Signatures of criticality arise in simple neural population models with correlations

Marcel Nonnenmacher[1,2,3*], Christian Behrens[3,4], Philipp Berens[3,4,5,6], Matthias Bethge[2,3,4], Jakob H. Macke[1,2,3*]

1 research center caesar, an associate of the Max Planck Society, Bonn, Germany
2 Max Planck Institute for Biological Cybernetics, Tübingen, Germany
3 Bernstein Center for Computational Neuroscience, Tübingen
4 Centre for Integrative Neuroscience and Institute of Theoretical Physics, University of Tübingen;
5 Institute of Opthalmic Research, University of Tübingen
6 Baylor College of Medicine, Houston, TX, USA

*** marcel.nonnenmacher@caesar.de, jakob.macke@caesar.de**

Abstract

Large-scale recordings of neuronal activity make it possible to gain insights into the collective activity of neural ensembles. It has been hypothesized that neural populations might be optimized to operate at a 'thermodynamic critical point', and that this property has implications for information processing. Support for this notion has come from a series of studies which identified statistical signatures of criticality in the ensemble activity of retinal ganglion cells. What are the underlying mechanisms that give rise to these observations?

Here we show that signatures of criticality arise even in simple feed-forward models of retinal population activity. In particular, they occur whenever neural population data exhibits correlations, and is randomly sub-sampled during data analysis. These results show that signatures of criticality are not necessarily indicative of an optimized coding strategy, and challenge the utility of analysis approaches based on equilibrium thermodynamics for understanding partially observed biological systems.

1 Introduction

Recent advances in neural recording technology [1, 2] and computational tools for describing neural population activity [3] make it possible to empirically examine the statistics of large neural populations and search for principles underlying their collective dynamics [4]. One intriguing hypothesis that has emerged from this approach is the idea that neural populations might be poised at a thermodynamic critical point [5–7], and that this might have important consequences for how neural populations process and encode sensory information [7, 8]. As similar observations have been made in other biological systems (e.g. [9–11]), it has been suggested that this might reflect a more general organizing principle [12].

In the case of neural coding, evidence in favour of this hypothesis has been put forward by a series of studies which measured neural activity from large populations of retinal ganglion cells and reported that their statistics resemble those of physical systems at a critical point [7,8]. Using large-scale multielectrode array recordings [2], spike-sorting methods that scale to large ($N > 100$) populations [2,13] and specially developed maximum entropy models [3,12,14–19], Tkačik et al. observed that the specific heat—a global population statistic which measures the range of probabilities of spike patterns—diverges as a function of population size. In addition, when an artificial temperature parameter is introduced, specific heat is maximised for the statistics of the observed data rather than for statistics which have been perturbed by changing the temperature parameter. These properties of retinal populations resemble the behaviour of physical systems at critical points, and gave rise to the hypothesis that neural systems might also be poised at critical points.

What neural mechanisms can explain these observations? It had been hypothesised that the properties of the system need to be finely tuned [7,12] to keep the system at a critical point, for example through adaptation [20]. A competing hypothesis [21–23] had stated that generic mechanisms based on latent-variable models could be sufficient to give rise to activity data with these statistics, but neither of these theoretical studies had investigated mechanistic models of retinal population activity. Thus, subsequent studies advocating criticality in the retina [7,8], continued to interpret their observations as indication for the retina to be poised at a special state that is advantageous for coding. It is therefore still an open question as to whether previously reported signatures of criticality reveal a new mechanism of retinal coding, or they are a direct consequence of the standard enconding models of retinal ganglion cell responses [24–28].

We here challenge the conclusion of studies which used tools from statistical physics to search for signatures of criticality by applying exactly the same data analysis approach to a simplistic feed-forward cascade model of retinal ganglion cell responses and showing that it exhibits the same effects. Focusing on how the specific heat of this simulated data varies with population size and temperature, we show that this simple model exhibits signatures of criticality and reproduces the experimentally reported dependence on different stimulus ensembles [7]. We provide a theoretical analysis of an analytically tractable model [21,29,30], and show mathematically that it exhibits signatures of criticality under a wide range of parameters.

This analysis also points to a subtle but important difference between how practical neural data analysis and theoretical studies often differ in how they study scaling behaviour of the system: Whereas many theoretical studies describe different systems of size N, in practical neural data analysis populations of different size are typically constructed by randomly subsampling a large (but fixed) recording of neural activity. We show that this sampling process produces 'signatures of criticality' whenever neural data has non-zero correlations, which could arise from a shared stimulus drive, recurrent connectivity or global state-fluctuations [31–34].

2 Results

2.1 Signatures of criticality arise in a simple model of retinal ganglion cell activity

A hallmark of criticality is that the specific heat of the model diverges when the temperature reaches the critical temperature [5]. Tkačik et al. [7] developed a statistical approach for translating this concept to neural data analysis. In their analysis, neural populations of different size n are generated from the full recording by randomly subsampling the entire population. The statistics of activity for each population of size n are characterized using a maximum entropy model [3,14,15,17,18]. Finally, the maximum entropy models are perturbed by introducing a temperature parameter, and specific heat is computed for each population size n and temperature T from the (perturbed) maximum entropy model fit. Divergence of specific heat with population size n, and a peak of the specific heat near unit temperature $T = 1$ (the 'temperature' of the original data) are interpreted as evidence for the system being at a critical point [7].

To test if these signatures of criticality can be reproduced by canonical properties of retinal circuits, we first created a simple phenomenological model of retinal ganglion cell (RGC) activity based on linear-nonlinear neuron models [24, 25, 28]. In this model (Fig. 1a), we assumed retinal ganglion cells to have centre-surround receptive fields [28, 35] with linear spatial integration [36], sigmoid nonlinearities and stochastic binary spikes, i.e. in each time bin of size 20ms, each neuron i either emitted a spike ($x_i = 1$) or not ($x_i = 0$). We used a sequence of natural images (see Methods 3.1 for details). In addition to the feed-forward drive by the stimulus, nearby neurons received shared Gaussian noise, mimicking common input from bipolar cells [37]. Thus, cross-neural correlations in the model arise from correlations in the stimulus, receptive-field overlap and shared noise, but not from lateral connections between RGCs. Parameters of the model were chosen to approximate the statistics of receptive-field centre locations of RGCs (Fig. 1b), as well as histograms of firing rates, pairwise correlation-coefficients and population spike-counts (Fig. 1d). Nevertheless, the model clearly cannot accurately capture all statistics of real RGC activity: Our goal was not to provide a realistic model of retinal processing. Rather, we wanted to directly test whether canonical mechanisms of retinal processing (overlapping centre-surround receptive fields, spiking nonlinearities, shared Gaussian noise) are sufficient for the signatures of criticality to arise, or whether this would require fine-tuning or sophisticated neural circuitry.

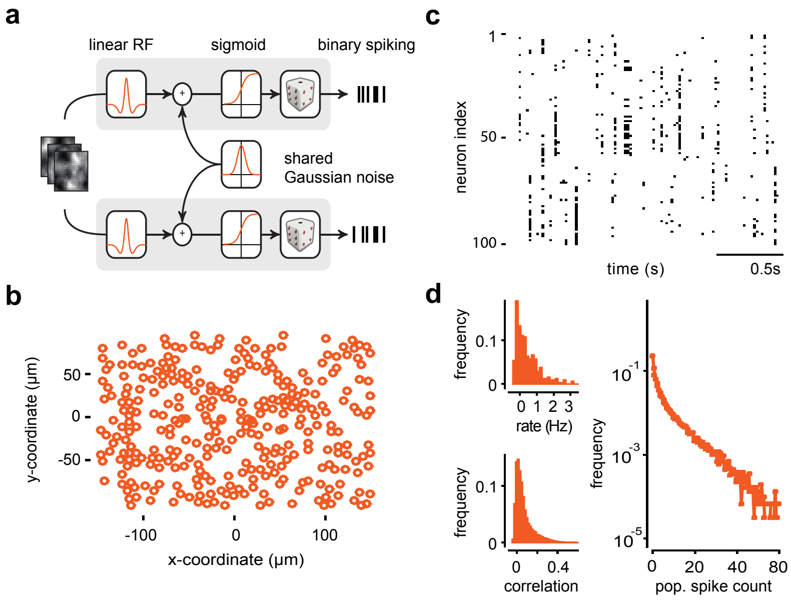

Figure 1. A simple phenomenological model of retinal ganglion cell activity **a)** Model schematic: Neurons have linear stimulus selectivity with centre-surround receptive fields and also receive correlated Gaussian noise. Neural activity is modelled in discrete time. **b)** Receptive field centres in simulation. **c)** Example raster plot of the simulated activity of 100 neurons in response to natural stimuli. **d)** Statistics of population activity in response to natural stimuli. Histogram of firing rates (top left), correlation coefficients (bottom left) and frequency of population spike-counts (right).

