How fast can we learn maximum entropy models of neural populations?

2009 J. Phys.: Conf. Ser. 197 012020

(http://iopscience.iop.org/1742-6596/197/1/012020)

International Workshop on Statistical-Mechanical Informatics 2009 (IW-SMI 2009) IOP Publishing
Journal of Physics: Conference Series **197** (2009) 012020 doi:10.1088/1742-6596/197/1/012020

How fast can we learn maximum entropy models of neural populations?

Elad Ganmor[1], **Ronen Segev**[2] **and Elad Schneidman**[1]

[1] Department of Neuroscience, Weizmann Institute of Science, Rehovot 76100, Israel
[2] Department of Life Sciences & Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Email: {elad.ganmor, elad.schneidman}@weizmann.ac.il

Abstract. Most of our knowledge about how the brain encodes information comes from recordings of single neurons. However, computations in the brain are carried out by large groups of neurons. Modelling the joint activity of many interacting elements is computationally hard because of the large number of possible activity patterns and limited experimental data. Recently it was shown in several different neural systems that maximum entropy pairwise models, which rely only on firing rates and pairwise correlations of neurons, are excellent models for the distribution of activity patterns of neural populations, and in particular, their responses to natural stimuli. Using simultaneous recordings of large groups of neurons in the vertebrate retina responding to naturalistic stimuli, we show here that the relevant statistics required for finding the pairwise model can be accurately estimated within seconds. Furthermore, while higher order statistics may, in theory, improve model accuracy, they are, in practice, harmful for times of up to 20 minutes due to sampling noise. Finally, we demonstrate that trading accuracy for entropy may actually improve model performance when data is limited, and suggest an optimization method that automatically adjusts model constraints in order to achieve good performance.

1. Introduction

Sensory information is encoded in the brain by the joint activity of groups of neurons that emit stereotyped electric pulses, termed spikes, as their output [1, 2]. Importantly, when sensory systems are presented repeatedly with identical stimuli, responses vary over presentations [3, 4]. This led many researchers to use a probabilistic approach to describe and study neural coding [5, 6]. Even if we simplify neural responses into a sequence of spikes (1's) and silence (-1's), we are still faced with a potentially exponential number of activity patterns.

Clearly it is not feasible to directly sample the joint distribution of large groups of neurons. Accordingly, many studies of neural populations have either focused on small sets of neurons, or tried to base models on statistics that can be accurately estimated from limited data. Recently it was shown, in several different neural systems, that for groups of ~10 neurons the maximum entropy model which takes into account only firing rates and pairwise correlation, out of the exponential number of possible correlation functions among neurons, provides an extremely accurate approximation of neural activity pattern distribution [7-9].

Correlation based models have a natural hierarchy, according to the order of correlation they rely on. This allows us to quantify the contribution of different orders of correlation to the total correlation in the network [10, 11]. The pairwise model is one member of this correlation based hierarchy, but

other more complex models which take into account higher order correlations may also be considered. As we move up in the hierarchy, adding higher orders of correlations, the resulting models are guaranteed to give more accurate results. However, building such complex models becomes more challenging – from computational and statistical standpoints. Here we examine the sampling properties of such models and the empirical statistics they rely on. Our results indicate that the pairwise model can be accurately learned within seconds, and furthermore, taking into account higher order correlations proves harmful due to sampling noise.

2. Maximum entropy based hierarchy of models of correlated neural population activity

To investigate the responses of groups of neurons we recorded the simultaneous activity of dozens of retinal ganglion cells in the isolated vertebrate retina presented with natural visual stimuli [12]. We discretize time into 20 ms bins and represent the state of the network at any moment by an *n*-bit binary word $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ (*n* being the number of neurons), where each bit corresponds to the firing of a single neuron.

Since pairs of neurons are typically weakly correlated [7, 13], it is tempting to assume that larger neural populations can be well described by an independent model, denoted $P^{(1)}$, in which the probability of a state $\{\sigma_i\}$ is given by $P^{(1)}(\{\sigma_i\}) = \Pi_{i=1}^{n} P(\sigma_i)$. $P^{(1)}$ is the minimal model that takes into account only the firing rates of individual neurons. It allows for efficient learning and inference, while avoiding the exponential complexity of the full joint distribution. However, small networks of ~10 neurons already display significantly correlated activity [7], and so $P^{(1)}$ makes large errors in describing the activity of even small networks. As neural correlations have been shown to carry information about stimuli, which may be behaviorally relevant [14-18] , a model that will take correlations into account is needed.

