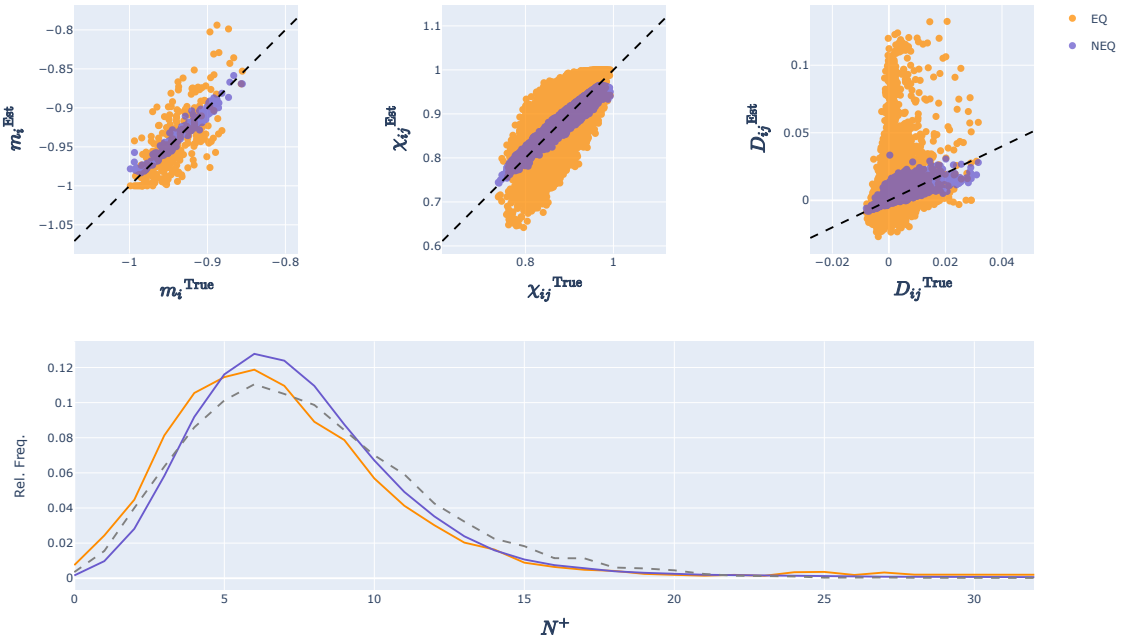


Figure 13: Scatters and relative frequencies illustrating the degree of agreement between the statistics of the fitted EQ and NEQ models to $N = 270$ recorded M2 units. Here, the inference pipeline involved getting a first approximation of the parameters using TAP, followed by multiple runs of gradient descent with learning rates starting at 0.001 for both models, but eventually lowered to 10^{-5} and 5×10^{-4} for the EQ and NEQ models, respectivelly. Model observables were calculated on the basis of $M = 150,000$ MCMC samples.

	RMSE(m_i)	RMSE(χ_{ij})	RMSE(D_{ij})	**p-values,** N^+	**Test stat.,** N^+
EQ	0.03	0.039	0.009	6.89×10^{-64}	8.10×10^{-2}
NEQ	0.007	0.008	0.002	6.83×10^{-21}	4.60×10^{-2}

Table 8: RMSE values and K-S test results corresponding to the scatters and relative frequencies depicted in Fig. 13.

Notably, the models capture not only the observables in the neural data that was used by the inference algorithms, but also generalize to unseen statistical features of the neural activity. As we saw in the bottom panel of Fig. 13, the NEQ model also distinguished itself in having better predictions of unseen statistics, and yielded a co-activation distribution, over the number N^+ of simultaneously active units that more closely agreed with the recording data than than did the EQ model. Another example are the third order covariances, depicted in Fig. 14. Note that, due to memory constraints, the third order covariances were calculated on the basis of recordings from a smaller sub-population, consisting of $N = 100$ neurons.

