Journal of Computational Neuroscience manuscript No.
(will be inserted by the editor)

The Ising decoder: reading out the activity of large neural ensembles

Michael T Schaub · Simon R Schultz

the date of receipt and acceptance should be inserted later

Abstract The Ising Model, a pairwise binary maximum entropy model, has recently received much attention for the statistical description of neural spike train data. In this paper, we propose and demonstrate its use for building decoders capable of predicting, on a millisecond timescale, the stimulus represented by a pattern of neural activity. After fitting to a training dataset, the Ising decoder can be applied "online" for instantaneous decoding of test data. While such models can be fit exactly using Boltzmann learning, this approach rapidly becomes computationally intractable as neural ensemble size increases. We show that several approaches, including the Thouless-Anderson-Palmer (TAP) mean field approach from statistical physics, and the recently developed Minimum Probability Flow Learning (MPFL) algorithm, can be used for rapid inference of model parameters in large-scale neural ensembles. Use of the Ising model for decoding, unlike other problems such as functional connectivity estimation, requires estimation of the partition function. As this involves summation over all possible responses, this step can be limiting. Mean field approaches avoid this problem by providing an analytical expression for the partition function. We demonstrate these decoding techniques by applying them to simulated neural ensemble responses from a mouse visual cortex model, finding an improvement in decoder performance for a model with heterogeneous as opposed to homogeneous neural tuning and response properties. In particular, we demonstrate decoding of spatial patterns of activity comprised of over 700 neurons with the TAP algorithm. Our results demonstrate the suitabilty of Ising and Ising-like models for reading out, or decoding, spatial patterns of activity in large neural ensembles.

Department of Bioengineering, Imperial College London,
South Kensington, London SW7 2AZ, UK.
E-mail: s.schultz@imperial.ac.uk

1 Introduction

Interpreting the patterns of activity fired by populations of neurons is one of the central challenges of modern systems neuroscience. The design of decoding algorithms capable of millisecond-by-millisecond readout of sensory or behavioural correlates of neuronal activity patterns would be a valuable step in this direction. Such decoding algorithms, as well as helping us to understand the neural code, may have further practical application, as the basis of communication neural prostheses for severely disabled patients such as those with "Locked In" syndrome.

At the heart of such a decoding algorithm must lie - whether explicit or implicit - a description of the conditional probability distribution of activity patterns given stimuli or behaviours. Making this description is nontrivial, as the brain, like other biological systems, exhibits an enormous complexity, arising from the complex interdependencies of a huge underlying network of interacting elements. This results in a very large number of possible states or configurations exhibited by the system, rapidly making the description of such systems by simply measuring the probabilities of each state unfeasible. Except for very small patterns, a model-based approach of some kind is essential.

New technologies in neuroscience such as high-density multi-electrode array recording and multi-photon calcium imaging now make it possible to monitor the activity of large numbers of neurons simultaneously. Analysis tools for such high dimensional data have however lagged behind the experimental technology, as most approaches are limited to very small population sizes. While considerable advances have been made in the use of information-theoretic approaches to characterise the statistical structure of small neural ensembles (Gawne et al. 1996, Panzeri, Schultz, Treves & Rolls 1999, Schultz & Panzeri 2001, Panzeri & Schultz 2001, Reich et al. 2001, Petersen et al. 2001, Pola et al. 2003, Montani

et al. 2007), finite sampling limitations have made results for larger ensembles much more difficult to obtain.

For the statistical description of multivariate neural spike train data, parametric models able to capture most of the interesting features of real data while still being of empirically accessible dimensionality are highly desirable. One promising approach has emerged from statistical mechanics: the use of Ising (or Ising-like) models, exploiting an analogy between populations of spike trains and ensembles of interacting magnetic spins (Shlens et al. 2006, Shlens et al. 2009, Schneidman et al. 2006). The Ising model is a pairwise binary maximum entropy model, and thus is a natural way to describe the statistics of binary neural spike patterns. Fitting of these models to observed neural data has the advantage that it does not implicitly assume some non-measured structure in the data, i.e. maximum entropy models express the most uncertainty about the modelled data given the chosen constraints (e.g. that certain moments of the measured distribution agree with the model distribution) (Jaynes 1957). It can be shown that this is mathematically equivalent to maximising the likelihood of the model parameters to explain the observed data (Berger et al. 1996). By using this approach to fit a model to the conditional activity pattern distribution, in conjunction with maximum a posteriori Bayesian decoding (Földiák 1993, Oram et al. 1998), it is possible to train a decoder which takes as its input a pattern of spiking activity, and gives as its output the stimulus that it determines to have elicited that spike pattern.

We are interested in devising an algorithm for "millisecond-by-millisecond" decoding, on the basis that information processing in the nervous system appears to make use of such fine temporal scales (Carr 1993, Bair & Koch 1996). The timescale of the "symbols" used in information processing is thus likely to be somewhere between 1 and 20 ms for most purposes (Butts et al. 2007). For time bins on this scale, neural spike trains can be binarized without loss of information, and the simplest model (in the maximum entropy sense) that captures pairwise correlations is the Ising model. Activity states in an Ising model are Boltzmann distributed, i.e. the are distributed according to the negative exponential of the "energy" associated with each state. This distribution, $P \propto \exp(-\sum_\mu \lambda_\mu f_\mu)$, is the maximum entropy distribution subject to the set of constraints imposed by Lagrange multipliers λ_μ on variables f_μ. Imposing these constraints upon firing rates and pairwise correlations gives

$$p_{\text{Ising}}(\mathbf{r}|s) = \frac{1}{\mathcal{Z}(s)} \exp\left(\sum_i h_i(s) r_i + \frac{1}{2}\sum_{i\neq j} J_{ij}(s) r_i r_j\right), \tag{1}$$

where $\mathbf{r} = (r_1, r_2, \ldots, r_C)^T$ and each binary response variable $r_i \in \{0, 1\}$ indicates the firing/not firing of neuron i in the observed time interval. $\mathcal{Z}(s)$ is the partition function, which acts as a normalisation factor. i.e.:

$$\mathcal{Z}(s) = \sum_{\mathbf{r}\in\mathcal{R}} \exp\left(\sum_i h_i(s) r_i + \frac{1}{2}\sum_{i\neq j} J_{ij}(s) r_i r_j\right). \tag{2}$$

Note that the first sum is over all *possible* (as opposed to observed) responses, given by the set $\mathcal{R}$.

The parameters h_i (known in statistical physics as 'external fields') and J_{ij} ('pairwise couplings') have to be fit to the data such that the model displays the same means and pairwise correlations as the data:

$$\langle r_i \rangle_{\text{Ising}} = \langle r_i \rangle_{\text{Data}}\,, \tag{3a}$$

$$\langle r_i r_j \rangle_{\text{Ising}} = \langle r_i r_j \rangle_{\text{Data}}\,, \tag{3b}$$

where $\langle\cdot\rangle_{\text{model}}$ denotes expectation with respect to the specified distribution.

In statistical physics it is more common to use a symmetric representation $\sigma_i \in \{-1, 1\}$ for the 'spins' that describe the activation of neuron i (with -1 indicating 'no spike' and 1 indicating 'spike'), which simply corresponds to a change of variables $\sigma_i = 2r_i - 1$. Accordingly the fields, couplings and partition functions change. As it is occasionally more convenient to work in one or the other representation we will denote the fields and couplings in the spin representation with $\tilde{h}_i$ and $\tilde{J}_{ij}$.

Standard Monte Carlo techniques for fitting these model parameters, such as Boltzmann learning, which can in principle provide an exact solution - given the number of samples is high enough - become computationally intractable as the number of cells increases. This poor scaling behaviour is due to the exponentially increasing number of states with the number of cells. Speeding up the model fitting process is hence an essential requirement to utilize Ising models for studies with large ensembles of neurons.

In the present study we have adopted several approaches for solving this problem. Firstly, we have made use of mean field approximations, including both the 'naive' mean field approximation and the Thouless et al. (1977) (TAP) extension to it, following Roudi, Tyrcha & Hertz (2009). Secondly, we compare this with the recently proposed Minimum Probability Flow Method (Sohl-Dickstein et al. 2009) for learning model parameters. To assess the relative performance of these approaches in the context of the discrete decoding problem, we simulated the activity of a population of neurons in layer V of the mouse visual cortex during an experiment in which a discrete set of drifting grating stimuli were presented. Using this simulation, we have evaluated the relative performance characteristics of the different decoding algorithms in the face of limited data, exploring decoding regimes with up to 200 neurons. Our findings point to the TAP approach being the most useful across a fairly wide parameter range, although in the presence of strong correlations, a regularized variant of MPFL shows its worth.