A natural extension of the independent model is to take into account correlations between pairs of neurons, in addition to the firing rates of individual neurons. The unique distribution which takes into account only pairwise correlations and firing rates, but makes no other implicit assumptions is the maximum entropy pairwise model ($P^{(2)}$) . The maximum entropy pairwise model can be found by methods of constrained optimization, specifically by finding the maximum of the following function:

$$\Lambda\left(P^{(2)}, \lambda\right) = H\left(P^{(2)}\right) - \sum_{i=1}^{n} h_i(< \sigma_i >_{P^{(2)}} - < \sigma_i >_{data}) - \sum_{i<j\leq n} J_{ij}(< \sigma_i\sigma_j >_{P^{(2)}} - < \sigma_i\sigma_j >_{data}) - \theta\left(\sum_{\{\sigma\}} P^{(2)}(\{\sigma\}) - 1\right) \quad (1.1)$$

where H denotes the entropy function $H(P) = -\sum_{\{\sigma\}} P(\{\sigma\}) \log_2 P(\{\sigma\})$ [19]. This is the unique model whose firing rates $< \sigma_i >$ and coincident firing rates $< \sigma_i\sigma_j >$ fit those in the data ($<\cdot>$ denotes averages over the empirical distribution P_{data} unless mentioned otherwise), and has maximal entropy. The h_i's and J_{ij}'s here are Lagrange multipliers, while θ multiplies the normalization constraint (giving rise to the partition function Z in the following formula). The solution is known to take the form [20, 21]:

$$P^{(2)} = \frac{1}{Z}\exp\left(\sum_{i=1}^{n} h_i\,\sigma_i + \sum_{i<j\leq n} J_{ij}\,\sigma_i\sigma_j\right) \quad (1.2)$$

Although a joint distribution over 10 neurons may be governed by higher orders of interactions [10, 22-25], it was found that pairwise models give a surprisingly accurate approximation of the joint distribution of neural responses in different neural systems [7-9]. The contribution of network correlation, or deviation from independence, can be quantified by $I_n = H\left(P^{(1)}\right) - H(P)$, and can be decomposed into a sum of contributions by different orders of correlations $I_n = \sum_{k=2}^{n} I^{(k)}$, where $I^{(k)}$ corresponds to the contribution of correlations of order *k* (see [11] for details). The contribution of

pairwise correlations to the total correlation structure is then quantified by $I^{(2)}/I_n = \left(H\left(P^{(1)}\right) - H\left(P^{(2)}\right)\right) / \left(H\left(P^{(1)}\right) - H(P)\right)$ (where P is the true distribution). Studies in several different neural systems have shown that pairwise correlations account for approximately 90% of the total correlation in networks of ~10 neurons [7-9].

3. Sampling accuracy and robustness of maximum entropy pairwise models of neural responses

As was previously shown [7-9], we find that although pairwise correlations are typically weak, the independence assumption fails. The distribution of synchronous spiking events (the number of spiking neurons in a single time bin, figure 1(b)) as measured in our experimental data, differs considerably from the distribution predicted by the independent model $P^{(1)}$. Again, in agreement with previous reports, the pairwise model gives a very good prediction to the empirically measured quantities (figure 1(b), (c)). We show that this result is robust to specific bin size selection. For bin sizes ranging from 5 to 40 ms, the contribution of pairwise correlations to the total network correlation (measured by $I^{(2)}/I_n$) is approximately 90%, and varies very little (figure 1(d)).