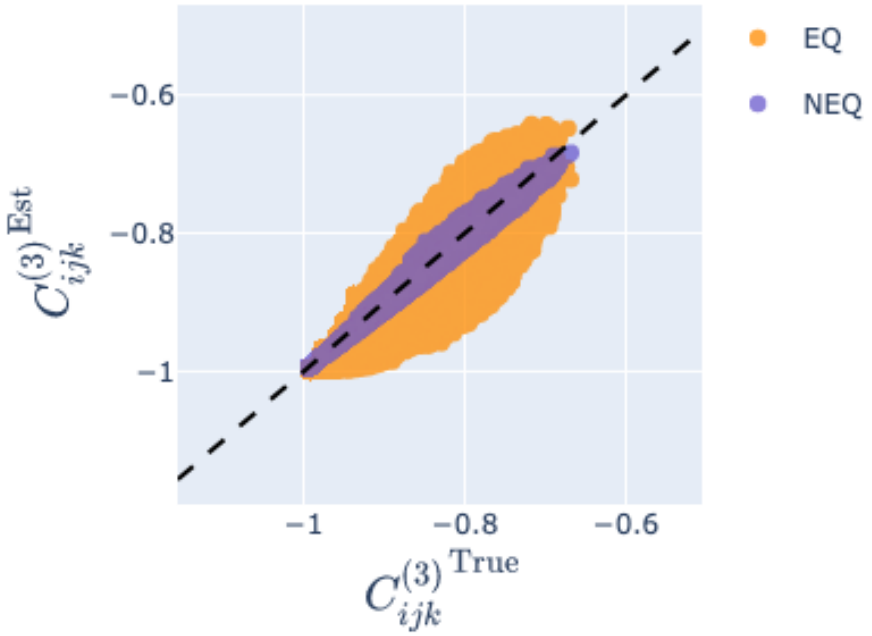
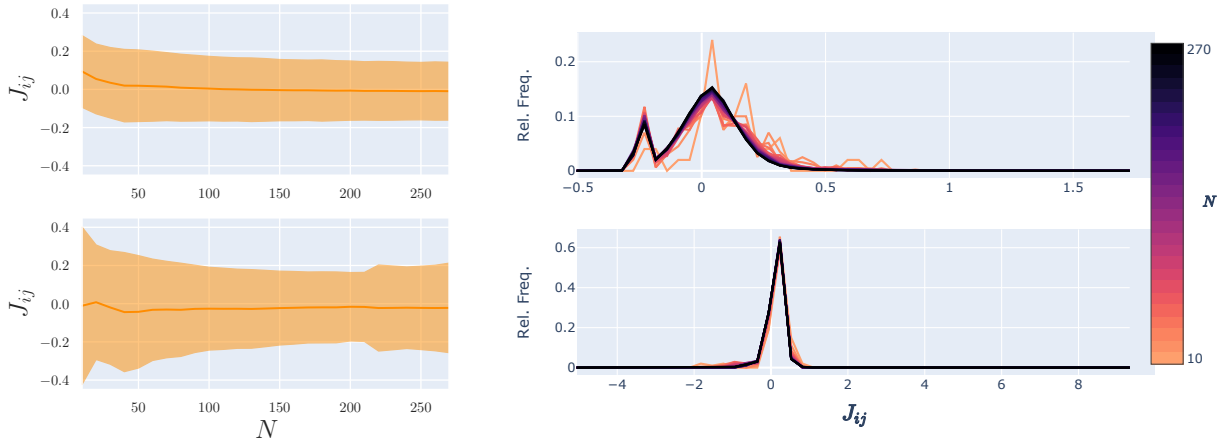


Figure 14: Scatters depicting the degree agreement between third order (non-centered) covariances calculated on the basis of the recording data itself and from $M = 150,000$ MCMC simulations from EQ and NEQ models fitted to $N = 100$ recorded M2 units. See Fig. 17a for scatters for the observables used in fitting the models. Consistent with what we see in the scatter the RMSE values were 0.038 and 0.008 for the EQ and NEQ models, respectively, indicating that the NEQ model does a better job of predicting statistical properties of the recording data not used by the inference algorithms.

In order to understand how the two models were influenced by network size, we also fitted them to smaller sub-populations of the recorded neurons, ranging in size from 10 to 270 neruons, in increments of 10 neurons[18]. The results of this are seen in Fig. 15.