2 Methods

2.1 Neural Decoding as a Discrete Classification Task

In this paper we consider the problem of the decoding which of a number of different stimuli has elicited a neural spike pattern. This can be seen as a discrete classification task: we have a set of S stimuli $s \in \mathcal{S} = \{s_1, s_2, \ldots, s_S\}$. Decoding in this scenario means that we have to provide a decision rule that estimates which stimulus has been the input to the system, given an observed spike pattern $\mathbf{r}_{\text{obs}}$. The particular example to which we apply this is a simulation of the spike pattern responses elicited by sinusoidal grating stimuli drifting across the receptive field of a visual cortical neuron in different directions: in this case each stimulus s_i represents a different grating direction. However, our main aim with this simulation was to validate our methodology in a coding regime that is a) neurophysiologically realistic, and b) relevant to particular multi-electrode array recording experiments that we plan to perform; we expect that the main results of the paper are not specific to this particular example.

For decoding we use the maximum a posteriori (MAP) rule:

$$\begin{aligned} \hat{s} = \arg\max_s p(s|\mathbf{r}_{\text{obs}}) &= \arg\max_s \frac{p(\mathbf{r}_{\text{obs}}|s)p(s)}{p(\mathbf{r}_{\text{obs}})} \\ &= \arg\max_s p(\mathbf{r}_{\text{obs}}|s)p(s), \end{aligned} \tag{4}$$

where the second step is the application of Bayes' theorem and the third equality holds because $p(\mathbf{r}_{\text{obs}})$ is independent of s and is hence irrelevant for maximising the given expression, i.e. just a constant factor with respect to s that scales the maximum accordingly. Since we are in control of the stimulus distribution $p(s)$, we can choose it to be uniform, i.e. to exhibit the same constant probability for each stimulus and therefore be independent of s, too. Hence our decoding rule simplifies further to the maximum likelihood (ML) rule:

$$\hat{s} = \arg\max_s p(\mathbf{r}_{\text{obs}}|s). \tag{5}$$

With this setting the task of creating a neural decoder reduces to the modelling of the stimulus dependent distributions $p(\mathbf{r}|s)$. Once these are obtained we can apply our ML decoding rule (Equation 5) to estimate the given input stimulus s.

We have used two different statistical models to fit the observed spike patterns for each stimulus. Firstly, we have used an Ising model for $p(\mathbf{r}|s)$, i.e. we assume that for each stimulus, the spike pattern distribution can be described by a (different) Ising model. Secondly, we have used an independent model distribution p_{ind}, assuming that given a stimulus, each cell is independent of the others:

$$p_{\text{ind}}(\mathbf{r}|s) = \prod_{i=1}^{C} p(r_i|s). \tag{6}$$

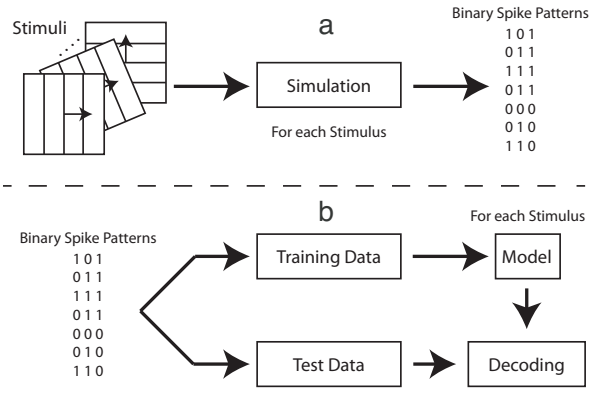

Fig. 1: Schematic description of the decoding process. **a**: For each stimulus we simulate neural responses which are represented as binary response vectors **r**. **b**: The simulated data is divided into test and training subsets. The training data is used to fit a probabilistic model for each stimulus; the test data is then used to assess the performance of the thus trained decoder. This assessment is 10-fold crossvalidated.

The independent model is the binary maximum entropy model of first order, i.e. it takes into account only the first order moments (the constraints on the means given by Equation 3a) and is therefore a natural comparison for the Ising model. As it is very easy to fit the independent model, we used this as a control method, to test whether the more complex Ising model could enhance decoding performance. Note that the numerical values for the probabilities can get very small for large cell ensembles, and therefore to evade finite precision problems we in this case use an equivalent log-likelihood decoding rule instead of the ML rule, i.e. maximise the logarithm of the likelihood instead of maximising the likelihood directly.

2.2 Training and Testing

To train and test the decoders, we proceed as follows:

1. For each stimulus we simulate a set of possible response vectors. The details of the simulation are described in the following subsection.
2. We separate the simulated response patterns into training data, which is used to fit the model and test data which we use evaluate the decoding performance of the obtained models.
3. The whole testing procedure is performed with 10 fold cross-validation, i.e. we divide the whole data for each stimulus into 10 equally sized parts. We then use 9 parts of the data to train our model and the remaining one for testing. We repeat this process again with all 10 possible test/training data set combinations of this kind to reveal if our results generalize to the whole dataset.

This procedure is summarized in Fig. 1.

2.3 Simulation of Evoked Spike Patterns from Mouse Primary Visual Cortex

We simulated the patterns of activity evoked by visual grating stimuli in layer V pyramidal neurons of the anaesthetized mouse visual cortex. The properties of our simulation are motivated by the results reported by Niell & Stryker (2008). Two variants of the model were simulated. In the first (Model A), neuronal direction preferences were uniformly spaced around the circle, and response properties were in general homogeneous. In the second (Model B), neuronal direction preferences were uniformly distributed around the circle, and response properties were in general heterogeneous (Figs. 2a and 3a).

Each neuron's tuning curve was a von Mises function (circular Gaussian) with half width at half maximum either set to the mean value extracted from Fig. 4c of (Niell & Stryker 2008) (Model A), or sampled from a gamma distribution fit to the data presented in that figure (Model B). Each neuron had a direction selectivity index either set to 0.1 (Model A) or obtained from a gamma distribution fit to Niell & Stryker (2008)'s Fig. 5b (Model B). Spontaneous firing rates were similarly set to the median of, or normally distributed (standard deviation 1 Hz, truncated to be never less than 0.3 Hz) around the reported median value for layer V neurons in (Niell & Stryker 2008) Fig. 8d. Transient and evoked sustained firing rates were also motivated by the reported values in Niell & Stryker (2008) publication (Fig.8). However, we selected a slightly higher evoked sustained firing rate set to 9.7 Hz in Model A or normally distributed around 9.7 Hz (standard deviation 4.12 Hz, truncated to be never less than 0.3 Hz), keeping the transient to evoked firing rate ratio at approximately 1.67, close to that inferred from Fig. 8 and supplementary Fig. S3 of Niell & Stryker (2008).

A time-varying response was simulated with a gaussian-shaped transient onset peak in instantaneous firing rate (standard deviation 10 ms) followed by a steady-state period of firing. Onset latency in response to sensory stimulation was 50 ms (Model A) or selected from a normal distribution with mean 50 ms and standard deviation 5 ms (Model B). However, for the purposes of the current study, a time window of duration 20 ms beginning at latency 50 ms was chosen (Figs. 2b-d and 3b-d).

Patterns of spikes fired by the neural population were simulated using a dichotomized Gaussian approach (Macke et al. 2009). Since we cannot estimate covariance matrices from experimental data directly, and not every positive definite symmetric matrix can be used as the covariance matrix of a multivariate binary distribution, we adapted the following approach. First we compute upper and lower covariance bounds for each pair of neurons, according to (Macke et al. 2009)

$$\begin{aligned}\max\left[-pq, -(1-p)(1-q)\right] &\leq \mathrm{Cov}(r_i, r_j) \\ &\leq \min\left[(1-q)p, (1-p)q\right], \end{aligned} \tag{7}$$

where p and q are the means (mean spiking probabilities) of neuron i and j, respectively. We then choose a random symmetric matrix A that lies between these bounds. As this in general does not result in a permissable correlation matrix, a Higham (2002) correction is applied to find the closest correlation matrix possible, to which we finally arithmetically add a random correlation matrix with uniformly distributed eigenvalues to adjust the mean correlation strength.

Where not otherwise stated in the text, 1000 trials per stimulus were simulated, allowing 900 training samples and 100 test samples with 10-fold crossvalidation. In the absence of a detailed characterisation of the correlation structure of neural responses in the mouse visual cortex, we assumed that the correlation in firing between each pair of neurons was weak and positive. Our simulation results in a mean correlation level of 0.11 and a standard deviation of 0.038 (measured with 100000 samples) for Model A and a mean correlation of 0.09 and standard deviation of approximately 0.043 (measured with 100000 samples) for Model B (Fig. 2e and Fig. 3e). Due to limitations of the dichotomized Gaussian simulation, we were not able to specify the correlations exactly, thus all reported correlations are measured and may be prone to small variations. However, such limitations would be inherent to any simulation approach, as i) the covariance structure is always constrained by the firing probabilities of the cells and can not be chosen independently of these firing rates (Macke et al. 2009), and ii) as we are operating in a regime with only a 1000 samples per stimulus, finite sampling effects will affect the simulated data, resulting in fluctuations in the correlation structure.

We were able to vary the (measured) mean correlation level in some simulations, allowing an assessment of the the relative effects of correlation strength for decoding. This included at one extreme bringing the mean level of correlation down to zero (some fraction of pairs thus being anti-correlated), in line with recent results on the distribution of correlations in the "UP" state in rat auditory cortex (Renart et al. 2010). The correlations were generated independently for each stimulus.