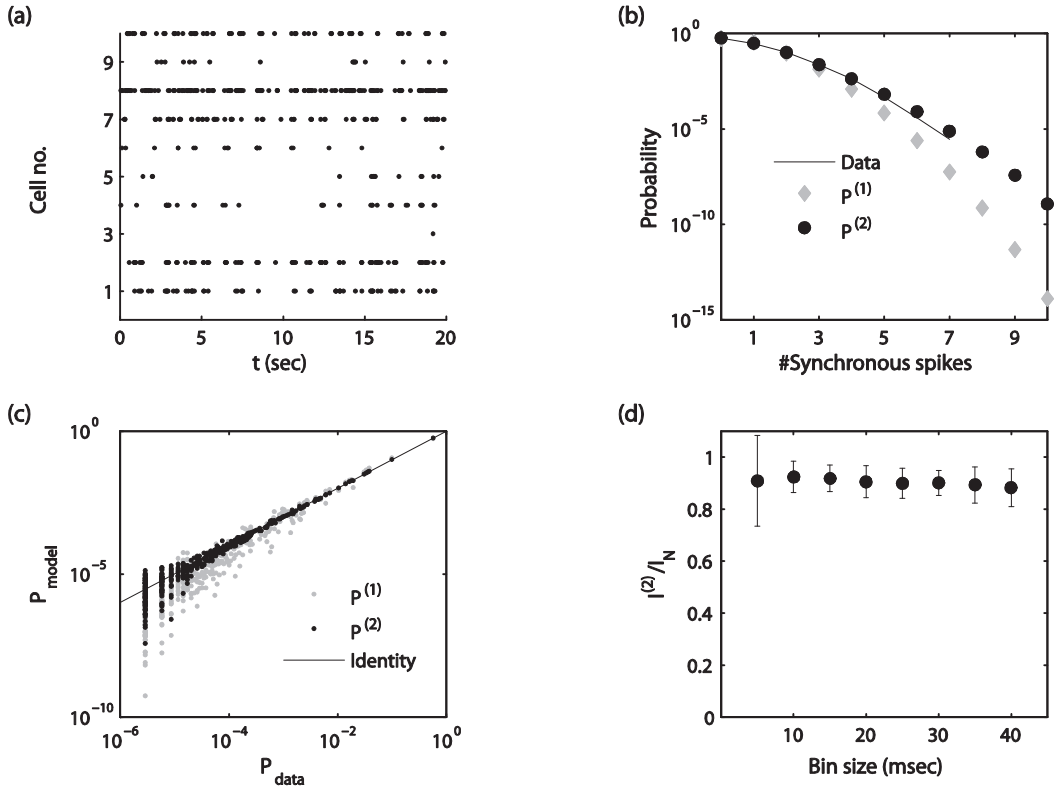


Figure 1. Comparison of independent and pairwise maximum entropy models of the joint spiking of a group of 10 neurons and the empirical distribution. **(a)** Raster plot of the response of 10 neurons to 20 sec of natural movie presentation. Each dot is a spike, the x-axis represents time, and the y-axis represents different neurons. **(b)** Distribution of synchronous spiking events. The probability of an event is plotted against the number of participating neurons. Black line connects the actual data points, black dots are the predictions of $P^{(2)}$, and gray diamonds are the predictions of $P^{(1)}$. **(c)** Predictions of $P^{(2)}$ (black dots) and $P^{(1)}$ (gray dots) for the observation frequency of each pattern in the data (ordinate), plotted against the empirically measured frequency (abscissa). Each dot represents a single pattern. Black line corresponds to identity. **(d)** The contribution of pairwise correlations to the total network correlation as measured by $I^{(2)}/I_n$ is plotted as a function of bin size used to collect spikes (error bars represent STD). Bin size has little effect on this quantity.

The only quantities we need to measure in order to uniquely determine $P^{(2)}$ are the firing rates $< \sigma_i >$ and coincident firing rates $< \sigma_i \sigma_j >$. These quantities can be accurately estimated using far fewer samples than are required in order to directly estimate the full joint distribution. Interestingly, due to the low firing rates of neurons, a characteristic common to many neural systems [26], the relative estimation error, measured by the Fano factor (variance over mean), is similar for both firing rates and coincident firing rates (figure 2(a)). This result does not hold for higher firing rates (figure 2(b), (c)).