Starting with Fig. 15a, we see that the mean connection strength for both models remained fairly stable as the system size was increased from $N = 10$ to $N = 270$, albeit with a downwards trend for the EQ model. This downwards trend is consistent with what was found in [42]. However, they were reconstructing a simulated network with couplings that were randomly sampled and without a spatial structure, which would likely carry over to the inferred model. In our case, on the other hand, we're dealing with a real biological neural network, and so it would be ill advised to assume the same assumptions should hold here. We also note that the decay in the mean coupling stength seen in our data seems to flatten out much quicker than in the aforementioned simulation study, and may simply be an instance of the smaller sub-populations in the sequence consisting of neurons whose firing-rate is below the population average.

Accross the network sizes, the NEQ model exhibited a greater standard deviation than did the EQ model. Whereas the standard deviation remained mostly stable for the EQ model, that of the NEQ model was seen to fluctuate, suggesting that the NEQ parameters may be detected features in the data that the EQ model misses. Consistent with this is the observabion, as in Fig.15b, the distribution of couplings in NEQ model is much more heavy-tailed than that of the EQ couplings. Despite it's longer vier tails, the NEQ couplings appear to remain much more consistent in overall shape as the system size is increased. The EQ couplings are especially noisy going from $N = 10$ to around $N = 50$, consistent with the shift in the mean observed in Fig. 15a, and the interpretation of this being due simply to small sample effects. Next, we observe that the EQ couplings has a bipolar distribution, which was also found in [44] and [45].

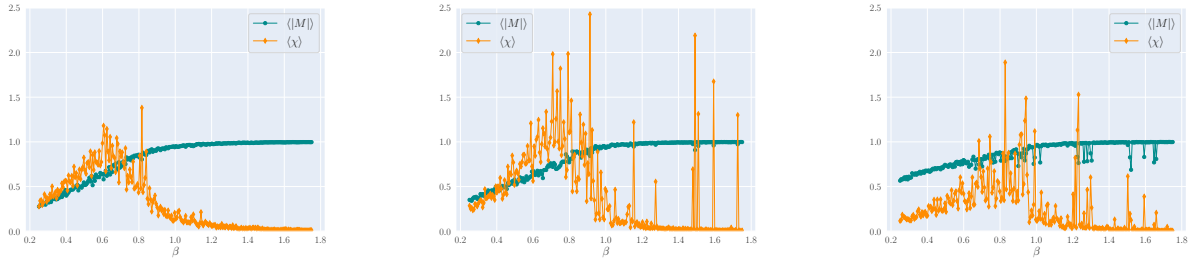


(a) (b)

Figure 15: The effect of system size on the distribution of inferred couplings in the EQ model (top panel) and NEQ model (bottom panel). (a) depicts the mean connection strengths as a function of the number N of recorded units used to fit the models, and with the width of the shaded regions on either side being the standard deviations. In (b), we show the relative frequencies of the coupling values for N ranging for 10 to 270 units. The latter model being the one depicted in Fig. 13

We now look again at the mean absolute magnetization $\langle |M| \rangle$ and the magnetic susceptibility χ, as we did in the introduction, but this time as a function of the inverse temperature β. See Fig. 16. Here, β plays the role of a scaling constant which we can use to rescale the inferred parameters: $\vec{h}, \mathbf{J} \to \beta\vec{h}, \beta\mathbf{J}$. Clearly then, the inferred model corresponds to $\beta = 1$.

[18]Examples of scatters, like in Fig. 13, for the cases $N = 50$ and $N = 100$ are depicted in Fig. 17.



(a) (b) (c)

Figure 16: The phase of the inferred EQ model. Similar to Fig. 4, but with the mean absolute magnetization and the magnetic susceptibility plotted as funcions of β. This is obtained, as in [Nguyen et al.] by scaling the inferred parameters by a factor β, so that weaker parameters correspond to $\beta < 1$, the inferred model to $\beta = 1$, and stronger parameters to $\beta > 1$. (a), (b) and (c) show this for $N = 50,\ 100$ and 270, respectively. The latter corresponding to the model whose observables are depicted in Fig. 13. See Fig. 17 for comparison scatters for the $N = 50$ and 100 model.

The inferred EQ model appears to be quite close to the critical point. This observation is consistent with findings [28, 46, 47, 48] which indicate that neural activity is in a critical state, striking a balance between being excessively orderly or disorderly, as well as the one discussed in [28], that β moves closer to 1 as the system size is increased. Does it therefore corroborate the *critical brain hypothesis* – that neural neural activity is situated in this critical state? [47] argue that this could endow neural activity with certain useful characteristics which facilitate long distance communication, information storage, computational power, and stability.