As a next step in the analysis, we subsampled populations of different size n by uniformly sampling cells from our simulated recording of size $N = 316$ neurons. For each population we fit a 'K-pairwise' maximum entropy model [3]. This model assigns a probability $P(\mathbf{x})$ to each spike-pattern $\mathbf{x}$. It is an extension of pairwise maximum entropy models (i.e. Ising models) [14,15] which reproduces the firing rates and pairwise covariances which has additional terms which make sure that the model also captures the population spike-counts of the data [3] (see Fig. 1d, and Methods 3.2 for details of model specification and parameterisation). As we needed to efficiently fit this model [38–40] to multiple simulated data sets, we developed an improved fitting algorithm based on maximum-likelihood techniques using Markov chain Monte Carlo (MCMC) techniques, building on work by [16]. In particular, we made the most computationally expensive component of the algorithm, the estimation of pairwise covariances via MCMC sampling, more efficient by using a 'pairwise' Gibbs-sampling scheme with Rao-Blackwellisation [41,42] (see Methods 3.2 for details). Rao-Blackwellisation resulted in a reduction of the number of samples (and computation time) needed for achieving low-variance estimates of the covariances by a factor of approximately 3 (Fig. 2a, Suppl. Inf. S1). After parameter fitting, the model reproduced the statistics of the simulated data relevant for the model (Fig. 2b). Using the formalism developed by Tkačik et al., we then introduced a temperature parameter which rescales the probabilities of the model,

$$P_T(\mathbf{x}) \propto P(\mathbf{x})^{1/T}. \tag{1}$$

Here, temperature $T = 1$ corresponds to the statistics of the empirical data. By changing T to other parameter values one can perturb the statistics of the system [43]: Increasing temperature leads to models with higher firing rates and weaker correlations (Fig. 2c), with $P_T(\mathbf{x})$ approaching the uniform distribution for very large T. If the temperature is decreased towards zero, $P_T(\mathbf{x})$ has most of its probability mass over the most probable spike patterns. In many probabilistic systems, lowering T leads to increasing correlations, as the systems then 'jumps' between several different patterns and thus the activation probabilities of different elements are strongly dependent on each other. However, for the simulated RGC activity, the sparsity of data leads to a decrease of correlations: At a bin size of 20 ms [14], the most probable state is the silent state, followed by patterns in which exactly one neurons spikes. In an example population of size $n = 100$, 53.8% of observed spike patterns contain at most one spike. When decreasing the temperature to $T < 1$, patterns with at most one spike dominate the systems even more strongly: For the same population and temperature $T = 0.8$, we find 95.6% of observed patterns to contain at most one spike. Thus when the temperature is lowered, the shift in probability mass to single-spike patterns decreases correlations.

We compute the specific heat of a population directly from the probabilistic model fit to data [7], using

$$c(T) = \frac{1}{n}\mathrm{Var}[\log P_T(X|\lambda)], \tag{2}$$

i.e. the variance of the log-probabilities of the model, normalised by n [7]. Specific heat is minimal for data in which all patterns $\mathbf{x}$ that occur in the data are equally probable, and big for data in which pattern-probabilities span a large range. We used MCMC-sampling to approximate the variance across all probabilities (see Methods 3.3), and used this approach to calculate, for each population of size n, the specific heat as a function of temperature (Fig. 2d).

We found that the temperature curves obtained from the simulated data qualitatively reproduces the critical features of those that had been observed for large-scale recordings in the salamander [7] and rat [8] retina: The peak of the curves diverges as the population size n is increased, and moves closer to unit temperature for increasing n (Fig. 2e). Consistent with experimental findings, we found that specific heat diverged linearly with population size (Fig. 2e). These results show that signatures of criticality arise in a simple feed-forward LN cascade model based on generic properties of retinal ganglion cells, and do not require finely tuned parameters or sophisticated circuitry.

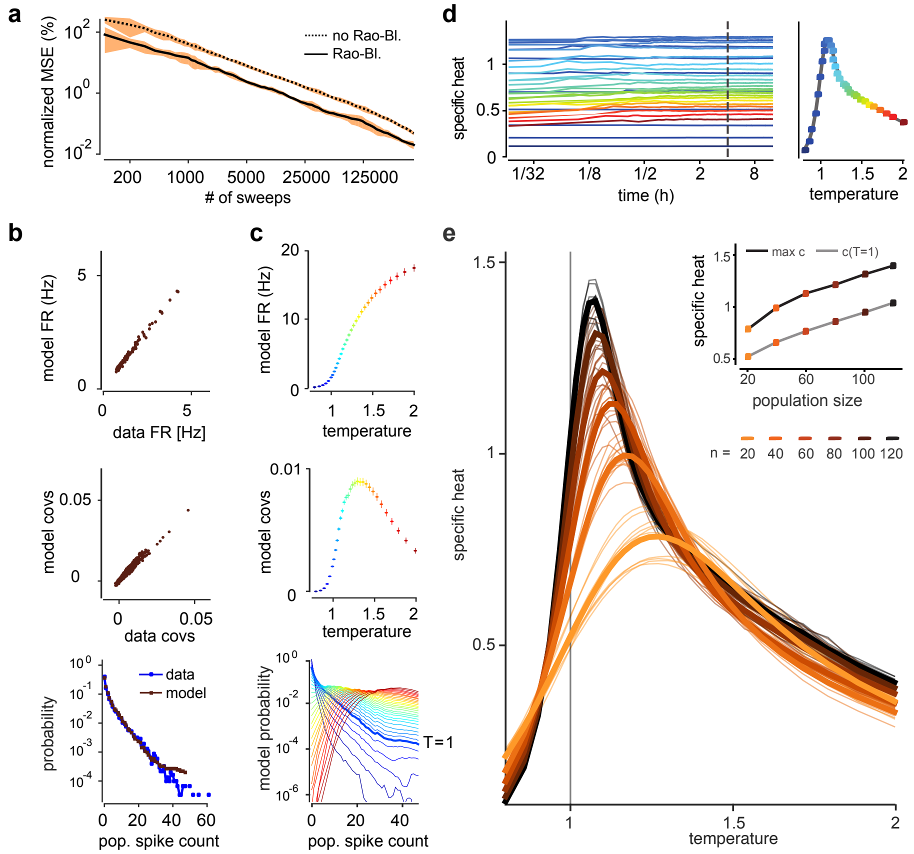

Figure 2. Signatures of criticality in a simple simulation of RGC activity a) Estimation-error (normalised mean square error) in pairwise covariances as function of sample size, averaged across 10 populations of size $n = 100$. Use of Rao-Blackwellization reduces the number of samples needed for a given level of accuracy by a factor of approximately 3, making it possible to explore multiple large population models. **b)** Quality of fit: After convergence, the population models (here $n = 100$, example population) capture the mean firing rates (top), covariances (centre) and spike count distribution (bottom) of the data. **c)** Changing the temperature parameter scales both mean firing rates (top), covariances (centre) and population spike-counts (bottom) **d)** Estimating specific heat via MCMC sampling: MCMC estimates of specific heat from a K-pairwise maximum entropy model fit to an example population (same model as in b-d). Final estimates were taken from the average over first 4h sampling time. Right: Resulting plot of specific heat as function of temperature. **e)** Divergence of specific heat: Average and individual traces for 10 randomly sampled populations for each of 6 different population sizes, exhibiting divergence of specific heat and peak in heat near unit temperature. Inset: Specific heat at unit temperature and at peak vs. population size.