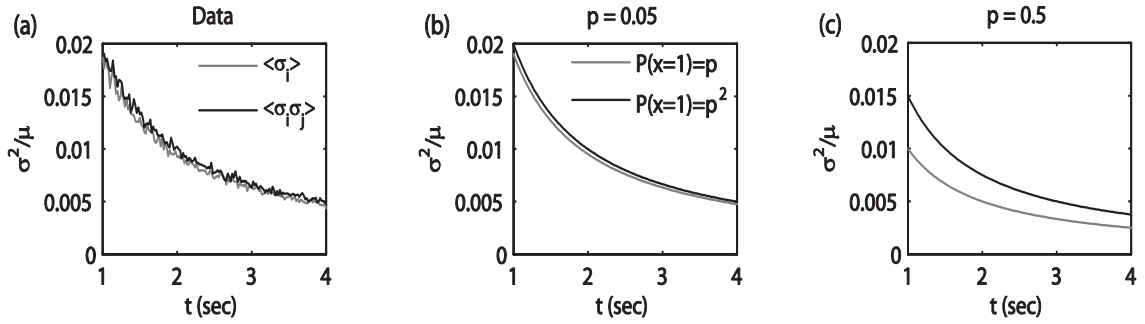


Figure 2. Sampling error for different firing rates. **(a)** Fano factor for the empirical estimates of firing rates (gray line) and coincident firing rates (black line). **(b)** The Fano factor as a function of sampling time (assuming 50 samples per second as in the experiment), for a random variable $Y = N^{-1} \sum_{i=1}^{N} X_i$, where $X \sim$ Bernoulli(p=0.05) (black line), and $X \sim$ Bernoulli(p^2) (gray line). The parameter p (which is analogous to firing rate) is set to be similar to the actual firing rates of neurons. **(c)** Same as **(b)** but for a higher probability of success (analogous to higher firing rate), p = 0.5.

Studies so far have used very large data sets in order to construct maximum entropy pairwise models. Clearly a large data set is necessary in order to construct the empirical distribution P_{data}, which enables us to measure such quantities as $I^{(2)}/I_n$. But how much data do we really need in order to construct an accurate pairwise model?

To answer this question, we constructed independent $P^{(1)}$, pairwise maximum entropy $P^{(2)}$ and triplewise maximum entropy $P^{(3)}$ models, for different amounts of data, and compared their performance. Figure 3(a) shows that the pairwise model converges to its final parameters after ~100 sec, as measured by the Kullback-Leibler divergence (D_{KL}, [21]) between $P^{(2)}$ and the empirical distribution constructed from the entire data set. A 50 minute natural movie was presented in a loop a little over two times, resulting in approximately 2 hours of continuous stimulus presentation, thus ~350,000 samples were collected from a distribution over 1024 states. We find that $P^{(2)}$ becomes a more accurate model of the population activity patterns than $P^{(1)}$ after only 35 sec, on average. Moreover, $P^{(2)}$ proves superior to the next model in the correlation based hierarchy, $P^{(3)}$, which is defined analogously to $P^{(2)}$, but with the addition of triplewise correlations $< \sigma_i \sigma_j \sigma_k >$, for times of up to 20 minutes. Thus, for a wide range of behaviorally relevant timescales, the most accurate model in the hierarchy one can construct for neural responses, and the stimuli they represent, is the pairwise model, as considering higher order correlations mainly introduces noise.

Can a pairwise maximum entropy model be learned in behaviorally relevant tasks on timescales of seconds? To test this, we presented a 50 sec long natural movie to the retina repetitively for 101 times. Offline we segmented the movie into 5 clips, each 10 sec long. We constructed a pairwise model for each clip presentation, i.e. 505 different models (101 models for each of the 5 clips). We then

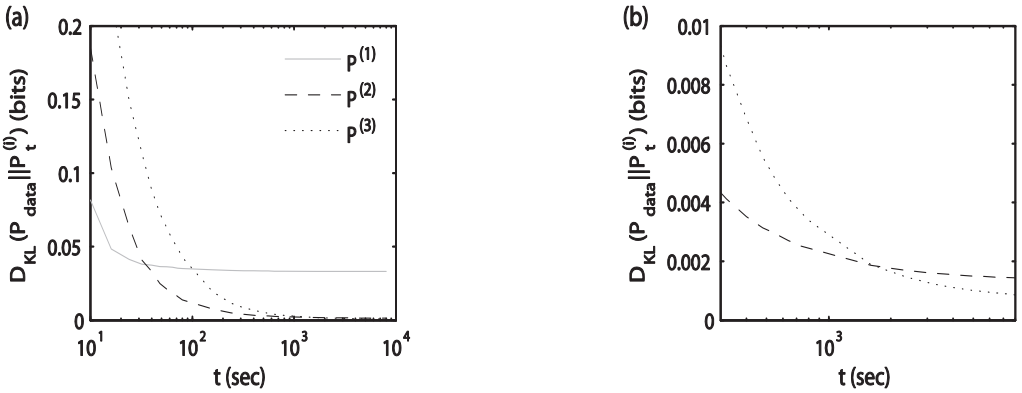


Figure 3. Sampling properties of maximum entropy models of different orders. **(a)** The Kullback-Leibler divergence of the different models in the correlation based hierarchy (see text) from the empirical distribution (estimated using approximately 2 hours of data, or 350,000 samples), is plotted as a function of the time used to estimated the relevant expected values (50 samples/sec). The pairwise model proves to be the best model in a wide range of sampling times. **(b)** Zoom in on longer time scales. $P^{(3)}$ passes $P^{(2)}$ only after over 1000 sec. Legend as in (a).