In [48], however, Mastromatteo et al. caution that models for which it is possible to accurately infer parameters on the basis of limited data are highly likely to appear close to the critical point, because this region of the parameter space is the only one that has a high density of models whose likelihood differs significantly between different parameter configurations. This observation, therefore, could well be an artefact of the inference procedure.

5 Concluding remarks

We set out to investigate the extent to which the equilibrium and non-equilibrium Ising models are able to capture the essential statistical features of cortical population activity measured using calcium imaging, and to assess the performance of the maximum likelihood and mean-field methods under different circumstances.

The latter objective was addressed using simulations from the models themselves, and it was observed that for both models, the overall performance of the inference methods followed the same ordering as has been found by others [28, 38], and that the mean field methods are more sensitive to β [28] as well as system size [38] compared with the maximum likelihood methods.

With the neural data, we were able to obtain satisfactory predictions for relatively large systems, ableit with significant noise in the case of the EQ model. This occurs, as discussed in Sec. 3.2.2, because in contrast with the Boltzmann learning algorithm, used for the EQ model, the NEQ maximum likelihood algorithm does not use sampling to estimate gradients at each step, thereby suffering from one less source of error than the EQ model. Sampling is nevertheless used in calculating observables to compare with the statistics of the neurall data, and here too the NEQ procedure performs better than the EQ one, with parallel dynamics leading to a greater effective sample size. Returning to the maximum likelihood methods, the third reason we've discussed regarding why the NEQ model performs better than the EQ one is that, unlike it's equilibrium counterpart, it uses the temporal information in the data, treating the neural time-series as markov-chains rather than independent snapshots of neuronal activity. While this latter point is important in reconstruction of known simulated models, it remains relevant when dealing with real neural data where the assuption of independent network configurations is unlikely to hold.

Both systems were able to predict statistical features of the data not used by the inference algorithms, and thereby exhibiting some degree of generalization, corroborating the findings of previous studies such as [49], where it was also fund that pairwise binary maximum entropy models (what we've referred to as the EQ model) is capable of predicting third order covariances and the co-activation distribution of neural activity using two-photon calcium imaging in the hippocampus. What we have shown is that this model can also be fruitfully applied to recordings from PPC neurons, recorded using one-photon calcium imaging, and that the same holds for the related kinetic Ising model.

Here too, however, the NEQ model displayed superior performance, with lower errors on both third-order covariances and the co-activation distribution. Additionally, others have found that whilst the EQ model fails to predict the synaptic connectivity of realistic cortical network models, the NEQ model was found to succeed in this task [38, 44, 45]. In our case, we did not have direct access to the the underlying synaptic connectivity, and so cannot validate our findings against any kind of ground truth. As mentioned, these are *in silico* studies, and so it would be interesting to do a similar kind of analysis, but with recordings from biological neural networks as well as anatomical information about the network structure.

Despite not knowing the true connectivity of the recorded networks, a solid estimate of the joint distribution of the network states remains a worthwhile object of study in and of itself and, by extention, the parameters of the inferred models. We therefore made an assessment of the inferred couplings, and found that there are clear differences between the connections obtained for the EQ and NEQ models. In particular, the NEQ couplings are more heavy-tailed, indicating that under this model, most neurons are not connected, or have very weak connections, while there are a small number of neurons with strong projections. Similar characteristics has also been observed in the synaptic connectivity visual cortex neurons in rats [50].

Besides the ones discussed so far, there are some additional limitaions of this work that we consider worth discussing here. While we've argued that the incarnation of the kinetic Ising model that was used in this project is based on more biologically realistic assumptions by virtue of allowing for asymmetric connectivity, and considering the dynamic nature of synaptic transmission. There are, however, other approaches to endowing the equilibrium model with temporal dynamics; one that possesses even greater biological realism than the one used here is to replace the parallel dynamics with asynchronous updating and let the neurons be updated independently and at random times. The resulting time series are then realizations of a doubly stochastic process. Besides this, there are many alternative paths that the data set used in this project might have permitted us to take, but which we did not endeavor to pursue due to time constraints. One surch path would be to investigate the relationship between the cortical recordings and behavioral data, and perhaps assess if the models can be used in decoders to predict the behavior based on network states, as was done in [51], where they built a decoder using the EQ model to decode simulated neural activity from layer

V of the mouse visual cortex and predict the orientation stimulus that elicited the response. We also had recordings from both PPC and M2, and so could in principle address questions regarding differences in the neural activity in these areas. The M2 and PPC recordings, however, were from different mice, and so such comparisons would be more difficult to interpret.