2.2 Specific heat diverges linearly in flat population models

In the phenomenological population model above, we observed that specific heat grew linearly with population size, as it did in previous studies built on experimental data [7, 8, 44]. Can we understand this phenomenon analytically in a simplified model? In particular, is the divergence indeed linear, and what determines its rate? To address these questions, we replaced the K-pairwise maximum entropy model by a model which only captures the distribution of population spike-count $K = \sum_i x_i$ [21, 29, 30, 32] of the data, and in which all neurons have the same mean firing rate and pairwise correlations. This 'flat' model can be fit to data by matching its parameters to the population spike-count distribution, side-stepping the computational challenges of the K-pairwise model (see Methods 3.4 for details). We here introduce a new parametrised flat model in which the spike-count distribution is given by beta-binomial distribution $P(K|\alpha, \beta, n)$, reducing the number of free parameters from n to 2. The beta-binomial model is a straightforward extension of an independent (i.e. binomial) population model: At each time-point, a new firing probability p is drawn from a beta-distribution with parameters α and β, and neurons then spike independently with probability p. The fact that the underlying fluctuations in p are shared across the population leads to correlations in neural activity. This beta-binomial model provided a good fit to the population spike-count distributions of the simulated data (Fig. 3a) across different population sizes n (Fig. 3b). The best-fitting parameters α and β did not vary systematically across population sizes, and converged to values of $\alpha = 0.38$ and $\beta = 12.35$ (Fig. 3 c), corresponding to an average firing rate of $\mu = 1.5$ Hz (i.e. each neuron has a probability of spiking of 0.03 in each bin) and average pairwise correlations of $\rho = 0.073$. The beta-binomial model also provided good fits to population spike-count distributions published in [30] and [32], [8] (Fig. 3d). When we applied this flat model to populations subsampled from the RGC simulation, we could qualitatively reproduce the heat curves of the K-pairwise model. In particular, we found a linearly diverging peak that moved closer to $T = 1$ as the population size was increased (Fig. 3 e). Thus, linear divergence of specific heat is qualitatively captured by flat models. We note that the absolute values of the specific heat do not match those of the K-pairwise model or simulated data, but are substantially bigger ($c_{max} = 4.02$ at $T = 1.07$).

One of the difficulties of interpreting the scaling behaviour of maximum entropy models fit to neural data is the fact that the construction of the limit in n differs from those studied in statistical physics: In statistical physics, different 'n' typically correspond to systems of different total size, and the parameters are scaled as a deterministic function of n (e.g. drawn from a Gaussian with variance proportional to $1/n$ in spin-glasses [45, 46]). In studies using maximum entropy models for neural data analysis, populations of different n are obtained by randomly subsampling a fixed large recording, and the parameters are fit to each subpopulation individually. Thus, there is no analytical relationship between population size and parameter values in this approach, and this has made it hard to determine whether the scaling observed in these studies is surprising or not.

With the flat model, it is possible to analytically characterise the behaviour of the specific heat for large population sizes for this sampling process: We assume that each population of size n is randomly drawn from an underlying, infinitely large flat population model [21, 29]. Using this approach, one can mathematically show (see Methods 3.4, Suppl. Inf. S2.3 and [21] for details) that for virtually all flat models, the specific heat diverges linearly at unit temperature, but not for any other temperature $T > 1$ or $T < 1$ (Suppl. Inf. S2.4). As a consequence, the peak must move to $T = 1$ as n is increased. Hence almost any data set analysed with the methods developed by [7] will under the flat model exhibit signatures of criticality. These results hold irrespective of the details of the properties of the full populations that the subpopulations are sampled from, including full populations that are more weakly or more strongly correlated than real neural populations, and even for models with unrealistic population spike-count distributions (see Suppl. Fig. S3 for an illustration). There are only two exceptions: The first one is a model in which all neurons are independent (i.e. a binomial population model), and the second one is a flat pairwise maximum entropy model—indeed, this is the only flat model with non-vanishing correlations for which the specific heat does not have its peak at unit temperature (see Fig. 3f and [21] for an illustration).

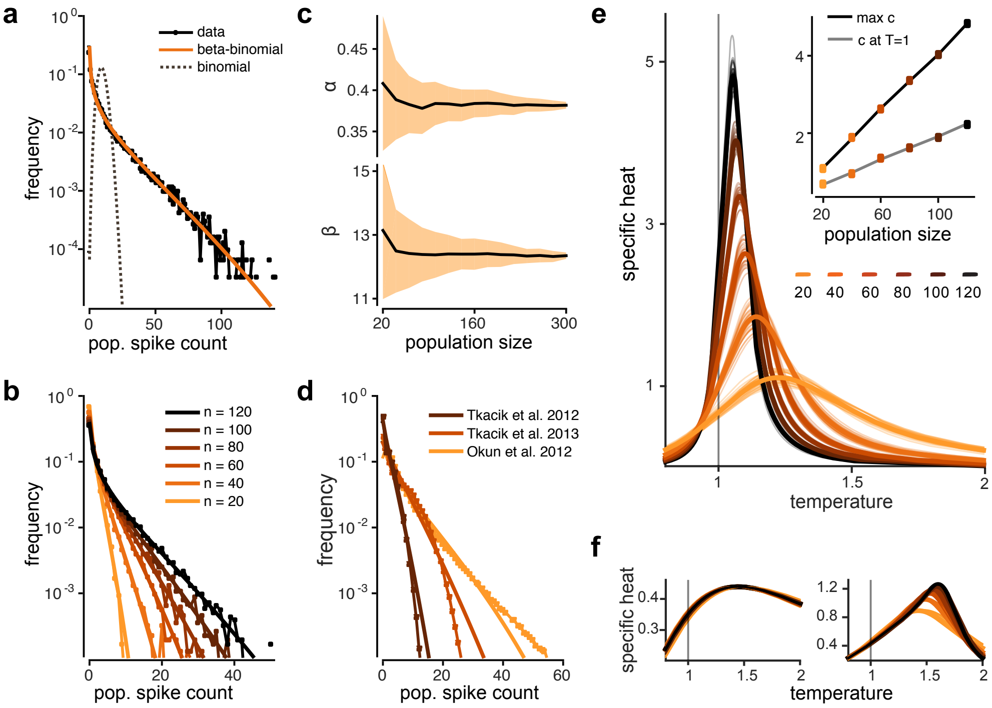

Figure 3. Signatures of criticality in a flat population model a) Population spike-count distribution in RGC simulation, and approximation by different models. Only the beta-binomial population model fits the simulated data accurately. **b)** Beta-binomial model fits for different population sizes, indicating the goodness-of-fit is robust across population size. **c)** Estimates for beta-binomial parameters α, β for data from the simulation for different population sizes (mean ± 1 s.t.d.), Best-fitting parameters do not vary systematically with population size. **d)** Beta-binomial model approximations to published empirically measured population spike-count distributions. **e)** Specific heat traces for the beta-binomial model, exhibiting signatures of criticality. Average and individual traces for 30 randomly sampled populations for each of 6 different population sizes. Inset: Specific heat at unit temperature and at peak vs. population size. **f)** Heat traces for independent model and flat pairwise maximum entropy model, which do not exhibit a divergence of the specific heat.

2.3 Strong neural correlations lead to fast divergence of specific heat.

The rate at which the specific heat diverges provides a mean of quantifying the 'strength' of criticality. What is the relationship between correlations in a neural population and the rate of divergence? To study how the specific heat rate $\tilde{c} = c(T = 1)/n$ depends on the strength of correlations, we used a beta-binomial model to generate simulated data with firing rate $\mu = 1.5$ Hz (i.e. each neuron has a probability of spiking of 0.03 in each bin), and different (population-wide, as all neuron pairs have the same correlation) pairwise correlation coefficients ρ ranging from $\rho = 0.01$ to $\rho = 0.25$ (Fig. 4a). We found that the heat curves had the same shape as in the analyses above, with a peak that increases and moves to unit temperature (Fig. 4b). Comparing the results for different specified correlation strengths within the populations, we found that the specific heat rates $\tilde{c}$ increased strictly monotonically with ρ (Fig. 4b,c). For the beta-binomial model, the large-n value of $\tilde{c}$ can be calculated analytically (see Suppl. Inf. S3.2 for details) as a function of the parameters α and β,

$$\begin{aligned}\tilde{c} = &\frac{\alpha(\alpha+1)\psi_1(\alpha+1) + \beta(\beta+1)\psi_1(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\ &+ \frac{\alpha\beta\left(\psi_0(\alpha+1) - \psi_0(\beta+1)\right)^2}{(\alpha+\beta)^2(\alpha+\beta+1)} - \psi_1(\alpha+\beta+1). \end{aligned} \tag{3}$$

This analytical evaluation of $\tilde{c}$ (valid for large n) was in good agreement with numerical simulations (Fig. 4c left). In the case of weak correlations ρ, equation 3 can be simplified: In this case, the specific heat rate is proportional to the strength of correlations (see Suppl. Inf. S3.1 for details), i.e.

$$\tilde{c} \approx \rho\ \mu(1-\mu)\left(\log\left(\frac{1-\mu}{\mu}\right)\right)^2 \tag{4}$$

This expression can also be derived from the Gaussian model in [8] equation (4), by inserting the expected values of the mean and variance of the population spike-count under random subsampling. Thus, at least for flat models and the analysis based on specific heat proposed previously, 'being very critical' is a consequence of 'being strongly correlated'.