calculated the Jensen-Shannon divergence (D_{JS}[3], [21, 27]) between each pair of models. When grouping models corresponding to the same visual stimulus together in the distance matrix, we clearly see that models constructed from the same stimulus are much more similar to each other than to models of other stimuli (figure 4(a)). This result demonstrates that pairwise models constructed using a very limited amount of data (only 10 sec) can be utilized in discrimination tasks. Nevertheless, as mentioned earlier, for such short periods of time the independent model is more accurate than the pairwise model. This can be seen by the higher ratio of the average distance between models corresponding to the same stimuli and the average distance between models of differing stimuli (figure 4(b)). Based on the results of figure 3, we expect that for clips which are a little longer than 35 seconds, $P^{(2)}$ will give a more accurate description of the response distribution than $P^{(1)}$. We emphasize that this discrimination does not rely on averaging across repeated presentations of the same stimulus, as is commonly done, but rather stimuli can be discriminated based on a single presentation.

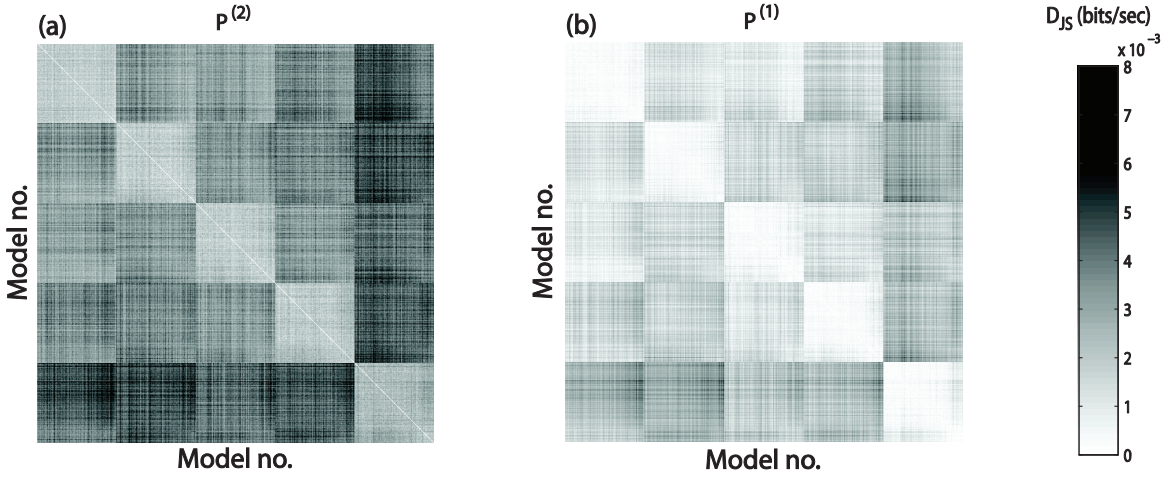
4. Relaxing constraints according to sampling certainty

Acknowledging that our empirical measurements of any observables are noisy may allow us to improve our model by trading model accuracy for increased entropy.

As evident from eq. (1.2), finding $P^{(2)}$ is reduced to finding the h_i's and J_{ij}'s. Since Λ is concave this can be done by gradient ascent based on the relevant derivatives:

$$\frac{\partial \Lambda}{\partial h_i} = < \sigma_i >_{data} - < \sigma_i >_{P^{(2)}}$$
$$\frac{\partial \Lambda}{\partial J_{ij}} = < \sigma_i \sigma_j >_{data} - < \sigma_i \sigma_j >_{P^{(2)}} \quad (4.1)$$

[3] The Jensen Shannon divergence is defined as $D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||Q) + \frac{1}{2} D_{KL}(Q||P)$. Unlike the Kullback-Leibler divergence, it provides a symmetric measure of similarity between probability distributions, which is bounded between 0 (identical distributions), and 1 (non-overlapping distributions).