Despite it's performance relative to the NEQ model, the EQ model still has some important advantages related to how much we know about systems at equilibrium and the ease of interpretation using the metaphor of energy landscapes. This points to the potential value for a better understanding, perhaps using the tools of statistical physics, of more realistic models, such as the NEQ model, and beyond.

Finally, we note many other variations of these models have been developed. Some of these may have properties, such as time (or stimulus) dependent model pa rameters [35, 36] that could have been useful for investigating the relationship between the behavioral conditions the animals were subjected to and the neural activity, as well as non-binary variables, which could potentially be better suited for modelling calcium imaging recordings [42].

References

[1] JW Deitmer, AJ Verkhratsky, and C Lohr. "Calcium signalling in glial cells". In: *Cell calcium* 24.5-6 (1998), pp. 405–416.

[2] Eric R Kandel et al. *Principles of neural science*. Vol. 4. McGraw-hill New York, 2000.

[3] Daniel J Amit and Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge university press, 1992.

[4] John A Hertz. *Introduction to the theory of neural computation*. CRC Press, 2018.

[5] Tuce Tombaz et al. "Action representation in the mouse parieto-frontal network". In: *Scientific reports* 10.1 (2020), pp. 1–14.

[6] Pengcheng Zhou et al. "Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data". In: *Elife* 7 (2018), e28728.

[7] Johannes Friedrich, Pengcheng Zhou, and Liam Paninski. "Fast online deconvolution of calcium imaging data". In: *PLoS computational biology* 13.3 (2017), e1005423.

[8] Robert M Dowben and Jerzy E Rose. "A metal-filled microelectrode". In: *Science* 118.3053 (1953), pp. 22–24.

[9] JD Green. "A simple microelectrode for recording from the central nervous system". In: *Nature* 182.4640 (1958), pp. 962–962.

[10] David H Hubel and Torsten N Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of physiology* 160.1 (1962), pp. 106–154.

[11] Bernt Christian Skottun. "Sound localization and neurons". In: *Nature* 393.6685 (1998), pp. 531–531.

[12] Trevor M Shackleton et al. "Interaural time difference discrimination thresholds for single neurons in the inferior colliculus of guinea pigs". In: *Journal of Neuroscience* 23.2 (2003), pp. 716–724.

[13] Romain Brette. "Is coding a relevant metaphor for the brain?" In: *Behavioral and Brain Sciences* 42 (2019).

[14] MB Cannell, JR Berlin, and WJ Lederer. "Intracellular calcium in cardiac myocytes: calcium transients measured using fluorescence imaging". In: *Soc Gen Physiol Ser* 42 (1987), pp. 201–214.

[15] Kensall D Wise, James B Angell, and Arnold Starr. "An integrated-circuit approach to extracellular microelectrodes". In: *IEEE transactions on biomedical engineering* 3 (1970), pp. 238–247.

[16] Roger D Rosenkrantz. *ET Jaynes: Papers on probability, statistics and statistical physics*. Vol. 158. Springer Science & Business Media, 2012.

[17] Michael S Gazzaniga. *The cognitive neurosciences*. MIT press, 2009.

[18] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[19] George Somjen. "Variables of the sensory code". In: *Sensory Coding in the mammalian nervous system*. Springer, 1972, pp. 33–49.

[20] David C Knill and Alexandre Pouget. "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *TRENDS in Neurosciences* 27.12 (2004), pp. 712–719.

[21] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[22] Edwin T Jaynes. "Information theory and statistical mechanics". In: *Physical review* 106.4 (1957), p. 620.

[23] Pierre Curie. *Propriétés magnétiques des corps a diverses températures*. 4. Gauthier-Villars et fils, 1895.

[24] Walter Greiner, Ludwig Neise, and Horst Stöcker. *Thermodynamics and statistical mechanics*. Springer Science & Business Media, 2012.