2.4 Specific heat depends on average correlation strength in K-pairwise model

Is the relationship between the strength and correlations and the 'strength' of criticality (i.e. the divergence rate of specific heat) also true in more general models? In the original study [7], specific heat was computed from K-pairwise model fits to RGC activity resulting from three different kind of stimuli: Checker-board stimuli (which do not have long-range spatial correlations, although stimulus-driven cross-neural correlations can arise from receptive field overlap), natural images, which exhibit strong spatial correlations, and full-field flicker (which constitutes an extreme case of spatial correlations since all pixels in the display are identical). Tkačik et al. found that specific heat diverges in all three conditions, and interpreted this as evidence that signatures of criticality are not 'inherited from the stimulus' [7]. Comparing the specific heat values for $n = 100$ reported in [7] across stimulus conditions, Tkačik et al. found the smallest peak for checkerboard stimuli ($c_{max} = 0.54$ for $n = 100$), intermediate for natural images ($c_{max} = 0.92$) and strongest for full-field flicker ($c_{max} = 2.4$).

We tested whether we find the same pattern of results in K-pairwise model fits to our retinal simulation. Specific heat divergence also followed the pattern predicted by the flat models (Fig. 4d): Checkerboard (which gave an average correlation between neural activity of $\rho = 0.033$) had the smallest peak (peak specific heat $c_{max} = 0.87$) followed by natural images ($\rho = 0.075$, $c_{max} = 1.32$) and full-field flicker ($\rho = 0.341$, $c_{max} = 3.09$). We conclude that the experimental evidence—which showed that the specific heat diverges, and how the speed of divergences depends on the stimulus ensemble—is entirely consistent with a simple, feed-forward phenomenological model of retinal processing.

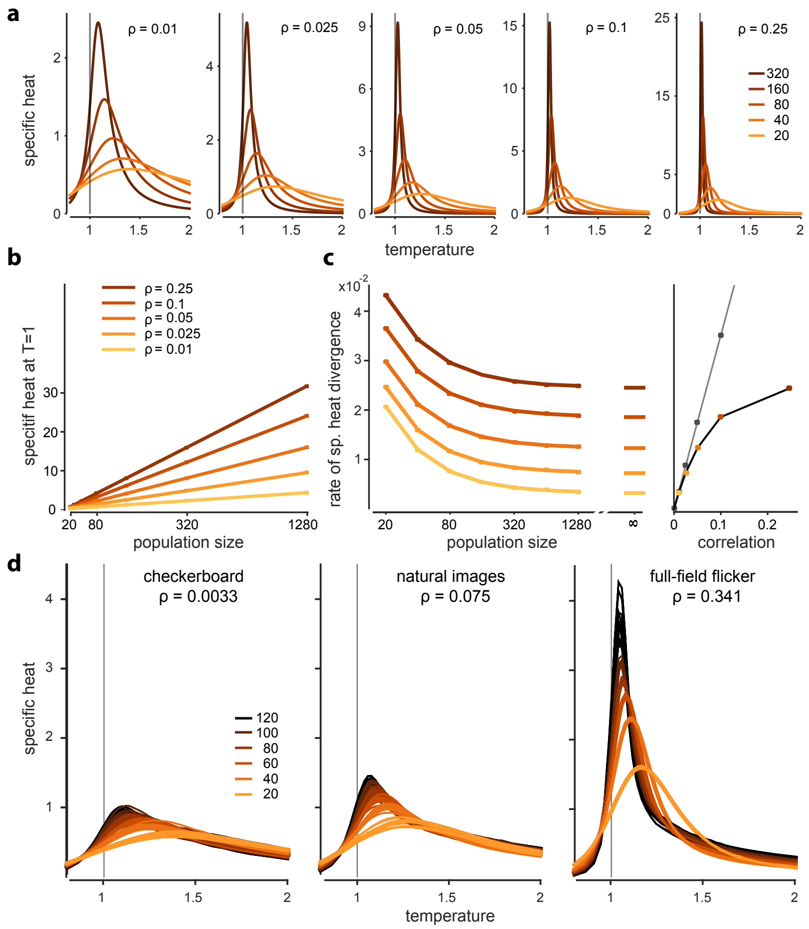

Figure 4. Relationship between correlations and criticality. a) Specific heat traces for beta-binomial model of different correlation strengths and population sizes. Heat traces are qualitatively similar, but differ markedly quantitatively (see y-axes). **b)** Specific heat diverges linearly, and the slope depends on the strength of correlations (left). Divergence rate of specific heat for beta-binomial model as a function of correlation strength (centre). Rightmost point (at infinity) corresponds to analytical prediction of large-n behaviour. Divergence rates are strictly increasing with correlation strength (right) which is captured by a weak-correlation approximation (dashed line). **c)** Specific heat increases with correlation in the K-pairwise maximum entropy model: average and individual traces for 10 randomly subsampled populations for 6 different population sizes. Left to right: checkerboard, natural images and full-field flicker stimuli presented to the population. Correlation strengths denote mean correlation coefficient in each population.

2.5 Sources of criticality-inducing correlations in neural activity

In the above, we showed that a beta-binomial spike count distribution can be sufficient for signatures of criticality to arise. For this to hold we need the variance of the population spike-count to grow at least quadratically with population size, i.e. $\mathrm{Var}(K) \propto n^2$. The variance of the population spike-count is equal to the sum of all variances and covariances in the population, $\mathrm{Var}(K) = \sum_{i=1}^{n} \mathrm{Var}(x_i) + \sum_{i \neq j} \mathrm{Cov}(x_i, x_j)$. A sufficient condition for signatures of criticality to arise in these models is that the average covariances (and hence correlations) between neurons are independent of n, $\frac{1}{n(n-1)}\sum_{i\neq j}\mathrm{Cov}(x_i,x_j)\approx$ constant [5, 6]. One possible correlation structure which has this properties are so called 'infinite range' correlations (Fig. 5a): correlation between neurons do not drop off to 0 for large spatial distances. In this case, adding more and more neurons to a population will not change the average pairwise correlation within the population (Fig. 5b).

In neural systems, there are at least two reasons that can facilitate the required correlation structure. First, as shown above, the choice of stimuli has a clear effect on the heat capacity indicating an important effect of input-induced correlations. In particular for full-field flicker stimuli infinite-range correlations are to be expected but also white noise input can generate correlations of considerable extent due to overlapping receptive fields. Second, even a neural population which does not have infinite range correlations can appear critical if it is randomly subsampled during analysis: Suppose that different populations of size n are obtained as above by (uniformly) subsampling a large recording of size N. Then, for any correlation structure on the full recording (including limited-range correlations, Fig. 5c), the average correlation in a population of size n will be independent of n (Fig. 5c): If neurons are randomly subsampled from the large recording, then the pairwise correlations in each subpopulation are also a random subsample of the large correlation matrix. As a consequence, the average correlation will be independent of n, and specific heat will diverge with constant slope (Fig. 5d). In contrast, if different population sizes are constructed by taking into account the spatial structure of the population (i.e. by iteratively adding neighbouring cells) then the average correlation in each subpopulation will drop with n, and the slope of specific heat growth will decrease with population size.

In our RGC simulation, correlations did drop off to zero with spatial distance for checkerboard and natural images, but not for full-field flicker (Fig. 5e). Correlations in the full-field flicker condition initially drop off due to distance-dependent shared noise, but eventually saturate at a level far above zero that is determined by the full-field stimulus. Due to these strong infinite-range correlations, both spatially structured sampling and uniform sampling then give rise to linear growth in specific heat (Fig. 5f left). For the other two stimulus conditions, however, the choice of subsampling scheme does result in markedly different behavior of the specific heat growth: Both for natural images and checkerboard stimuli, we can see the rate of growth decreases for large n under spatially structured subsampling (Fig. 5f centre, right). This effect will be more pronounced for larger simulations, and in additional simulations we found specific heat to saturate completely once populations are substantially bigger than the spatial range of correlations.