2.4 Fitting the Ising Model Parameters

For fitting the model parameters in the Ising model case we use two different strategies: mean field approximations and minimum probability flow learning. In earlier work we used Boltzmann learning (Seiler et al. 2009), however this becomes computationally intractable as the number of neurons exceed 30 or so, and thus we have not reported it here.

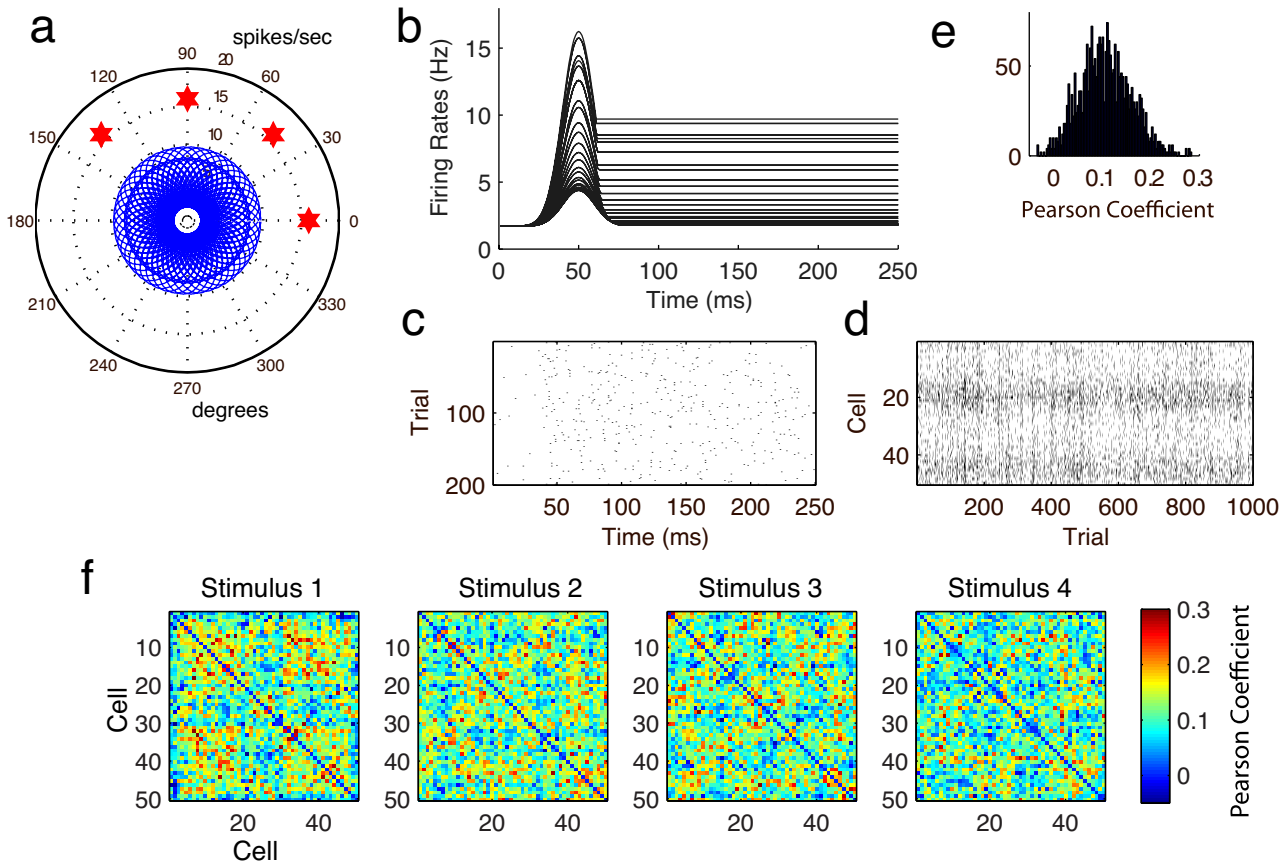

Fig. 2: Neural ensemble responses simulated for Model A (homogeneous scenario). **a** Tuning curves for each cell. Stimulus directions for this example are indicated by red star symbols. **b** Time course of ensemble instantaneous firing rates for one stimulus. **c** Spike rastergram for the same stimulus as (b). **d** A typical dataset (one stimulus, 1000 samples, a black dot indicating a spike fired in the time window between 50 and 70 ms. **e** The average pairwise cross-correlation generated (Pearson correlation coefficient computed with 100000 samples to avoid finite sampling effects, auto-correlation terms excluded). **f** The full correlation matrix for each stimulus (computed with 100000 samples). Diagonal terms set to zero for visualisation purposes only.

2.4.1 Mean Field Methods

Different Mean Field approaches for fitting the model parameters of an Ising model from statistical physics have been recently assessed by Roudi, Tyrcha & Hertz (2009). For fitting the parameters we have compared two different mean field approximations. The 'naive' Mean Field approach, following the nomenclature of Roudi, Tyrcha & Hertz (2009), yields the following expressions for the Ising model parameters

$$\tilde{\mathbf{J}} = \mathbf{P}^{-1} - \mathbf{C}^{-1}, \tag{8a}$$

$$\tilde{h}_i = \tanh^{-1} m_i - \sum_j \tilde{J}_{ij} m_j, \tag{8b}$$

where $\tilde{\mathbf{J}}$ is the estimated coupling matrix with elements $\tilde{J}_{ij}$, the 'magnetization' $m_i = \langle \sigma_i \rangle$, the covariance matrix $\mathbf{C}$ is defined by $C_{ij} = \langle \sigma_i \sigma_j \rangle - m_i m_j$, and similarly $P_{ij} = (1 - m_i^2)\delta_{ij}$.

Use of the naive Mean Field approach for decoding allows the average amount of correlation to be taken into account for decoding, with a small self-coupling correction term given by P_{ii} (referred to as "diagonal weight-trick" in Tanaka (1998)) but no further aspect of its statistical structure. To make use of such information, it must be extended. Again following one of the approaches analysed in Roudi, Tyrcha & Hertz (2009), we use inversion of the TAP equations (Thouless et al. 1977) which can be regarded as a correction to the naive mean field approach, to fit the model parameters (Roudi, Tyrcha & Hertz 2009, Tanaka 1998, Kappen & Rodríguez 1998). Using the TAP approach the equations for the model parameters read:

$$2\tilde{J}_{ij}^2 m_i m_j + \tilde{J}_{ij} + (\mathbf{C}^{-1})_{ij} = 0 \quad (i \neq j), \tag{9a}$$

$$\tilde{h}_i = \tanh^{-1} m_i - \sum_j \tilde{J}_{ij} m_j + m_i \sum_j \tilde{J}_{ij}^2 (1 - m_j^2), \tag{9b}$$