[25] Martin Niss. "History of the Lenz-Ising model 1920–1950: from ferromagnetic to cooperative phenomena". In: *Archive for history of exact sciences* 59.3 (2005), pp. 267–318.

[26] Rudolf Peierls. "On Ising's model of ferromagnetism". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 32. 3. Cambridge University Press. 1936, pp. 477–481.

[27] Lars Onsager. "Crystal statistics. I. A two-dimensional model with an order-disorder transition". In: *Physical Review* 65.3-4 (1944), p. 117.

[28] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. "Inverse statistical problems: from the inverse Ising problem to data science". In: *Advances in Physics* 66.3 (2017), pp. 197–261.

[29] John J Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.

[30] Geoffrey E Hinton and Terrence J Sejnowski. "Optimal perceptual inference". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. Vol. 448. Citeseer. 1983.

[31] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. "A learning algorithm for Boltzmann machines". In: *Cognitive science* 9.1 (1985), pp. 147–169.

[32] Paul Smolensky. *Information processing in dynamical systems: Foundations of harmony theory*. Tech. rep. Colorado Univ at Boulder Dept of Computer Science, 1986.

[33] Geoffrey E Hinton. "Training products of experts by minimizing contrastive divergence". In: *Neural computation* 14.8 (2002), pp. 1771–1800.

[34] Elad Schneidman et al. "Weak pairwise correlations imply strongly correlated network states in a neural population". In: *Nature* 440.7087 (2006), pp. 1007–1012.

[35] Rodrigo Quian Quiroga and Stefano Panzeri. *Principles of neural coding*. CRC Press, 2013.

[36] Yasser Roudi and John Hertz. "Mean field theory for nonequilibrium network reconstruction". In: *Physical review letters* 106.4 (2011), p. 048702.

[37] Yasser Roudi and John Hertz. "Dynamical TAP equations for non-equilibrium Ising spin glasses". In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.03 (2011), P03031.

[38] Yasser Roudi, Joanna Tyrcha, and John Hertz. "Ising model for neural data: model quality and approximate methods for extracting functional connectivity". In: *Physical Review E* 79.5 (2009), p. 051915.

[39] Timm Plefka. "Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model". In: *Journal of Physics A: Mathematical and general* 15.6 (1982), p. 1971.

[40] HJ Kappen and JJ Spanjers. "Mean field theory for asymmetric neural networks". In: *Physical Review E* 61.5 (2000), p. 5658.

[41] Daniel L Stein and Charles M Newman. *Spin glasses and complexity*. Vol. 4. Princeton University Press, 2013.

[42] Yasser Roudi, Erik Aurell, and John A Hertz. "Statistical physics of pairwise probability models". In: *Frontiers in computational neuroscience* 3 (2009), p. 22.

[43] Yasser Roudi, Sheila Nirenberg, and Peter E Latham. "Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't". In: *PLoS computational biology* 5.5 (2009), e1000380.

[44] Joanna Tyrcha et al. "The effect of nonstationarity on models inferred from neural data". In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.03 (2013), P03005.

[45] John A Hertz et al. "Inferring network connectivity using kinetic Ising models". In: *BMC neuroscience* 11 (2010), pp. 1–2.

[46] Gasper Tkacik et al. "Ising models for networks of real neurons". In: *arXiv preprint q-bio/0611072* (2006).

[47] John M Beggs. "The criticality hypothesis: how local cortical networks might optimize information processing". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366.1864 (2008), pp. 329–343.

[48] Iacopo Mastromatteo and Matteo Marsili. "On the criticality of inferred models". In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.10 (2011), P10012.

[49] Leenoy Meshulam et al. “Collective behavior of place and non-place neurons in the hippocampal network”. In: *Neuron* 96.5 (2017), pp. 1178–1191.

[50] Sen Song et al. “Highly nonrandom features of synaptic connectivity in local cortical circuits”. In: *PLoS biology* 3.3 (2005), e68.

[51] Michael T Schaub and Simon R Schultz. “The Ising decoder: reading out the activity of large neural ensembles”. In: *Journal of computational neuroscience* 32 (2012), pp. 101–118.