$1.9\cdot10^{-3}$	$3.2\cdot10^{-3}$	$3.1\cdot10^{-3}$	$3.6\cdot10^{-3}$	$4.6\cdot10^{-3}$
	$2.0\cdot10^{-3}$	$3.4\cdot10^{-3}$	$3.7\cdot10^{-3}$	$4.7\cdot10^{-3}$
		$2.0\cdot10^{-3}$	$3.3\cdot10^{-3}$	$4.2\cdot10^{-3}$
			$2.1\cdot10^{-3}$	$4.5\cdot10^{-3}$
				$2.5\cdot10^{-3}$

$3\cdot10^{-4}$	$1.3\cdot10^{-3}$	$1.1\cdot10^{-3}$	$1.7\cdot10^{-3}$	$2.5\cdot10^{-3}$
	$5\cdot10^{-4}$	$1.6\cdot10^{-3}$	$1.8\cdot10^{-3}$	$2.8\cdot10^{-3}$
		$4\cdot10^{-4}$	$1.2\cdot10^{-3}$	$2.0\cdot10^{-3}$
			$5\cdot10^{-4}$	$2.4\cdot10^{-3}$
				$5\cdot10^{-4}$

Figure 4. Similarity map of maximum entropy models for different visual stimuli, learned using single stimulus presentation. **(a)** Top: The distance matrix of different models to different clips of a repeating natural movie. Distance was measured by the Jensen-Shannon divergence between the pairwise models generated from the expected values estimated during each 10 sec segment. Models for the same visual stimuli are grouped together. Colorbar on the right. Bottom: Average distance between models of movie clips (each clip corresponds to the same visual stimulus). Models are grouped as in above matrix. **(b)** Same as **(a)**, but using the independent model. Clearly models of the same visual stimulus are more similar than models of differing visual stimuli.

Denoting the vector of parameters by $\bar{\lambda}$, i.e. $\bar{\lambda} = (h_1 \dots h_n, J_{12} \dots J_{(n-1)n})$, we can iteratively update the model parameters according to he following formula - $\bar{\lambda}_{t+1} = \bar{\lambda}_t + \eta\nabla\Lambda$ (where $\nabla\Lambda$ denotes the gradient of Λ, $(\nabla\Lambda)_i = \frac{\partial\Lambda}{\partial\bar{\lambda}_i}$. For sufficiently small learning rate (η) this process is guaranteed to converge, thus we can find a solution that is arbitrarily close to the true maximum entropy distribution. In practice, to terminate learning in finite time we must set some convergence threshold, a predetermined condition which indicates that we are close enough to the desired solution. Such a condition is commonly set according to the magnitude of the gradient, i.e., the optimization halts when the overall gradient is smaller than some threshold, $\|\nabla\Lambda\| < \delta$. Generally, we would like δ to be as small as possible in order to achieve maximal accuracy, while the only reason to increase δ is to allow for faster termination of the optimization procedure. However, in our case it proves beneficial to increase δ when sampling noise is high. We see that when only a few samples are available, corresponding to short learning periods, a model with higher δ is in fact closer to the empirical distribution than a more stringent model with smaller δ (figure 5(a)). This is because higher δ results in higher model entropy (figure 5(b)), thus less of an effort is made to concentrate the probability mass according to the (noisy) observations, and more of the probability mass is distributed over all possible patterns. Higher δ values result in higher entropy because of the initial conditions used in our optimization procedure. Since we initialize our parameters so they correspond to a uniform distribution, higher δ values tend to lead to higher entropy distributions upon termination of the optimization.

As model accuracy improves if we adjust our model tolerances according to the noise in our estimations, we would like to automatically adjust the convergence threshold according to the confidence in our data. Every expected value estimated from our data can be approximated by a mean of a sum of Bernoulli random variables with a parameter p equal to the expected value itself. Therefore, if we require that our model's expected values are within one standard deviation from the empirical estimates, we can set the convergence threshold as follows $(\nabla\Lambda)_i \leq \sqrt{< \sigma_i >_{data} \cdot (1 - < \sigma_i >_{data})/N}$, where N is the number of data samples. The above is correct for h_i's, to adjust for J_{ij}'s the firing rates need to be substituted for the appropriate coincident firing rates. Hence, we have an individual convergence threshold for each parameter, which takes into account the confidence in the estimate of the relevant statistic. We denote this automatic adjustment of constraint flexibility *AutoFlex*, and show that the *AutoFlex* model is more accurate than the fixed threshold models we tested throughout our sampling range (figure 5(a)). We note that a more general form of using soft constraints to improve the model was presented recently by [28], where they directly maximize the entropy given the flexibility of the constraints, while here the increase in entropy is a result of the initial conditions.