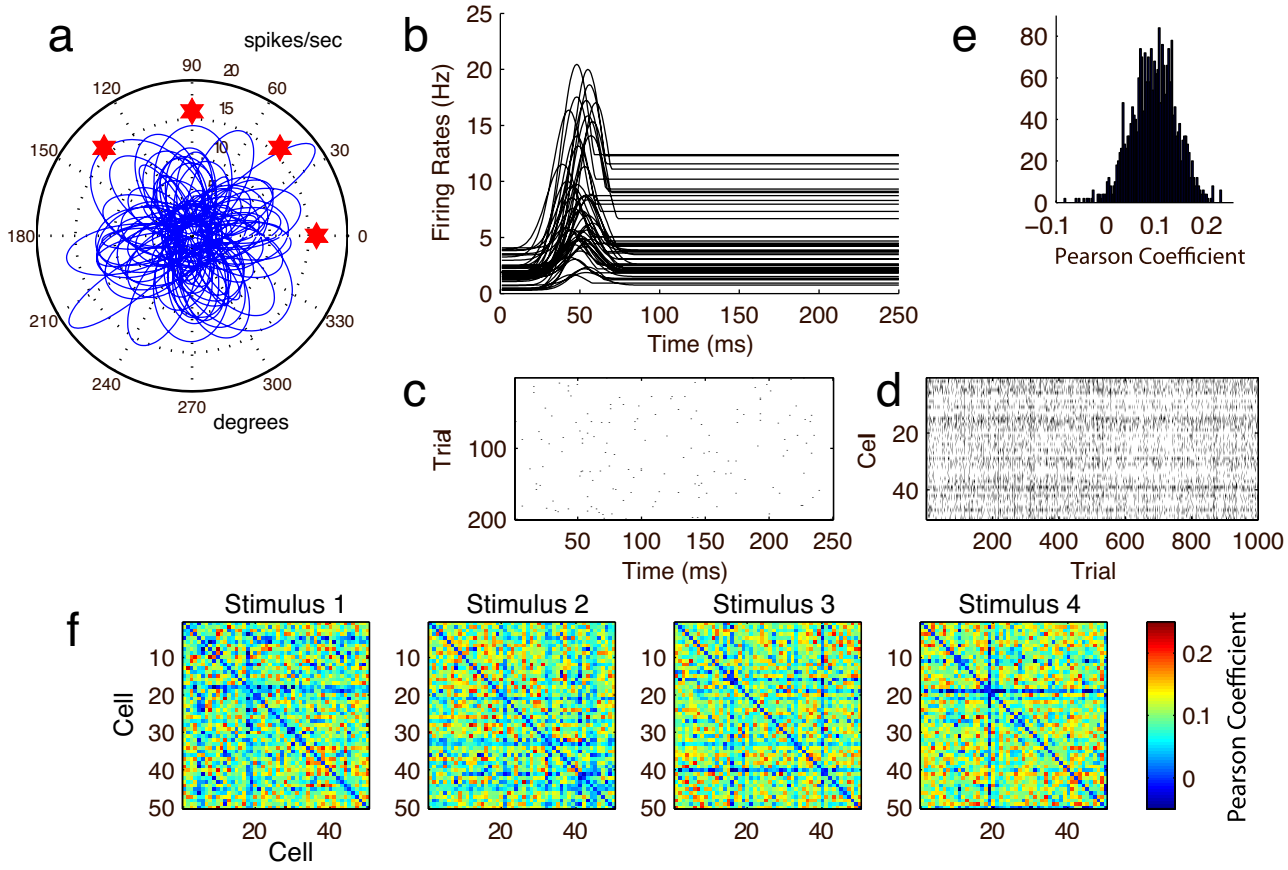

Fig. 3: Neural ensemble responses simulated for Model B (heterogenous scenario). For panel descriptions see Fig. 2

where the first equation can be solved for the pairwise couplings J_{ij} and the correct solution has to be chosen according to continuity conditions outlined in Tanaka (1998), from which then the external fields h_i can be computed. More precisely, if $m_i m_j(\mathbf{C}^{-1})_{ij} < 0$ we choose the solution, which is closer to the original mean field solution. If $m_i m_j(\mathbf{C}^{-1})_{ij} > 0$ we use the standard mean field solution. We use this method as it avoids pairwise couplings becoming complex and respects the continuity of the inverse Ising problem for J_{ij} as a function of $(\mathbf{C}^{-1})_{ij}$.

2.4.2 Minimum Probability Flow Learning

Sohl-Dickstein et al. (2009) recently proposed the Minimum Probability Flow Learning (MPFL) technique, which provides a general framework for learning model parameters. As this technique is also applicable to the Ising model, we have used it to learn external fields and pairwise couplings for our model. However, as the sampling regime usually feasible in neurophysiological experiments dictates a small number of samples compared to the number of parameters in the model (which is C^2 with C cells), the learning problem for the parameters becomes under-constrained already at intermediate neural ensemble sizes, i.e. we are likely to have more parameters to fit than there are samples.

Therefore we introduced a regularization term to their original objective function to penalize model parameters growing to large numbers, i.e. to avoid overfitting. Given the original objective function $K(\theta)$ with θ being the parameters of our model, our regularized objective function reads:

$$K_{\text{reg}}(\theta) = K(\theta) + \frac{\lambda}{2}\|\theta\|_2^2, \tag{10}$$

where $\|\cdot\|_2$ is the L_2 norm, which is a common choice of regularization term (Bishop 2007). So for the Ising model case we have:

$$\|\theta\|_2^2 = \sum_i h_i^2 + \sum_{i \neq j} J_{ij}^2.$$

For the present work the regularization parameter was set to $\lambda = 0.0075$, after systematically assessing different settings for an ensemble size of 50 cells. We refer to this learning algorithm as rMPFL in the rest of the manuscript.

Other choices for the regularization term are possible and might even result in better performance for decoding

purposes, e.g. two independent penalty terms for the external fields and the pairwise couplings. However an extensive assessment of different parameter settings would be very time consuming due to the cost of calculating the invoked partition function. While it is not necessary to compute the partition function for some applications of MPFL (e.g. if learning the J_{ij} parameters is the end in itself), it is required for decoding. We therefore have not performed an exhaustive analysis of regularization.

2.5 Partition Function Estimation

Estimating the partition function is computationally expensive task if done analytically, since the set of possible responses $\mathcal{R}$ grows exponentially with the number of cells C, rendering an analytical computation (Equation 2) intractable for large neural ensembles.

As MPFL learning does not provide an estimate of the partition function, we use the Ogata-Tanemura partition function estimator (Ogata & Tanemura 1984, Huang & Ogata 2001), which is based on Markov Chain Monte Carlo (MCMC) techniques. Since MCMC is still a very time consuming technique, we estimate the partition function only once for each stimulus in a 10 fold cross-validation run when using MPFL, as ideally all samples should come from the same distribution, thus approximately sharing the same partition function.

When fitting the model parameters with mean field theoretic approaches, we computed the (true) partition function $\mathcal{Z}(s)$ in the mean field approximation, as reported in Kappen & Rodríguez (1998) and Thouless et al. (1977).

For the 'naive' mean field approach this yields:

$$\begin{aligned}\log \mathcal{Z} = &\sum_i \log\left(2\cosh\left(\tilde{h}_i + W_i\right)\right) \ \ldots \\ &\ldots - \sum_i W_i m_i + \sum_{i<j} \tilde{J}_{ij} m_i m_j,\end{aligned} \tag{11}$$

with

$$W_i = \sum_{j\neq i} \tilde{J}_{ij} m_j.$$

Here each of the parameters is actually a function of the stimulus s, which we omit for clarity.

For the TAP approach the corresponding equation becomes:

$$\begin{aligned}&\log \mathcal{Z} = \sum_i \log\left(2\cosh\left(\tilde{h}_i + L_i\right)\right) - \sum_i L_i m_i \ \ldots \\ &\ldots + \frac{1}{2}\sum_{i,j} \tilde{J}_{ij} m_i m_j + \frac{1}{4}\sum_{i,j} \tilde{J}_{ij}^2 (1-m_i^2)(1-m_j^2),\end{aligned} \tag{12}$$

with

$$L_i = \sum_{j\neq i} \tilde{J}_{ij} m_j - m_i \sum_{j\neq i} \tilde{J}_{ij}^2 (1-m_j^2).$$

3 Results

We performed computer simulations as described in Methods, to generate datasets for training and testing decoding algorithms. 10-20 simulations were performed with different random number seeds, in order to characterize decoding performance. A number of metrics, including the fraction of correct decodings (accuracy) and mutual information between decoded and presented stimulus distributions, were used in order to characterize and compare decoding performance.

Homogeneous Scenario

The Ising model based decoders show better performance than the independent decoder in nearly all cases (Fig. 4a), in terms of the average fraction of correctly decoded stimuli. The performance of the standard (non-regularized) MPFL technique however falls away relatively quickly as the number of cells increases, failing to better the independent decoder after approximately 100 cells. This behaviour can be explained by considering that the problem of parameter estimation becomes more and more underconstraint as the number of cells increases while holding the number of training samples fixed. Falsely learned model parameters moreover affect the decoding performance by influencing the estimated partition function and thus worsen the decoder performance even further. As we have only estimated the partition function once per stimulus when using MPFL, large fluctuations in the training dataset can potentially have a big effect. To compensate for this behaviour a regularization term can be included, which can stabilize the performance up to a significantly larger number of neurons (as described in Methods).

As the fraction correct or accuracy does not provide a complete description of the decoder, we computed the mutual information between the encoded and decoded stimulus (Panzeri, Treves, Schultz & Rolls 1999) to characterise the performance further. This provides a compact description of the decoding confusion matrix (which gives a full description of decoding behaviour).

We can write the mutual information (measured in bits) as:

$$I(s,\hat{s}) = H(s) - H(s|\hat{s}), \tag{13}$$

where $H(s)$ is the entropy of the encoded stimulus

$$H(s) = -\sum_{s\in\mathcal{S}} p(s)\log_2 p(s), \tag{14}$$

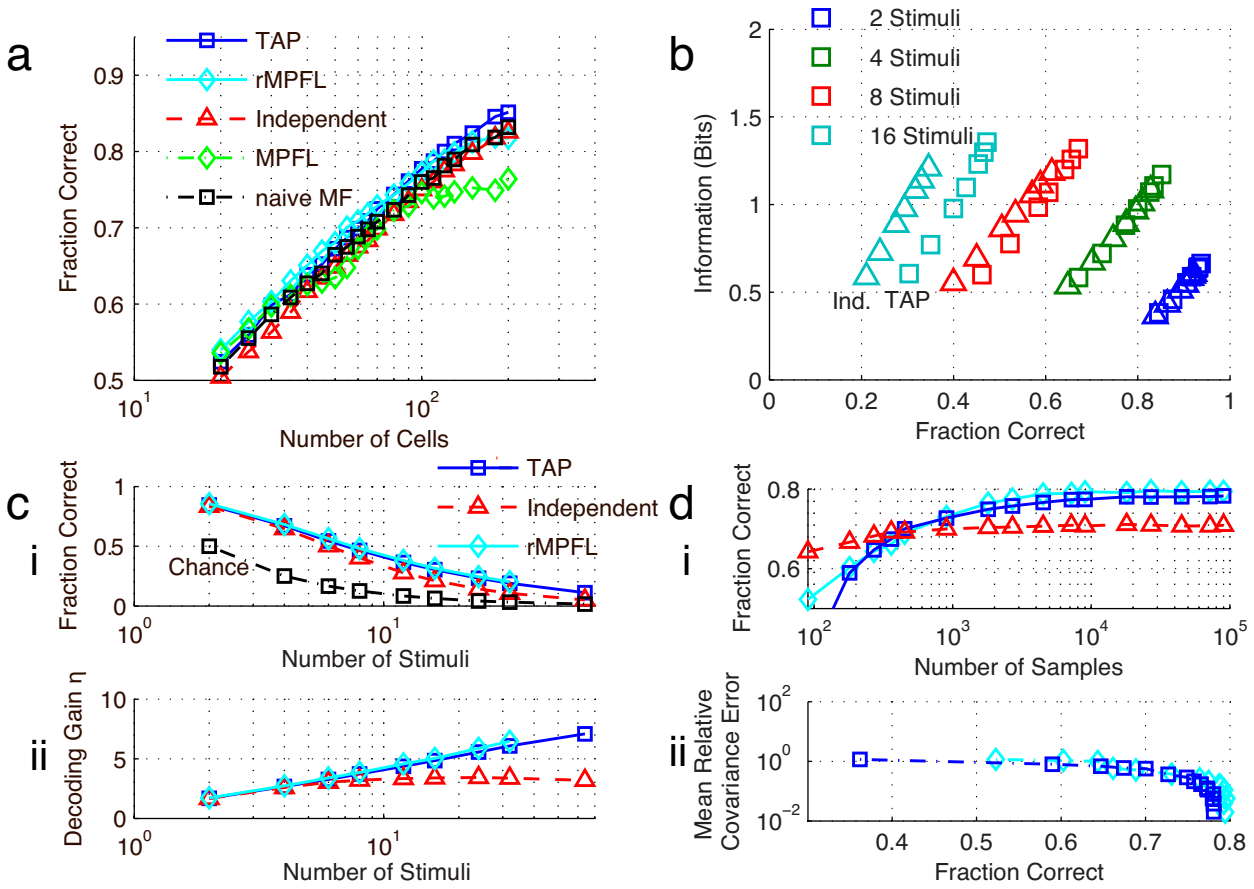

Fig. 4: Performance of decoding algorithms in the homogeneous scenario (Model A). **a** Fraction of correct decodings versus neural ensemble size, for a training dataset size of 900 samples per stimulus. **b** The relationship between fraction correct and mutual information $I(s, \hat{s})$ for varying stimulus set and ensemble size ($C = \{50, 70, 100, 120, 150, 170, 200\}$ varying from bottom left to top right for each symbol type). Triangles denote the performance of the Independent decoder, squares the TAP Ising Decoder. Average of 20 simulations. **c** The dependence of decoding performance on stimulus set size. **i** TAP, Independent and rMPFL decoders compared to random selection of stimuli. This is replotted in **ii** as the gain in fraction correct over chance performance, $\eta = p_{\text{dec}}/p_{\text{guess}}$, making the performance saturation for the independent decoder as problem difficulty increases more apparent. **d** Dependence of decoder performance on test trials on training set sample size, for 4 stimuli and a neural ensemble size of 70. **i** Fraction correct as a function of number of training samples. Below 450 training samples the Ising decoder fails to better the independent decoder. **ii** Relation between mean relative covariance error E as a measure of finite sampling effects and fraction correct.

and $H(s|\hat{s})$ is the conditional entropy describing the distribution of stimuli s that have been observed to elicit each decoded state $\hat{s}$,

$$H(s|\hat{s}) = -\sum_{s,\hat{s}\in\mathcal{S}} p(s,\hat{s}) \log_2 p(s|\hat{s}). \tag{15}$$

Since we have in the current study opted for a uniform stimulus distribution, the entropy $H(s)$ is simply given by

$$H(s) = \log_2 S. \tag{16}$$

In general the conditional entropy $H(s|\hat{s})$ has to be computed from the confusion matrix. We note that if we were to assume that the correctly decoded stimuli and errors are uniformly distributed for all stimuli, i.e. that the conditional distribution $p(\hat{s}|s)$ is of the form

$$p(\hat{s}|s) = \begin{cases} f_c & \text{for } \hat{s} = s \\ \dfrac{1-f_c}{S-1} & \text{for } \hat{s} \neq s, \end{cases}$$

then the conditional entropy simplifies to

$$H(s|\hat{s}) = -f_c \log_2 f_c - (1-f_c) \log_2 \left(\frac{1-f_c}{S-1}\right).$$

This simplified expression has been used to characterize decoder performance in the Brain Computer Interface literature (Wolpaw et al. 2002). Here, we present this equation only to make apparent the scaling behaviour, and compute the decoded information using the more general expression.

The decoded information analysis reveals that the difference in the decoding performance of the independent and Ising (as exemplified by TAP) models becomes more pronounced as the number of stimuli is increased, as shown in Fig. 4b,c. As the number of stimuli increase the independent and Ising decoder curves separate, indicating not only a difference in the accuracy (fraction correct) of both decoders but also a difference in the confusion matrices, i.e. in the distribution of errors between the two approaches. This is considered in more detail in section 3.1. A further interesting behaviour of the Ising decoder is apparent in Fig. 4c: as the number of stimuli increases, the relative decoding gain η, (hich we define as the ratio between the "actual" and "chance" fraction correct) keeps increasing with the number of stimuli for the Ising model case, whereas it saturates for the independent decoder. This suggests that the Ising decoder may be particularly advantages as the decoding problem becomes more difficult.

Performance of the Ising decoder is strongly dependent on the number of training trials available (Fig. 4d i). Here we found, for 70 neurons, that around 400 training samples were required to allow the Ising decoder to outperform independent decoding. The independent decoder will necessarily have better sampling performance, as it relies only upon lower order response statistics. (It is worth considering that a "full" decoder which made use of all aspects of spike pattern structure would of course have far worse scaling behaviour than either). Another way of looking at this is to examine the estimated covariance matrix from the simulated data (Fig. 4d ii). We define the mean relative error of the covariance matrix as

$$E = \frac{1}{S}\sum_{s\in\mathcal{S}}\frac{1}{C^2}\sum_{i,j}\left|\frac{C^{\text{data}}_{ij,s} - C^{\text{as}}_{ij,s}}{C^{\text{as}}_{ij,s}}\right|, \qquad (17)$$

where $C^{\text{data}}_{ij,s}$ is the estimated covariance between unit i and j under stimulus s from the (finite) simulated data (normally 1000 samples), and $C^{\text{as}}_{ij,s}$ is the asymptotic covariance, defined in the same way but computed with 100,000 samples. This error provides a measure of the finite sampling bias we encounter for fitting the model.

With 1000 samples per stimulus, the mean relative error between the estimated and the "true", asymptotic covariance is comparably large. However we have used this sample size in this study where not otherwise shown as it is probably the maximum number of samples that can realistically be obtained in many types of neuophysiological experiment. The error in estimating the covariance matrix reduces the advantage of the Ising decoder approach as the model is fitted to a set of measured (biased) correlations, which do not necessarily correspond to the correlations in the real distribution of the spike patterns. With increasing number of samples and reduced covariance error the difference between the Ising model decoder and independent decoder increases. This analysis shows that in many neurophysiological experiments or Brain-Machine Interfacing applications, we may need to operate close to the minimal feasible sample size for the Ising decoder.

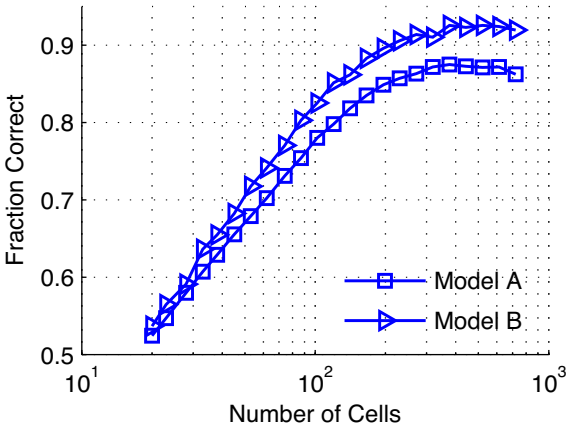

Fig. 5: Decoder performance is enhanced in the heterogeneous scenario (Model B). Comparison of the fraction of correct decodings vs. number of cells curves for Models A and B. The TAP algorithm was used to train the decoder.

Heterogeneous Scenario

The performance characteristics for the heterogeneous scenario (Model B) are fairly similar to the homogeneous case, so we report here only on the observed differences. The overall classification performance both in terms of fraction correct and mutual information, is slightly better for both decoder types with heterogeneous neural ensembles. As an example the fraction correct as a function of cells for both models is compared in Fig. 5 for the TAP approach. The better performance for Model B can be explained by the greater variability of cell properties allowing more specific response patterns than in the homogeneous scenario. This scenario is presumably more relevant to many real-world decoding problems.