A Miscellaneous derivations

A.1 Analytic expressions for the IP observables

Here, we'll derive analytic expressions for the magnetizations and covariances for the independent-pair (IP) model, which is characterized by having units that that have a non-zero connection with at most one other unit. Let $\mathcal{P}$ denote the set of pairs, and assume each pair only occurs once in $\mathcal{P}$, *i.e.*,

$$(k,l) \in \mathcal{P} \implies (l,k) \notin \mathcal{P}.$$

.

Denoting $\mathcal{H}_{k,l}(\sigma_k, \sigma_l) = J_{kl}\sigma_k\sigma_l + h_k\sigma_k + h_l\sigma_l$, we have:

$$\begin{aligned}
\mathcal{H}(\vec{\sigma}) &= -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \\
&= -\sum_{(k,l)\in\mathcal{P}} J_{kl}\sigma_k\sigma_l - \sum_i h_i\sigma_i \\
&= -\sum_{(k,l)\in\mathcal{P}} (J_{kl}\sigma_k\sigma_l + h_k\sigma_k + h_l\sigma_l) \\
&= \sum_{(k,l)\in\mathcal{P}} \mathcal{H}_{k,l}(\sigma_k, \sigma_l).
\end{aligned}$$

Moreover, letting $Z_{kl} = \sum_{\sigma_k=\pm1}\sum_{\sigma_l=\pm1} \exp\big(\mathcal{H}_{k,l}(\sigma_k,\sigma_l)\big)$, we can simplify the expression for the partition function:

$$\begin{aligned}
Z &= \sum_{\vec{\sigma}} \exp\big(-\mathcal{H}(\vec{\sigma})\big) \\
&= \sum_{\vec{\sigma}} \exp\big(-\sum_{(k,l)\in\mathcal{P}} \mathcal{H}_{k,l}(\sigma_k,\sigma_l)\big) \\
&= \sum_{\vec{\sigma}} \prod_{(k,l)\in\mathcal{P}} \exp\big(-\mathcal{H}_{k,l}(\sigma_k,\sigma_l)\big) \\
&= \prod_{(k,l)\in\mathcal{P}} \sum_{\sigma_k=\pm1}\sum_{\sigma_l=\pm1} \exp\big(-\mathcal{H}_{k,l}(\sigma_k,\sigma_l)\big) \\
&= \prod_{(k,l)\in\mathcal{P}} Z_{kl}
\end{aligned}$$

Next, we turn to the observables. Assuming units i and j are connected, we have:

$$\begin{aligned}
\langle\sigma_i\rangle &= \frac{1}{Z}\sum_{\vec{\sigma}} \sigma_i \exp\left(-\mathcal{H}(\vec{\sigma})\right) \\
&= \frac{\sum_{\vec{\sigma}} \sigma_i \prod_{(k,l)\in\mathcal{P}} \exp\big(-\mathcal{H}_{k,l}(\sigma_k,\sigma_l)\big)}{\prod_{(k,l)\in\mathcal{P}} Z_{kl}} \\
&= \frac{\sum_{\sigma_i,\sigma_j} \exp\big(-\mathcal{H}_{ij}(\sigma_i,\sigma_j)\big) \prod_{(k,l)\in\mathcal{P}\setminus(i,j)} \sum_{\sigma_k,\sigma_l} \exp\big(-\mathcal{H}_{kl}(\sigma_k,\sigma_l)\big)}{Z_{ij}\prod_{(k,l)\in\mathcal{P}\setminus(i,j)} Z_{kl}} \\
&= \frac{1}{Z_{ij}}\sum_{\sigma_i,\sigma_j} \exp\big(-\mathcal{H}_{ij}(\sigma_i,\sigma_j)\big) \\
&= \frac{e^{-\mathcal{H}_{ij}(+1,+1)} + e^{-\mathcal{H}_{ij}(+1,-1)} - e^{-\mathcal{H}_{ij}(-1,+1)} - e^{-\mathcal{H}_{ij}(-1,-1)}}{e^{-\mathcal{H}_{ij}(+1,+1)} + e^{-\mathcal{H}_{ij}(+1,-1)} + e^{-\mathcal{H}_{ij}(-1,+1)} - e^{-\mathcal{H}_{ij}(-1,-1)}}.
\end{aligned}$$