In summary, populations will exhibit critical behaviour if correlations have infinite range (over the size of the recording), irrespective of the sampling scheme. In addition, if a population is randomly subsampled (as was done in [7, 8]), then signatures of criticality will arise even if the underlying correlations have limited range.

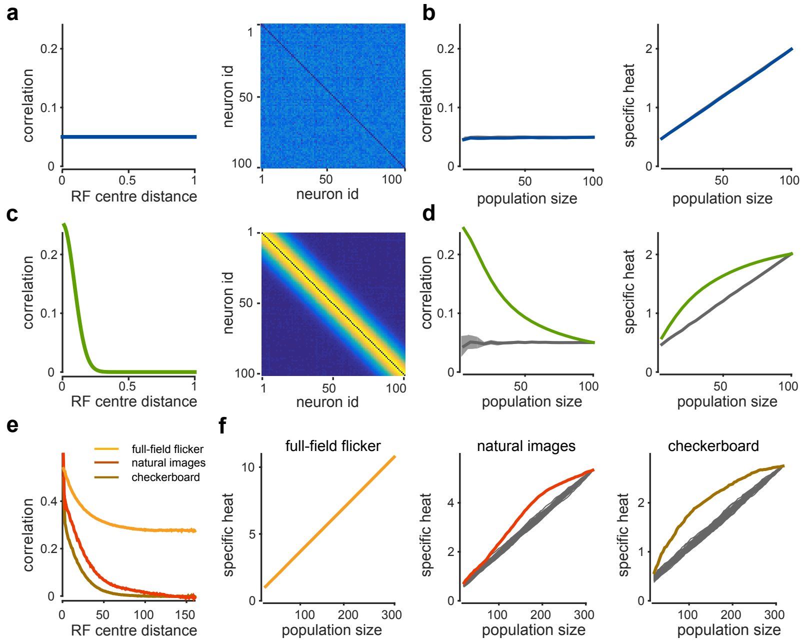

Figure 5. Random subsampling leads to criticality-inducing correlations. a) Illustration: A population with 100 neurons and infinite-range correlations, the average correlation between any pair of neurons is close to 0.05. Correlation as function of inter-neuron distance (left) and full correlation matrix (right). **b)** Average correlation in subpopulation of different size n (left) and specific heat as function of n (right), when neurons are sampled from 1 to 100. Random sampling gives identical results (not shown). **c)** Population with limited-range correlations, same plots as in panel a. **d)** Left: Average correlation as function of population size for ordered sampling (green) and uniform subsampling (gray). Right: Specific heat grows linearly for random subsampling, but shows signs of saturation for ordered sampling. **e)** Average correlation as function of inter-neuron distance in RGC simulation. For checkerboard and natural images, correlations drop to 0 for large distances. **f)** Specific heat for different stimulation conditions, for ordered (colour) or random subsampling (gray).

3 Materials and Methods

3.1 Numerical simulation of retinal ganglion cell activity

Retina simulation: We simulated a population of $N = 316$ retinal ganglion cells as linear threshold neurons whose receptive fields were modelled by difference-of-Gaussian filters with ON-centres [25,28,36]. The simulation comprised two subgroups of cells with different receptive field sizes (surrounds 56µm and 30µm in retinal space, centres 28µm and 15µm, respectively, one third cells with large receptive fields). For both subgroups, the weight of the surround was 0.5 of the centre weight. Locations of receptive field centres were based on a reconstruction of 518 soma locations from a patch of mouse retina [47]. As the reconstructed locations in that data set also comprised about 40% amacrine cell somata, we randomly discarded 40% of the cell locations. The resulting patch of retina covered an area of 200×300µm^2, corresponding to 100×150 pixels in stimulus space. Correlated noise across neurons was modelled using correlated additive Gaussian noise. Correlations dropped off exponentially with soma distance with a decay constant of $\tau = 30$µm i.e. noise covariance matrix was chosen as $\Sigma = \sigma^2_{noise}(aI_n + be^{-\Delta/\tau})$, where

Δ_{ij} is the distance between neurons i and j and $a^2 + b^2 = 1$. We set $\sigma_{noise} = 0.022$ and $a = 0.45$. We modelled neural spiking in discrete time using 20ms bins. In each bin t, the total input $z_i(t)$ to neuron i was given by $z_i(t) = w_i^\top s(t) + \epsilon_i(t)$, where w_i is the receptive field of neuron i, $s(t)$ the vectorised stimulus and $\epsilon_i(t)$ the input noise of neuron i. A neuron in a given bin is active ($x_i = 1$) if $z_i + d > 0.5$ and inactive ($x_i = 0$) otherwise, with offset $d = 0.168$. Parameters of the simulation (centre and surround sizes, relative strength of centre and surround, magnitude and correlations of noise, spiking threshold) were chosen to roughly match the statistics of neural spiking (firing rates, pairwise correlations, population activity counts) reported in studies of salamander retinal ganglion cells [2, 3, 14]. (Code will be available at www.mackelab.org/code).

Stimuli: We used three types of stimuli for this study: natural images, checkerboard patterns and full-field flicker. For natural image stimuli, we used a sequence of 101 images of meadow sceneries taken from low hight. Each image was 400 × 400 pixels, and each image was presented for 20ms with 300 repetitions total. The luminance histograms of the images were transformed to a normal distribution with mean 0.5 and pixel values between 0 and 1.

For the full-field flicker stimulus, luminance levels were drawn from a Gaussian distribution with mean $\mu = 0.5$ and variance $\sigma^2 = 0.06$. Checkerboard stimuli consisted of 80 × 80 tiles of size 5 × 5 pixels each. Luminance levels (from within the interval $[0, 1]$) of each tile were chosen to be either 0.15 or 0.77 with probability 0.5. The parameters of both stimulus sets were chosen to match the dynamic range of the simulated retinal ganglion cells. For both types of stimuli, 2000 images were generated and the image sequences were presented with 10 repetitions. To calculate specific heat as function of increasing population size, we randomly selected 10 subsamples of the full simulated population of $N = 316$ cells at population sizes $n \in \{20, 40, 60, 80, 100, 120\}$ by uniformly drawing n neurons out of the full population without replacement.

3.2 Modeling neural population data with maximum entropy models

Model definition: We modelled retinal ganglion cell activity by using a 'K-pairwise' maximum entropy model [3]. In a maximum entropy model [48], the probability of observing the binary spike word $\mathbf{x} \in \{0, 1\}^n$ for parameters $\lambda = \{h, J, V\}$ is given by

$$P(\mathbf{x}|\lambda) = \frac{1}{Z(\lambda)} \exp\left(h^\top \mathbf{x} + \mathbf{x}^\top J \mathbf{x} + \sum_{k=0}^{n} V_k \delta\left(K(\mathbf{x}) = k\right)\right) \tag{5}$$

Here, the parameter vector h (of size $n \times 1$) and the upper-triangular matrix $J \in \mathbb{R}^{n \times n}$ correspond to the bias-terms and interaction terms in a pairwise maximum entropy model (also known as an Ising model or spin-glass) [14]. The term $K(\mathbf{x}) = \sum_{i=1}^n x_i$ denotes the population spike-count, i.e. the total number of spikes across the population within a single time bin, and the indicator-term $\delta\left(K = k\right)$ is 1 whenever the population spike-count equals k, and is 0 otherwise. The term $\sum_{k=0}^n V_k \delta\left(K = k\right)$ was introduced by [3] to ensure that the model precisely captures the population spike-count distribution of the data using n additional free parameters. The partition function Z for given λ is chosen such that the probabilities of the model sum to 1.

Parameter fitting: To fit the model parameters $\lambda = \{h, J, V\}$ to a data set $D = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\}$, we maximised the penalised log-likelihood [49, 50] of the data D

under the model,

$$L(h,J,V) := \sum_{m=1}^{M} \log P(\mathbf{x}^{(m)}|h,J,V) - \frac{1}{\sigma_h}\|h\|_1 - \frac{1}{\sigma_J}\|J\|_1 - \frac{1}{2}V^T\Sigma^{-1}V. \qquad (6)$$

Here, the $l1$-penalty controlled the magnitudes of parameters h, J, the term $\|J\|_1$ favoured sparse coupling matrices, and the regularisation term Σ on the V-parameters ensures that the terms controlling the spike count distribution vary smoothly in k (Suppl. Inf. S1). This smoothness prior is particularly important for large spike counts, as it makes it possible to interpolate parameters for which the number of observed counts is small.