However, the increased variability also raises the likelihood of encountering neurons in our simulations that exhibit a very low spiking probability. This makes it harder to estimate the correct means and correlations from the finite training data set and therefore harder to fit the right model parameters. As a result the benefits of the Ising model approach compared to the independent decoder reduce for larger neural ensembles ($\approx$ 100 cells, not shown). The MPFL approach even breaks down after an ensemble size of 100 cells, which could be explained by a combination of the discussed finite sampling effects, the under-constrained learning problem, and once per stimulus computed partition function, all of which degrade the performance of the decoder.

It is also apparent from Fig. 5 that for both the homogeneous and heterogeneous simulations, it was practical to decode spatial patterns of activity comprised of over 700 neurons with the TAP method. The time to train these decoders was relatively modest, even on a desktop PC several years old at the time of writing. The limitation upon the size of neural ensemble that could be decoded was largely due to the number of trials available for training - 900 in the present situation. It can be seen that performance levels off in Fig. 5 as the number of neurons approaches that order, suggesting that as a rule of thumb, the number of training trials should be at a minimum several time the number of neurons in the ensemble.

3.1 Confusion Matrix Analysis

The confusion matrix provides complete information about the decoding error distribution. Confusion matrices for the Ising decoder and the Independent decoder respectively are shown in Fig. 6. The Ising decoder has higher diagonal terms which corresponds to the better accuracy (fraction correct) of the decoder. The overall appearance of the Ising decoder confusion matrix is fairly similar to the independent decoder. However, by comparing the two confusion matrices it can be seen that the Ising decoder mainly gains its performance benefit over the independent decoder by avoiding confusion of adjacent stimuli. This shows that as the difference in adjacent stimulus directions becomes less with an increasing number of stimuli, the Ising model decoder can utilize the correlation patterns to enhance the decoding accuracy, i.e. to distinguish between adjacent stimulus directions more precisely. This effect of course depends on the correlation model used - here the correlation between each pair of neurons was resampled for each stimulus in the simulation. If correlations were not at all stimulus-dependent, or if the model was quite different, then the Ising decoder may not be able to take advantage of this potential performance advantage.

3.2 Dependence on level of correlation

As the advantage of the Ising Decoder over the Independent Decoder stems from its ability to take advantage of information contained in pairwise correlations, we examined the dependence of this advantage on the average strength of correlation. Although we have set the average level of correlation to what has traditionally been thought to be a reasonable level for cells in the same vicinity (Zohary & Shadlen 1994), there is ongoing debate about the level and stimulus dependence of correlation relevant for cortical function (Renart et al. 2010, Bair et al. 2001, Nase et al. 2003).

While our ability to manipulate the average level and spread of the distribution of correlation is limited by fundamental constraints upon the covariance of a binary distribution (Macke et al. 2009), it was possible to adjust the average level and spread of correlation (Fig. 7), although not independently. It can be seen that as the level of correlation increases for specified spike count, the independent decoder loses some discriminative capacity. The Ising decoders however can take advantage of the higher level and spread of correlation values for discrimination between stimuli, and in particular regularized MPFL continues to improve as the level of correlation is increased. As the level of correlation is decreased below the level (0.11) used in the rest of this study, the Independent decoder behaves as expected, approximately converging in performance with the Ising decoder. Around zero correlation level the Ising decoder however does something that is at first glance unexpected: its performance actually increases. This increase is explained by the fact that zero average correlation does not actually mean zero correlation: in fact, there is a spread (standard deviation ~ 0.04) around zero, such that any pair of cells are equally likely to be anticorrelated as correlated. The Ising decoder is able to take advantage of any stimulus dependence in the correlation matrix due to such effects.

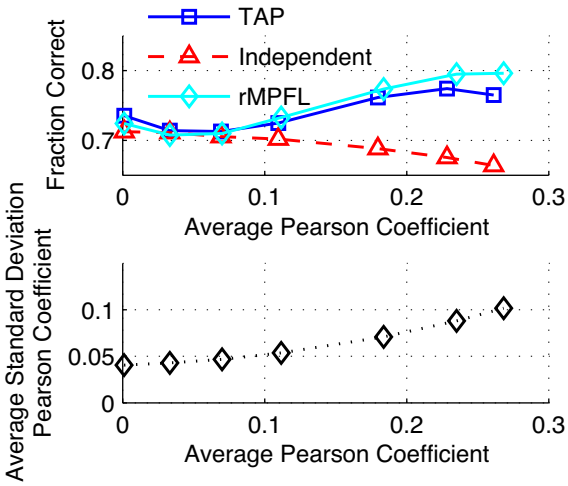

Fig. 7: MPFL learning benefits from a high correlation regime. Upper panel: Fraction correct as a function of the average Pearson correlation coefficient for three decoding algorithms. Lower panel: Average Standard Deviation of the Pearson Correlation coefficient vs. average Pearson correlation coefficient. All averages taken over 20 simulations; Model A; correlations measured with 1000 samples.

3.3 Comparison with linear decoding

While most work in the Brain-Machine Interface literature has focused on continuous decoders, there has been some work on the discrete decoding problem, although to date with a focus on the analysis of either continuous data, such

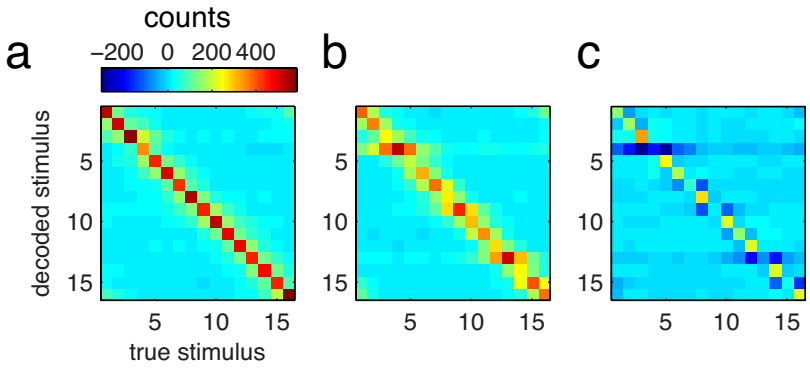

Fig. 6: Decoder confusion matrices. The case illustrated is for 16 Stimuli, 200-neuron ensembles, under the homogeneous scenario (Model A). Counts of the number of times each stimulus is decoded, for 1000 stimulus presentations. **a** Ising model decoder. **b** Independent decoder. **c** A difference plot between **a** and **b**.

as electroencephalographic (EEG) data (Wolpaw et al. 2002), or on longer time windows of multi-electrode array data, in which spike counts are far from binary (Santhanam et al. 2009). As the discrete decoding problem can be viewed as a classification problem (in the same sense as the continuous decoding problem can be seen as a regression problem), it is of interest to compare the performance of our approach with traditional classification approaches such as the Optimal Linear Classifer (OLC).

Following Bishop (2007), each stimulus class can be described by its own linear model, so that

$$y_k = \mathbf{w}_s^T \boldsymbol{\sigma} + w_{k_0}, \tag{18}$$

where $s = 1, \ldots, S$. Using a 1-of-S binary coding scheme (i.e. we denote stimulus class s_i by a "target" column vector $\mathbf{t}$ with all zeros except the i-th entry, which is one) the weights $\mathbf{w}_s$ can be trained such as to minimize a sum of squares error function for the target stimulus vector.

This is done in Fig. 8. It can be seen that, under the conditions we test here (1000 trials, simulated data as described previously for Model A), the OLC underperforms both the TAP and Independent classifiers. The former is not unexpected, but it may seem initially counter-intuitive that the OLC does not yield identical performance to the Independent decoder, as the latter is in effect performing a linear classification. However, several differences must be noted that these two algorithms. Firstly, while neither algorithm takes into account correlation, their performance can be differently affected by it. Secondly, as their implementation details differ, they may have markedly different sensitivity to limited sampling - the independent decoder, as we have constructed it here (product of marginals) has remarkable sampling efficiency. Finally the implementation of the OLC decoder assumes a Gaussian error in the stimulus class target vectors, which is clearly not valid for the here considered binary class vectors.

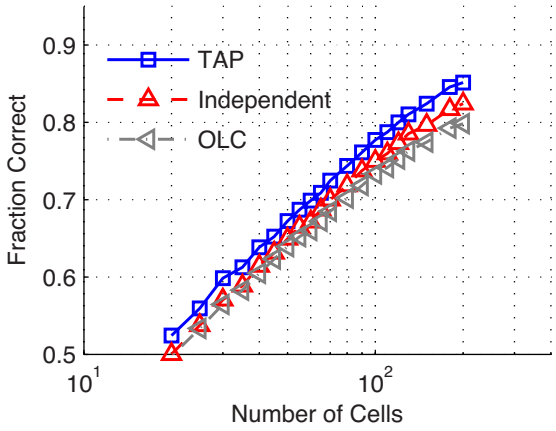

Fig. 8: Comparison of the TAP Ising Decoder and Independent Decoder with the Optimal Linear Classifier (OLC).