Repeating this exercise for the covariances, we obtain:

$$\langle \sigma_i \sigma_j \rangle = \frac{e^{-\mathcal{H}_{ij}(+1,+1)} - e^{-\mathcal{H}_{ij}(+1,-1)} - e^{-\mathcal{H}_{ij}(-1,+1)} + e^{-\mathcal{H}_{ij}(-1,-1)}}{e^{-\mathcal{H}_{ij}(+1,+1)} + e^{-\mathcal{H}_{ij}(+1,-1)} + e^{-\mathcal{H}_{ij}(-1,+1)} - e^{-\mathcal{H}_{ij}(-1,-1)}}.$$

Finally, being non-centered covariances, if $k \neq j$ we have $\langle \sigma_i \sigma_k \rangle = \langle \sigma_i \rangle \langle \sigma_k \rangle$.

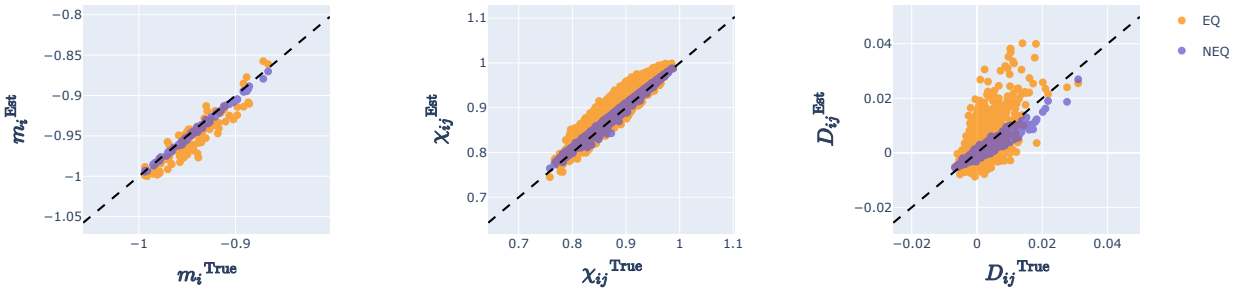
A.2 Expected activations under the non-equilibrium model

Under non-equilibrium model, the expected activation for a given unit i conditioned on the previous state of the network is easily derived. Starting from Eq. 13, we observe that,

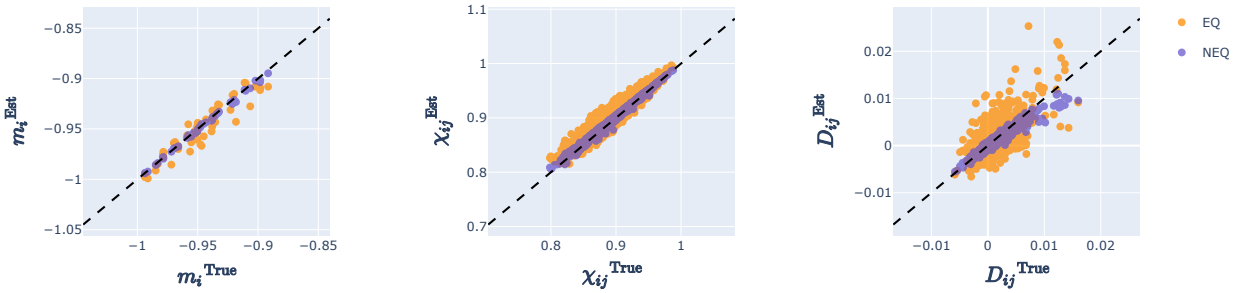
$$\begin{aligned} \langle \sigma_i(t+1) | \vec{\sigma}(t) \rangle &= (+1) P(\sigma_i(t+1) = +1 | \vec{\sigma}(t)) + (-1) P(\sigma_i(t+1) = -1 | \vec{\sigma}(t)) \\ &= \frac{e^{H_i(t)}}{2 \cosh H_i(t)} - \frac{e^{-H_i(t)}}{2 \cosh H_i(t)} = \tanh(H_i(t)). \end{aligned}$$

Note also that $\langle \sigma_i(t+1) \sigma_j(t) | \vec{\sigma}(t) \rangle = \tanh(H_i(t)) \sigma_j(t)$.

B Additional Figures



(a)



(b)

Figure 17: Same as the top panel of Fig. 13, but for $N = 100$ in (a) and $N = 50$ in (b).