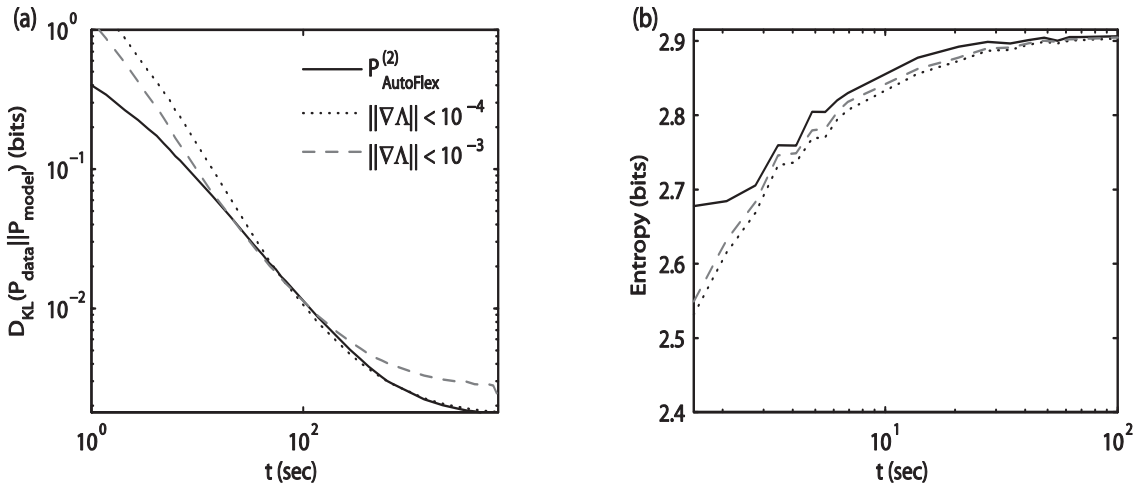


Figure 5. Flexible constraints result in a more accurate model with higher entropy. **(a)** Comparison of the accuracy of different pairwise models calculated using either fixed convergence thresholds (dashed, dotted lines) or using sample size adjusted flexible constraints (solid line), as a function of the time used to estimate firing rates and coincident firing rates (50 samples/sec). Accuracy was measured using the Kullback-Leibler divergence from the empirical distribution estimated using the entire data set (estimated using approximately 2 hours of data, or 350,000 samples). **(b)** Entropy of the corresponding models from **(a)**. Evidently, for noisy data, higher entropy results in better accuracy. The *AutoFlex* approach adjusts the convergence thresholds to get the best results throughout the sampling range.

5. Summary

Modelling the joint distribution of many interacting elements is known to be a hard computational problem encountered in many fields, such as genetics [29], statistical physics [30], machine learning [31], and image processing [32] among others [33]. As experimental technology advances, simultaneous recordings of many neurons become more readily available [34, 35]. Understanding and modelling the joint activity of groups of neurons is therefore becoming a central question in neuroscience.

Here we asked how much data is actually needed in order to construct maximum entropy models of neural population activity patterns, depending on the order of correlations that they rely on. We

showed that accurate pairwise models of the responses of groups of 10 neurons can be constructed within several seconds, and that higher order correlations, which can, in theory, improve model accuracy by about 10%, are not useful for times of up to 20 minutes, due to sampling noise. Furthermore we demonstrate that by trading the accuracy of expected value reconstruction for model entropy, we can actually improve model accuracy. Finally we suggested how to adjust optimization parameters in order to rapidly construct accurate models for neural responses.

In this study we considered groups of 10 neurons, which may display up to 1024 different activity patterns. Our experimental data which consisted of 350,000 samples from a single experiment was sufficient to estimate the full probability distribution of activity patterns of this population. However, functional networks in the brain are comprised of much larger groups of neurons. It is possible that for larger networks the contribution of higher order correlations will become much more significant and pairwise models will no longer suffice, which would require different, more subtle ways to learn the nature of population activity patterns [24, 25, 36-39].

Acknowledgements
This work was supported by grants from the Israel Science Foundation (1525/08 to ES, and 502/07 and 1619/07 to RS), fellowships from the Center for Complexity Sciences (to ES and to RS), the Clore center for Biological Physics and a Peter and Patricia Award (ES), and the Zlotowski center for Neuroscience (RS).