4 Discussion

We have demonstrated, for the first time, the use of the Ising model to decode discrete stimulus states from (simulated) large-scale neural population activity. To do this, we have had to overcome several technical obstacles, namely the poor scaling properties of previously used algorithms for learning model parameters, and similarly the poor scaling behaviour of methods for estimating the partition function, which although not necessary for some applications of the Ising model in neuroscience, is required for decoding. The Ising model has one particular advantage over a simpler independent decoding algorithm: that it can take advantage of stimulus dependence in the correlation structure of neuronal responses, where it exists. With the aid of a statistical simulation of neuronal ensemble spiking responses in the mouse visual cortex, we have demonstrated that correlational information can be taken advantage of for decoding the activity of neuronal ensembles of size in the hundreds by several algorithms, in-

cluding the naive Mean Field algorithm, the TAP algorithm, and regularized MPFL. In particular, we have shown that the TAP algorithm can be used to decode spatial patterns of activity comprised of over 700 neurons.

Ising models have gained much attraction recently in neuroscience to describe the spike train statistics of neural ensembles (Schneidman et al. 2006, Shlens et al. 2006, Shlens et al. 2009). However, these findings have largely been made only in relatively small neural ensembles (at most a few tens of cells), from which an extrapolation to larger ensemble sizes might not be wise (Roudi, Nirenberg & Latham 2009, Roudi, Aurell & Hertz 2009). The principal reason for this limit has been the poor scaling of the computational load of fitting the Ising model parameters, when algorithms such as Boltzmann learning are used. Moreover, new findings suggests that pairwise correlations (and thus Ising models) might not be sufficient to predict spike patterns of small scale local clusters of neurons ($< 300\mu m$ apart), which have been observed to provide evidence of higher order interactions (Ohiorhenuan et al. 2010). While the formalism for higher order models may be similar, scaling properties are guaranteed to be even worse. There is thus the pressing need to develop for better algorithms for learning the parameters of Ising and Ising-like models.

The development of discrete neural population decoding algorithms has two motivations. The first motivation is the desire to develop brain-computer communication devices for cognitively intact patients with severe motor disabilities (Mak & Wolpaw 2009). In this type of application, an algorithm such as those we describe could be used together with multi-electrode brain recordings to allow the user to select one of a number of options (for instance a letter from a virtual keyboard), or even in the longer term to communicate sequences of symbols from an optimised code directly into a computer system or communications protocol. Given the short timescale to which the Ising decoder can be applied (we have fixed this at 20 ms here), and sufficiently large recorded ensembles to saturate decoder performance, very high bit rates could potentially be achieved.

The second motivation is more scientific: to use such decoding algorithms to probe the organisation and mechanisms of information processing in neural systems. It should be immediately be apparent that the Ising decoder and related models can be used to ask questions about the neural representation of sensory stimuli, motor states or other behavioural correlates, by comparing decoding performance under different sets of assumptions (for instance, by changing the constraints in an Ising model to exclude correlations, include correlations within 50 μm, etc). This (commonly referred to as the "encoding problem") is essentially the same use to which Shannon information theory has been applied in neuroscience (see e.g. Schultz et al. (2009) for a recent example), with simply a different summary measure. Use of decoding performance may be an intuitively convenient way to ask such questions, but it is still asking exactly the same question. However, there are other uses to which such algorithms can be applied. For instance, combining sensory and learning/memory experimental paradigms, once a decoder has been trained, it could be used subsequently to read out activity patterns in different brain states such as sleep, or following some period of time - for instance, to "read out memories" by decoding the patterns of activity that represent them. The decoding approach may thus have much to offer the study of information processing in neural circuits.

Our results show that decoding performance is critically dependent on the sample size used for training the decoder, as the knowledge of the exact pairwise correlations is essential to fit a model that correctly matches to the statistical structure of unobserved data. In the "encoding problem", such finite sampling constraints result in a biased estimation of the entropy of the system. For the decoding problem considered here, finite sampling leads to an overfitting of the model to the observed training data, such that it does not generalize to the unobserved data, and accordingly fails to predict stimulus classes correctly during test trials. Such finite sampling constraints mean that below a particular sampling size - which we found to be 400 trials for 70 neurons in one particular example we studied - there is no point in using a model which attempts to fit pairwise (or above) correlations, one may as well just use an independent model. This has implications for experimental design. However, it should be noted that the real brain has no such limitation - in effect, many thousands or millions of trials are available over development, and so a biological system should certainly be capable of learning the correct correlations from the data (Bi & Poo 2001) and thus may well be able to operate in a regime where decoding benefits more clearly from known correlation structure.

We have additionally shown that incorporating correlations in the decoding process might be especially relevant for 'hard' decoding problems, i.e. multi-class discrimination problems in which stimuli are not easily distinguishable by just observing individual neuronal firing rates. In this scenario including correlations could be a means to enhance the precision of the decoding process by increasing the discriminability between adjacent or similar stimuli. Including correlations could potentially made the pattern distribution more flat, or uniform, with low firing rates, leading to greater energy efficiency of information coding (Baddeley et al. 1997).

One caveat to the advantage provided by correlations of using Ising over Independent decoders is that it depends entirely upon the extent to which correlations are found to depend upon the stimulus variables of interest. While a previous study at longer timescales has found correlations to improve neural decoding (Chen et al. 2006), the jury is still out

on the prevalence of stimulus dependence of pairwise and higher order correlations in the mammalian cortex. Stimulus-dependent correlations have been found in both mouse (Nase et al. 2003) and monkey (Kohn & Smith 2005, Samonds & Bonds 2005, Kohn et al. 2009), where they have been shown to contribute to information encoding (Montani et al. 2007), but most recordings to date have sampled relatively sparsely from the local cortical circuit, due to limitations in multi-electrode array hardware. It is possible that if one were to be able to record from a greater proportion of neurons in the local circuit, then stronger stimulus-dependent correlations might be observable.

Several recommendations can be made on the basis of our study of neural decoding. The first is that if a relatively modest number of trials is available, then no advantage is gained by attempting to fit a more complex model (the actual number will depend upon the ensemble size as well as response statistics). In such as situation, the physiologist should assume independence on the grounds of insufficient data to do anything else. The second is that, if there is sufficient data to train a correlational decoder, there still remains a choice of what type of decoder to train. For very small ensembles, any algorithm will do, For strong correlations, or moderate sized ensembles (< 100 cells), the MPFL algorithm performed well. However, until a good approach can be found for efficient partition function estimation to couple with the MPFL training algorithm, computational constraints will forbid the MPFL approach to be used with excessively large ensembles. Finally, if there is the desire both to consider correlations and to decode with very large neural populations, the TAP approach would appear to demonstrate clear advantages over the other approaches considered here.

A number of avenues present themselves for future development of decoding algorithms. Firstly, algorithms for reducing model dimensionality without losing discriminatory power, may prove advantageous. These may include graph and hypergraph theoretic techniques (Aghagolzadeh et al. 2010) for pruning out uninformative dimensions (edges and nodes), and factor analysis methods for modeling conditional dependencies (Santhanam et al. 2009). Such an approach may be particularly advantageous when experimental trials are limited, as the dimensionality of the parameter set is the main reason for Ising model performance not exceeding the Independent model for limited trials. One difficulty with the use of graph pruning approaches is that the usual pairwise correlation matrix of neural recordings, unlike the graph in many network analysis problems, tends not to be sparse. It is of course a functional, as opposed to synaptic, connectivity matrix, and one reason for this lack of sparseness is its symmetric nature. It has recently been proposed that the symmetry property of the J_{ij} matrix can be relaxed in the context of the (non-equilibrium) Kinetic Ising model, which also provides a convenient way to take into account space-time dependencies, or causal relationships (Hertz et al. 2010). Use of the Kinetic Ising model framework for decoding would appear to be an interesting future direction to pursue.

While new experimental technologies are yielding increasingly high dimensional multivariate neurophysiological datasets, usually without concomitant increases in the duration of data that can be collected, we conclude that there is some reason for optimism that we will be able to develop new data analysis methods capable of taking advantage of this data. Maximum entropy approaches to the fitting of structured parametric models such as the Ising model and its extensions would appear to be one approach likely to yield progress.

Acknowledgements

We thank Phil Bream, Hélène Seiler and Yang Zhang for their contributions to earlier work leading up to that reported here, and Aman Saleem for useful discussions. We also thank Yasser Roudi for useful comments on the TAP equations, and Jascha Sohl-Dickstein, Peter Battaglino, and Mike DeWeese for useful MATLAB code and discussion of the MPFL technique. This work was supported by EPSRC grant EP/E002331/1 to SRS.

References

Aghagolzadeh, M., Eldawlatly, S. & Oweiss, K. (2010), 'Synergistic coding by cortical neural ensembles.', *IEEE Trans Inf Theory* **56**(2), 875–899.

Baddeley, R., Abbott, L. F., Booth, M. C., Sengpiel, F., Freeman, T., Wakeman, E. A. & Rolls, E. T. (1997), 'Responses of neurons in primary and inferior temporal visual cortices to natural scenes.', *Proc Biol Sci* **264**(1389), 1775–83.

Bair, W. & Koch, C. (1996), 'Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey.', *Neural Comput* **8**(6), 1185–202.