In maximum entropy models, exact evaluation of the penalised log-likelihood and its gradients requires the calculation of expectations under the model, $\mathrm{E}[x_i]$, $\mathrm{E}[x_i x_j]$ or equivalently $\mathrm{cov}(x_i, x_j)$, and $P(K=k)$ (Suppl. Inf. S1.1), which in turn requires summations over all 2^n possible states $\mathbf{x}$ and is prohibitive for $n > 20$. Following previous work [16], we used Gibbs sampling to approximate the relevant expectations (Suppl. Inf. S1.1 for derivations and implementation details). We used two modifications over previous applications of Gibbs sampling to fitting maximum entropy models to neural population spike train data, with the goals of speeding up parameter learning and alleviating memory usage:

First, we use Rao-Blackwellisation [41, 42] to speed up convergence of the estimation of covariances of $\mathbf{x}$. We used pairwise Gibbs sampling (blocked Gibbs with block size 2), where each new sample in the MCMC chain was obtained by updating two entries i and j of $\mathbf{x}$ at a time, rather than just a single entry. This allowed us to get estimates of the conditional probabilities $P(x_i x_j = 1|x_{\sim\{i,j\}})$, and to use them to speed up the estimation of the second moment $\mathrm{E}[x_i x_j]$ from empirical average of these conditional probabilities (Suppl. Inf. S1.1).

Second, we used a variant of coordinate ascent that calculated all relevant quantities as running averages over the MCMC sample, and thereby avoided having to store the entire $n \times \tilde{M}$ MCMC sample in memory [16], where $\tilde{M}$ is the length of the sample. Because all features of the maximum entropy model are either 0 or 1 (x_i, $x_i x_j$ and the indicator function for the spike count), the gain in log-likelihood obtainable from either updating a single element of h or J [16, 40], or from updating all V simultaneously (but not from updating multiple entries of h and J) can be computed directly from MCMC estimates of $\mathrm{E}[x_i]$, $\mathrm{E}[x_i x_j]$ and $P(K=k)$ (Suppl. Inf. S1.2). For each iteration, we calculated the gain in log-likelihood for each possible update of h_i, J_{ij} and full V, and picked the update which led to the largest gain [16, 51].

We measured the length of Markov chains in sweeps, where one sweep corresponds to one round of $n(n-1)/2$ Markov chain updates that encompasses all pairs of entries of $\mathbf{x}$ in random order. We set a learning schedule that started at 800 sweeps for the first parameter update and doubled the number of sweeps in the chain after each set of 1000 parameter updates. We monitored convergence of the algorithm using a normalised mean square error between empirical $\mathrm{E}[x_i]$, $\mathrm{cov}(x_i, x_j)$, $P(K=k)$ and their estimates from the MCMC sample. For normalisation, we used the average squared values of the target quantity, e.g. $\frac{1}{n}\sum_{i=1}^{n} < x_i^2 >$ for the firing rates. We stopped the algorithm when a pre-set threshold was reached (0.01%, 0.25%, 0.01% for $\mathrm{E}[x_i]$, $\mathrm{cov}(x_i, x_j)$, $P(K=k)$, respectively), or when the fitting algorithm took more than $\left(\frac{n}{100}\right)^2 \times 72$h of computation time on a single core (2.294 GHz AMD Opteron(TM) Processor 6276) (Suppl. Fig. S1). For 10 populations of size $n = 100$ (for natural images), the normalised MSEs after model-fitting were 0.43%, 2.80%, 0.42%). An implementation of the fitting algorithms in MATLAB will be available at www.mackelab.org/code.

3.3 Calculating specific heat and temperature curves

Specific heat calculations: To investigate thermodynamic properties of neural population codes, Tkačik et al [7] introduced a temperature parameter T for equation 5:

$$P_T\left(\mathbf{x}|\lambda\right) = \frac{1}{Z_T}\exp\left(\frac{1}{T}\left(h^\top\mathbf{x} + \mathbf{x}^\top J\mathbf{x} + \sum_{k=0}^{n} V_k \delta\left(K(\mathbf{x}) = k\right)\right)\right) \tag{7}$$

Model fits are obtained at $T = 1$, and the temperature parameter T is scaled to study the system (i.e. characterised by $P_T\left(\mathbf{x}|h, J, V\right)$ for $T \ = 1$). We note that varying T, in effect, modulates probabilities by exponentiating them with $1/T$,

$$P_T(\mathbf{x}) \propto \left(P_{T=1}(\mathbf{x})\right)^{1/T}, \tag{8}$$

and that the family of probability distributions obtained by varying T can be constructed for any distribution, not just maximum entropy models. For large temperatures P_T approaches a uniform distribution ($P_T(\mathbf{x}) \approx 2^{-n}$ for each $\mathbf{x}$), whereas for small temperatures it converges to a singleton, $P_T(\mathbf{x}^*) \approx 1$ with $\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x}}(P_{T=1}(\mathbf{x}))$.

The specific heat, as given in equation 2, can be obtained from the variance of the log-probabilities of the model. As the variance in practice can not be outright computed for n beyond 20, we obtained estimates of $c(T)$ using a pairwise Gibbs sampler. We note that the specific heat does not depend on Z_T, as changing Z_T results in a constant, additive shift in log-probabilities which does not affect the variance. We tracked the variance of log-probabilities over an MCMC chain $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\tilde{M})}$ of length $\tilde{M}$ sampled at temperature T,

$$c(T) \approx \frac{1}{n}\left(\frac{1}{\tilde{M}}\sum_{m=1}^{\tilde{M}}\left(\log P_T\left(\mathbf{x}^{(m)}|\lambda\right)\right)^2 - \left(\frac{1}{\tilde{M}}\sum_{m=1}^{\tilde{M}}\log P_T\left(\mathbf{x}^{(m)}|\lambda\right)\right)^2\right). \tag{9}$$

For each population, we evaluated $c(T)$ for 31 temperatures between $T = 0.8$ and $T = 2$, and found the Gibbs sampler to provide reliable estimates over this temperature range. We used a burn-in of $2 * 10^4$ sweeps, and ran the sampler for $\left(\frac{n}{100}\right)^2 \times 4$h of CPU time, resulting in between 9.97e5 and 1.72e6 sweeps (mean ± std) for $n = 100$ (i.e. between 4.94e9 and 8.52e9 sampled individual spike words).

3.4 Simplified population models and the beta-binomial model

For the theoretical analysis, we adopted a class of population models (here referred to as 'flat' models) in which all neurons have identical mean firing rates, pairwise correlations and higher-order correlations [3, 21, 29, 52, 53]. Such a model is fully specified by the population spike-count distribution $P(K = k)$, and all spike words with the same spike count are equally probable. As a result, the probabilities of individual patterns $\mathbf{x}$ can be read off from the spike count distribution by

$$P(\mathbf{x}) = \binom{n}{k}^{-1} P(K = k) \tag{10}$$

whenever $\sum_{i=1}^{n} x_i = k$. In a maximum entropy formalism, this model can be obtained by setting $h_i = 0$ and $J_{ij} = 0$ for all $i, j \in \{1, \ldots, n\}$ and only optimising entries of V. Without loss of generality, we fixed fixed $V_0 = 0$ [30], resulting in n degrees of freedom for the model.

In flat models, it is possible to explicitly construct a limit $n \to \infty$ which will help us understand population analyses performed on experimental data: We assume that there is a spike count density $f(r)$, $r \in [0,1]$, which describes the population spike-count distribution of an infinitely large population. $f(r)$ denotes the probability density of a fraction of r neurons spiking simultaneously. Finite-size populations of n cells are then obtained as random subsamples out of this infinitely large system. Based on previous findings by [21], we show in Suppl. Inf. S2.3 that, in this construction, flat models always exhibit a linear divergence of specific heat, unless the limit $f(r)$ is given by either a single delta peak or a mixture of two symmetric delta peaks. These two models corresponds to systems that (for large n) either behave like a fully independent population (whose spike count distribution converges to a single delta peak), or a population described by a pure pairwise maximum entropy model (which converges to two delta peaks). In particular, any flat model with higher-order correlations [18, 27, 52, 53], or a non-degenerate $f(r)$, will exhibit 'signatures of criticality'. Furthermore, we show that, for continuous $f(r)$, $c(T)$ does not diverge for any $T \neq 1$. In combination, these results show that the peak of the specific heat is mathematically bound to converge to $T = 1$ for $n \to \infty$ in this model class.