References

[1] Adrian E 1928 *The Basis of Sensation* (London: Christophers)
[2] Georgopoulos A P, Schwartz A B and Kettner R E 1986 *Science* **233** 1416-9
[3] Mainen Z F and Sejnowski T J 1995 *Science* **268** 1503-6
[4] de Ruyter van Steveninck R R, Lewen G D, Strong S P, Koberle R and Bialek W 1997 *Science* **275** 1805-8
[5] Rieke F, Warland D, de Ruyter van Steveninck R and Bialek W 1999 *Spikes: Exploring the Neural Code* (Cambridge, MA: The MIT Press)
[6] Yang T and Shadlen M N 2007 *Nature* **447** 1075-80
[7] Schneidman E, Berry II M J, Segev R and Bialek W 2006 *Nature* **440** 1007-12
[8] Shlens J, Field G D, Gauthier J L, Grivich M I, Petrusca D, Sher A, Litke A M and Chichilnisky E J 2006 *J. Neurosci.* **26** 8254
[9] Tang A, et al. 2008 *J. Neurosci.* **28** 505-18
[10] Amari S 2001 *ITIT* **47** 1701-11
[11] Schneidman E, Still S, Berry II M J and Bialek W 2003 *Phys. Rev. Lett.* **91** 238701
[12] Segev R, Goodhouse J, Puchalla J and Berry II M J 2004 *Nat. Neurosci.* **7** 1154-61
[13] Schneidman E, Bialek W and Berry II M J 2003 *J. Neurosci.* **23** 11539
[14] Dan Y, Alonso J M, Usrey W M and Reid R C 1998 *Nat. Neurosci.* **1** 501-7
[15] Maynard E M, Hatsopoulos N G, Ojakangas C L, Acuna B D, Sanes J N, Normann R A and Donoghue J P 1999 *J. Neurosci.* **19** 8083-93
[16] Romo R, Hernandez A, Zainos A and Salinas E 2003 *Neuron* **38** 649-57
[17] Stark E, Globerson A, Asher I and Abeles M 2008 *J. Neurosci.* **28** 10618-30
[18] Fujisawa S, Amarasingham A, Harrison M T and Buzsaki G 2008 *Nat. Neurosci.* **11** 823-33
[19] Shannon C E 1948 *Bell System Tech. J.* **27** 379-423
[20] Jaynes E T 1957 *Phys. Rev.* **106** 620-30
[21] Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley-Interscience)
[22] Bohte S M, Spekreijse H and Roelfsema P R 2000 *Neural Comput.* **12** 153-79
[23] Martignon L, Deco G, Laskey K, Diamond M, Freiwald W and Vaadia E 2000 *Neural Comput.* **12** 2621-53
[24] Amari S, Nakahara H, Wu S and Sakai Y 2003 *Neural Comput.* **15** 127-42

[25] Montani F, Ince R A A, Senatore R, Arabzadeh E, Diamond M E and Panzeri S 2009 *Phil. Trans. Royal Soc.* A **367** 3297-310
[26] Olshausen B A and Field D J 2004 *Curr. Opin. Neurobiol.* **14** 481-7
[27] Lin J 1991 *ITIT* **37** 145-51
[28] Dudík M, Phillips S J and Schapire R E 2004 *Proc. 17th Ann. Conf. Comput. Learning Theory* (New York: ACM Press) pp 655-62
[29] Shen-Orr S S, Milo R, Mangan S and Alon U 2002 *Nat. Genet.* **31** 64-8
[30] Binder K 1986 *Monte Carlo Methods in Statistical Physics* (New York: Springer-Verlag)
[31] Ackley D H, Hinton G E and Sejnowski T J 1985 *Cogn. Sci.* **9** 147-69
[32] Li S Z 2001 *Markov Random Field Modeling in Image Analysis* (Tokyo: Springer-Verlag)
[33] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 *Science* **298** 824-7
[34] Meister M, Pine J and Baylor D A 1994 *J. Neurosci. Methods* **51** 95-106
[35] Nicolelis M A L 1999 *Methods for Neural Ensemble Recordings* (Boca Raton, FL: CRC press)
[36] Tkacik G, Schneidman E, Berry II M J and Bialek W 2006 *q-bio/0611072*
[37] Bethge M and Berens P 2008 *Advances in Neural Information Processing Systems* vol 20 ed Platt J C, Koller D, Singer Y and Roweis S T (Cambridge: MIT press) pp 97-104
[38] Shlens J, Field G D, Gauthier J L, Greschner M, Sher A, Litke A M and Chichilnisky E J 2009 *J. Neurosci.* **29** 5022-31
[39] Roudi Y, Nirenberg S and Latham P E 2009 *PLoS Comput. Biol.* **5** e1000380