Bair, W., Zohary, E. & Newsome, W. T. (2001), 'Correlated firing in macaque visual area mt: time scales and relationship to behavior.', *J Neurosci* **21**(5), 1676–97.

Berger, A. L., Pietra, V. J. D. & Pietra, S. A. D. (1996), 'A maximum entropy approach to natural language processing', *Comput. Linguist.* **22**(1), 39–71.

Bi, G.-q. & Poo, M.-m. (2001), 'Synaptic modification by correlated activity: Hebb's postulate revisited', *Annual Review of Neuroscience* **24**(1), 139–166.

Bishop, C. M. (2007), *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. 2006. corr. 2nd printing edn, Springer.

Butts, D. A., Weng, C., Jin, J., Yeh, C.-I. I., Lesica, N. A., Alonso, J.-M. M. & Stanley, G. B. (2007), 'Temporal precision in the neural code and the timescales of natural vision.', *Nature* **449**(7158), 92–5.

Carr, C. E. (1993), 'Processing of temporal information in the brain', *Annu Rev Neurosci* **16**, 223–43.

Chen, Y., Geisler, W. S. & Seidemann, E. (2006), 'Optimal decoding of correlated neural population responses in the primate visual cortex.', *Nat Neurosci* **9**(11), 1412–20.

Földiák, P. (1993), The 'ideal homunculus': statistical inference from neural population responses, *in* 'Computation and Neural Systems', Kluwer Academic Publishers, Norwell, MA, pp. 55–60.

Gawne, T. J., Kjaer, T. W., Hertz, J. A. & Richmond, B. J. (1996), 'Adjacent visual cortical complex cells share about 20% of their stimulus-related information.', *Cereb Cortex* **6**(3), 482–9.

Hertz, J., Roudi, Y., Thorning, A., Tyrcha, J., Aurell, E. & Zeng, H. L. (2010), 'Inferring network connectivity using kinetic ising models', *BMC Neuroscience* **11**(Suppl 1), P51.

Higham, N. J. (2002), 'Computing the nearest correlation matrix–a problem from finance', *IMA Journal of Numerical Analysis* **22**(3), 329.

Huang, F. & Ogata, Y. (2001), 'Comparison of two methods for calculating the partition functions of various spatial statistical models', *Australian & New Zealand Journal of Statistics* **43**(1), 47–65.

Jaynes, E. T. (1957), 'Information theory and statistical mechanics', *Phys. Rev.* **106**(4), 620–630.

Kappen, H. J. & Rodríguez, F. B. (1998), 'Efficient learning in boltzmann machines using linear response theory', *Neural Computation* **10**(5), 1137–1156.

Kohn, A. & Smith, M. A. (2005), 'Stimulus dependence of neuronal correlation in primary visual cortex of the macaque.', *J Neurosci* **25**(14), 3661–73.

Kohn, A., Zandvakili, A. & Smith, M. A. (2009), 'Correlations and brain states: from electrophysiology to functional imaging.', *Curr Opin Neurobiol* **19**(4), 434–8.

Macke, J. H., Berens, P., Ecker, A. S., Tolias, A. S. & Bethge, M. (2009), 'Generating spike trains with specified correlation coefficients', *Neural Computation* **21**(2), 397–423.

Mak, J. N. & Wolpaw, J. R. (2009), 'Clinical applications of brain-computer interfaces: Current state and future prospects.', *IEEE Rev Biomed Eng* **2**, 187–199.

Montani, F., Kohn, A., Smith, M. A. & Schultz, S. R. (2007), 'The Role of Correlations in Direction and Contrast Coding in the Primary Visual Cortex', *J. Neurosci.* **27**(9), 2338–2348.

Nase, G., Singer, W., Monyer, H. & Engel, A. K. (2003), 'Features of Neuronal Synchrony in Mouse Visual Cortex', *J Neurophysiol* **90**(2), 1115–1123.

Niell, C. M. & Stryker, M. P. (2008), 'Highly Selective Receptive Fields in Mouse Visual Cortex', *J. Neurosci.* **28**(30), 7520–7536.

Ogata, Y. & Tanemura, M. (1984), 'Likelihood analysis of spatial point patterns', *Journal of the royal statistical Society. Series B.* **46**(3), 496–518.

Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q. & Victor, J. D. (2010), 'Sparse coding and high-order correlations in fine-scale cortical networks', *Nature* **466**(7306), 617–621.

Oram, M. W., Földiák, P., Perrett, D. I., Oram, M. W. & Sengpiel, F. (1998), 'The 'ideal homunculus': decoding neural population signals', *Trends in Neurosciences* **21**(6), 259 – 265.

Panzeri, S. & Schultz, S. R. (2001), 'A unified approach to the study of temporal, correlational, and rate coding.', *Neural Comput* **13**(6), 1311–49.

Panzeri, S., Schultz, S. R., Treves, A. & Rolls, E. T. (1999), 'Correlations and the encoding of information in the nervous system', *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**(1423), 1001–1012.

Panzeri, S., Treves, A., Schultz, S. & Rolls, E. T. (1999), 'On decoding the responses of a population of neurons from short time windows.', *Neural Comput* **11**(7), 1553–77.

Petersen, R. S., Panzeri, S. & Diamond, M. E. (2001), 'Population coding of stimulus location in rat somatosensory cortex', *Neuron* **32**(3), 503 – 514.

Pola, G., Thiele, A., Hoffmann, K. & Panzeri, S. (2003), 'An exact method to quantify the information transmitted by different mechanisms of correlational coding', *Network-Computation in Neural Systems* **14**(1), 35–60.

Reich, D. S., Mechler, F. & Victor, J. D. (2001), 'Independent and Redundant Information in Nearby Cortical Neurons', *Science* **294**(5551), 2566–2568.

Renart, A., de la Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A. & Harris, K. D. (2010), 'The asynchronous state in cortical circuits.', *Science* **327**(5965), 587–90.

Roudi, Y., Aurell, E. & Hertz, J. A. (2009), 'Statistical physics of pairwise probability models', *Frontiers in Computational Neuroscience* **3:22**, 1–15.

Roudi, Y., Nirenberg, S. & Latham, P. E. (2009), 'Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't', *PLoS Comput Biol* **5**(5), e1000380.

Roudi, Y., Tyrcha, J. & Hertz, J. (2009), 'Ising model for neural data: Model quality and approximate methods for extracting functional connectivity', *Phys. Rev. E* **79**(5), 051915 (12 pages).

Samonds, J. M. & Bonds, A. B. (2005), 'Gamma oscillation maintains stimulus structure-dependent synchronization in cat visual cortex.', *J Neurophysiol* **93**(1), 223–36.

Santhanam, G., Yu, B. M., Gilja, V., Ryu, S. I., Afshar, A., Sahani, M. & Shenoy, K. V. (2009), 'Factor-analysis methods for higher-performance neural prostheses.', *J Neurophysiol* **102**(2), 1315–30.

Schneidman, E., Berry II, M. J., Segev, R. & Bialek, W. (2006), 'Weak pairwise correlations imply strongly correlated network states in a neural population', *Nature* **440**(7087), 1007–1012.

Schultz, S. R., Kitamura, K., Post-Uiterweer, A., Krupic, J. & Hausser, M. (2009), 'Spatial Pattern Coding of Sensory Information by Climbing Fiber-Evoked Calcium Signals in Networks of Neighboring Cerebellar Purkinje Cells', *J. Neurosci.* **29**(25), 8005–8015.

Schultz, S. R. & Panzeri, S. (2001), 'Temporal correlations and neural spike train entropy.', *Phys Rev Lett* **86**(25), 5823–6.

Seiler, H., Zhang, Y., Saleem, A., Bream, P., Apergis-Schoute, J. & Schultz, S. R. (2009), 'Maximum entropy decoding of multivariate neural spike trains', *BMC Neuroscience* **10**(Suppl 1), P107.

Shlens, J., Field, G. D., Gauthier, J. L., Greschner, M., Sher, A., Litke, A. M. & Chichilnisky, E. J. (2009), 'The Structure of Large-Scale Synchronized Firing in Primate Retina', *J. Neurosci.* **29**(15), 5022–5031.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M. & Chichilnisky, E. J. (2006), 'The Structure of Multi-Neuron Firing Patterns in Primate Retina', *J. Neurosci.* **26**(32), 8254–8266.

Sohl-Dickstein, J., Battaglino, P. & DeWeese, M. R. (2009), 'Minimum Probability Flow Learning', http://arxiv.org/abs/0906.4779.

Tanaka, T. (1998), 'Mean-field theory of boltzmann machine learning', *Phys. Rev. E* **58**(2), 2302–2310.

Thouless, D. J., Anderson, P. W. & Palmer, R. G. (1977), 'Solution of solvable model of a spin glass', *Philos Mag* **35**(3), 593–601.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. (2002), 'Brain-computer interfaces for communication and control.', *Clin Neurophysiol* **113**(6), 767–91.

Zohary, E. & Shadlen, M. (1994), 'Correlated neuronal discharge rate and its implications for psychophysical performance', *Nature* **370**(6485), 140–143.