We further simplified the flat model by re-parametrising $P(K = k)$ by a beta-binomial distribution, thereby reducing the number of parameters from n to two, and—importantly—obtaining parameters which do not explicitly depend on n. In this model,

$$P(K = k) = \binom{n}{k} \frac{\mathrm{Beta}(\alpha + k, \beta + n - k)}{\mathrm{Beta}(\alpha, \beta)} \tag{11}$$

and

$$f(r) = \frac{1}{\mathrm{Beta}(\alpha, \beta)} r^{\alpha - 1}(1 - r)^{\beta - 1}. \tag{12}$$

For simulated data, we found values for α, β extracted from the beta-binomial fits to populations of different sizes n to be stable over a large range of n (Fig. 3b). We used the beta-binomial parameters obtained from the largest investigated n to estimate the divergence rate $\tilde{c}$ for $n \to \infty$.

4 Discussion

An intriguing hypothesis about the collective activity of large neural populations has been the idea that their statistics resemble those of physical systems at a critical point. Using a definition of criticality which is based on temporal dynamics with power-law statistics, numerous studies have reported and studied critical behaviour in neural population activity [8, 20, 54, 55]. Multiple possible mechanisms for these dynamics have been proposed (e.g. [20, 56, 57]). It has been argued that such temporal dynamics might be beneficial for neural computation and communication [20, 58, 59] (see [5] for an overview). More recently, a second line of studies [5–8, 11, 12] has studied the statistics of time-instantenous patterns of neural activity using tools from statistical mechanics, and argued that they also exhibit critical behaviour. This hypothesis could open up further questions on how the system maintains its critical state, and what implications this observation has for how neural populations encode sensory information and perform computations on it. Similarly, signatures of criticality have also been observed in natural images [11] and small cortical populations [6], and have been studied using the theory of finite-size scaling and critical exponents [6]. It has been argued that systems close to a critical point might be optimally sensitive to external perturbations [6] and that the large dynamic range of the code (i.e. the large variance of log-probabilities) might be beneficial for encoding sensory events which likewise have a large distribution of occurrence-probabilities [17].

Alternatively, generic mechanisms could be sufficient to give rise to activity data with these statistics. We had demonstrated in a previous theoretical study [21] that a simple models with common input can exhibit signatures of criticality. More recently, Schwab et al. [22] and Aitchison et al. [23] elaborated on these findings, showing that common input (or other latent variables which lead to shared modulations in firing rates) can give rise to Zipf-like scaling of pattern probabilities (a second signature of criticality). Mathematically, Zipf's Law is equivalent to stating that the plot of entropy vs energy (i.e. log-probability) is a straight line with unit slope [22, 23]. Schwab et al [22] showed that particular latent variable models give rise to Zipf's law. This result was generalized by [23] which showed that, under fairly general circumstances, high-dimensional latent variable models exhibit a wide distribution of energies (i.e. log-probabilities) and hence a large specific heat. In addition, they showed that large fluctuations in the specific heat are (under some additional assumptions) sufficient to achieve Zipfian scaling. While it has also been argued that the use of data-sets which are too small might give rise to spuriously big specific heats [60]– while this is true in principle, additional analyses e.g. in [7] show that their results are robust with respect to data-set size.

However, neither of these previous theoretical studies analysed mechanistic models of neural population activity, nor did they have tools for studying population statistics in large simulations or recordings, and they were therefore limited to studying very small ($N < 20$) systems. It has thus been an open question of whether and how these theoretical considerations can account for effects observed in retinal ganglion cells. We here showed that surprisingly simple mechanisms are sufficient for two key signatures of thermodynamic criticality—a divergence of specific heat and a peak of the specific heat near unit temperature— to arise.

We found that neural population activity exhibits signatures of criticality whenever the average correlation in population of different sizes is larger than zero and does not depend on population size. In the thermodynamic analysis of physical systems at equilibrium, long-range correlations typically vanish in the thermodynamic limit. In neural systems, however, such 'criticality-inducing' correlations can arise as a consequence of various factors: In a local patch of retina, retinal ganglion cells have a large degree of receptive field overlap, and natural stimuli also contain strong spatial correlations. This can lead to correlations which do have unlimited range within the experimentally accessible length scales. Thus, fluctuations in the stimulus will lead to common activity modulations amongst neurons within the population. Empirically, activity correlations between pairs of retinal ganglion cells only fall of slowly with the distance between somas (or receptive field centres) [28]. Similarly, Mora et al [8] used a moving-bar stimulus with strong temporal correlations, and found that including activity from multiple time-lags markedly increase the strength of specific heat. We hypothesise that this increase in specific heat is a consequence of temporal correlations being stronger than inter-neural correlations in this stimulus condition. In addition, firing rates of cortical neurons are modulated by global fluctuations in excitability [31–34], resulting in neural correlations with infinite range.

Finally, we showed that criticality-inducing correlations arise as a consequence of constructing different subpopulations by uniformly subsampling a large recording with correlations. Signatures of criticality are entirely consistent with canonical properties of neural population activity, and require neither finely-tuned parameters in the population, nor sophisticated circuitry or active mechanisms for keeping the system at the critical point. Signatures of criticality are likely going to be found not just in retinal ganglion cells, but in multiple brain areas and model systems. These observation raise the question of whether signatures of criticality are really indicative of an underlying principle, or rather are a consequence of viewing the statistics of neural populations through the lens of equilibrium thermodynamics. In order to realise the potential of large-scale recordings of neural activity in the search of a theory of neural computation, we will need data-analysis methods which are adapted to the specific properties of biological data [4,19].

5 Acknowledgments

Work was funded by the German Federal Ministry of Education and Research (BMBF; FKZ: 01GQ1002, Bernstein Center Tübingen), the German Research Foundation (BE 5601/1-1 to PB) and the Max Planck Society. We thank F. Franzen for help with figures and cluster computing, S. Buchholz, D. Greenberg and S. Turaga for comments on the manuscript and useful discussions.

References

1. Kerr JND, Denk W. Imaging in vivo: watching the brain in action. Nature Reviews Neurosci. 2008;9(3):195–205.

2. Marre O, Amodei D, Deshmukh N, Sadeghi K, Soo F, Holy TE, et al. Mapping a complete neural population in the retina. The Journal of Neuroscience. 2012;32(43):14859–14873.

3. Tkačik G, Marre O, Amodei D, Schneidman E, Bialek W, Berry MJ 2nd. Searching for collective behavior in a large network of sensory neurons. PLoS Comput Biol. 2014;10(1):e1003408.

4. Gao P, Ganguli S. On simplicity and complexity in the brave new world of large-scale neuroscience. Current opinion in neurobiology. 2015;32:148–155.

5. Beggs JM, Timme N. Being critical of criticality in the brain. Frontiers in physiology. 2012;3.

6. Yu S, Yang H, Shriki O, Plenz D. Universal organization of resting brain activity at the thermodynamic critical point. Front Syst Neurosci. 2013;7:42.

7. Tkačik G, Mora T, Marre O, Amodei D, Palmer SE, Berry MJ, et al. Thermodynamics and signatures of criticality in a network of neurons. Proceedings of the National Academy of Sciences. 2015;112(37):11508–11513.

8. Mora T, Deny S, Marre O. Dynamical criticality in the collective activity of a population of retinal neurons. Physical review letters. 2015;114(7):078105.

9. Mora T, Walczak AM, Bialek W, Callan CG. Maximum entropy models for antibody diversity. Proceedings of the National Academy of Sciences. 2010;107(12):5405–5410.

10. Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M, et al. Statistical mechanics for natural flocks of birds. Proceedings of the National Academy of Sciences. 2012;109(13):4786–4791.

11. Stephens GJ, Mora T, Tkačik G, Bialek W. Statistical thermodynamics of natural images. Phys Rev Lett. 2013 Jan;110(1):018701.

12. Mora T, Bialek W. Are biological systems poised at criticality? Journal of Statistical Physics. 2011;144(2):268–302.

13. Segev R, Goodhouse J, Puchalla J, Berry MJn. Recording spikes from a large fraction of the ganglion cells in a retinal patch. Nature Neuroscience. 2004;7(10):1154–1161.

14. Schneidman E, Berry MJn, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006;440(7087):1007–12.

15. Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, et al. The structure of multi-neuron firing patterns in primate retina. J Neurosci. 2006;26(32):8254–66.

16. Broderick T, Dudik M, Tkacik G, Schapire RE, Bialek W. Faster solutions of the inverse pairwise Ising problem. arXiv. 2007;0712.2437v2.

17. Tkacik G, Schneidman E, Berry MJ II, Bialek W. Spin glass models for a network of real neurons. arXiv:q-bio/0611072v2. 2009;.

18. Ohiorhenuan IE, Mechler F, Purpura KP, Schmid AM, Hu Q, Victor JD. Sparse coding and high-order correlations in fine-scale cortical networks. Nature. 2010;466(7306):617–621.

19. Roudi Y, Dunn B, Hertz J. Multi-neuronal activity and functional connectivity in cell assemblies. Current opinion in neurobiology. 2015;32:38–44.

20. Shew WL, Clawson WP, Pobst J, Karimipanah Y, Wright NC, Wessel R. Adaptation to sensory input tunes visual cortex to criticality. Nature Physics. 2015;11(8):659–663.

21. Macke JH, Opper M, Bethge M. Common input explains higher-order correlations and entropy in a simple model of neural population activity. Physical Review Letters. 2011;106(20):208102.

22. Schwab DJ, Nemenman I, Mehta P. Zipf's law and criticality in multivariate data without fine-tuning. Physical review letters. 2014;113(6):068102.

23. Aitchison L, Corradi N, Latham PE. Zipf's law arises naturally in structured, high-dimensional data. arXiv preprint. 2014;1407.7135.

24. Chichilnisky E. A simple white noise analysis of neuronal light responses. Network: Computation in Neural Systems. 2001;12(2):199–213.

25. Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, et al. Do we know what the early visual system does? J Neurosci. 2005;25(46):10577–10597.

26. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature. 2008;454(7207):995–9.

27. Leen DA, Shea-Brown E. A Simple Mechanism for Beyond-Pairwise Correlations in Integrate-and-Fire Neurons. J Math Neurosci. 2015;5(1):30.

28. Pitkow X, Meister M. Decorrelation and efficient coding by retinal ganglion cells. Nature neuroscience. 2012;15(4):628–635.

29. Amari Si, Nakahara H, Wu S, Sakai Y. Synchronous firing and higher-order interactions in neuron pool. Neural Computation. 2003;15(1):127–142.

30. Tkačik G, Marre O, Mora T, Amodei D, Berry II MJ, Bialek W. The simplest maximum entropy model for collective behavior in a neural network. Journal of Statistical Mechanics: Theory and Experiment. 2013;2013(03):P03011.

31. Harris KD, Thiele A. Cortical state and attention. Nat Rev Neurosci. 2011;12(9):509–523.

32. Okun M, Yger P, Marguet SL, Gerard-Mercier F, Benucci A, Katzner S, et al. Population rate dynamics and multineuron firing patterns in sensory cortex. J Neurosci. 2012;32(48):17108–19.

33. Ecker AS, Berens P, Cotton RJ, Subramaniyan M, Denfield GH, Cadwell CR, et al. State dependence of noise correlations in macaque primary visual cortex. Neuron. 2014;82(1):235–48.

34. Schölvinck ML, Saleem AB, Benucci A, Harris KD, Carandini M. Cortical state determines global variability and correlations in visual cortex. J Neurosci. 2015 Jan;35(1):170–8.

35. Kuffler SW. Discharge patterns and functional organization of mammalian retina. Journal of neurophysiology. 1953;16(1):37–68.

36. Rodieck RW. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vision research. 1965;5(12):583–601.

37. Trong PK, Rieke F. Origin of correlated activity between parasol retinal ganglion cells. Nature Neuroscience. 2008;11(11):1343–1351.

38. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. Physical review letters. 1988;61(23):2635.

39. Sohl-Dickstein J, Battaglino PB, DeWeese MR. New method for parameter estimation in probabilistic models: minimum probability flow. Physical review letters. 2011;107(22):220601.

40. Schwartz G, Macke J, Amodei D, Tang H, Berry MJ 2nd. Low error discrimination using a correlated population code. J Neurophysiol. 2012;108(4):1069–88.

41. Radhakrishna Rao C. Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society. 1945;37(3):81–91.

42. Blackwell D. Conditional expectation and unbiased sequential estimation. The Annals of Mathematical Statistics. 1947;p. 105–110.

43. Kirkpatrick S, Gelatt CD, Vecchi MP, et al. Optimization by simulated annealing. science. 1983;220(4598):671–680.

44. Tkacik G, Schneidman E, Berry II MJ, Bialek W. Ising models for networks of real neurons. arXiv preprint. 2006;0611072v1.

45. Sherrington D, Kirkpatrick S. Solvable model of a spin-glass. Physical review letters. 1975;35(26):1792.

46. Mezard M, Parisi G, Virasoro M. Spin Glass Theory and Beyond (Singapore: Word Scientific); 1987.

47. Baden T, Berens P, Roman-Roson M, Bethge M, Euler T. The functional diversity of mouse retinal ganglion cells. Nature. (in press);.

48. Jaynes ET. Information theory and statistical mechanics. Physical review. 1957;106(4):620.

49. Dudík M, Schapire RE. Maximum entropy distribution estimation with generalized regularization. In: Learning Theory. Springer; 2006. p. 123–138.

50. Altun Y, Smola A. Unifying divergence minimization and statistical inference via convex duality. In: Learning theory. Springer; 2006. p. 139–153.

51. Dudik M, Phillips SJ, Schapire RE. Performance guarantees for regularized maximum entropy density estimation. In: Learning Theory. Springer; 2004. p. 472–486.

52. Yu S, Yang H, Nakahara H, Santos GS, Nikolic D, Plenz D. Higher-order interactions characterized in cortical activity. J Neurosci. 2011;31(48):17514–17526.

53. Barreiro AK, Gjorgjieva J, Rieke F, Shea-Brown E. When do microcircuits produce beyond-pairwise correlations? Front Comput Neurosci. 2014;8:10.

54. Beggs JM, Plenz D. Neuronal avalanches in neocortical circuits. The Journal of neuroscience. 2003;23(35):11167–11177.

55. Petermann T, Thiagarajan TC, Lebedev MA, Nicolelis MAL, Chialvo DR, Plenz D. Spontaneous cortical activity in awake monkeys composed of neuronal avalanches. Proc Natl Acad Sci U S A. 2009 Sep;106(37):15921–6.

56. Levina A, Herrmann JM, Geisel T. Dynamical synapses causing self-organized criticality in neural networks. Nature physics. 2007;3(12):857–860.

57. Markram H, Muller E, Ramaswamy S, Reimann MW, Abdellah M, Sanchez CA, et al. Reconstruction and Simulation of Neocortical Microcircuitry. Cell. 2015;163(2):456–92.

58. Bertschinger N, Natschläger T. Real-time computation at the edge of chaos in recurrent neural networks. Neural computation. 2004;16(7):1413–1436.

59. Shew WL, Yang H, Yu S, Roy R, Plenz D. Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. J Neurosci. 2011;31(1):55–63.

60. Saremi S, Sejnowski TJ. On Criticality in High-Dimensional Data. Neural Comput. 2014 Jul;26(7):1329–1339.

Supporting Information

S1 Fitting the K-pairwise maximum entropy model to data

To identify the values $\hat{\lambda}$ of the model parameters which yield the best fit of the maximum entropy model to data, we maximise the log-likelihood of the model given the data. The general form of the log-likelihood of a maximum entropy model parametrised by vector λ is given by

$$L(\lambda) = \sum_{m=1}^{M} \log P(\mathbf{x}^{(m)}|\lambda) = -M \log Z_\lambda + \sum_{m=1}^{M} \lambda^T f(\mathbf{x}^{(m)}) \tag{13}$$

for the spike-data vectors $x^{(m)} \in \{0,1\}^n$, $m = 1, \ldots, M$. Any choice of the feature function f defines a specific maximum entropy model over this n-dimensional binary space. For the K-pairwise maximum entropy model used in this paper, $f(\mathbf{x}) \in \{0,1\}^{n(n+3)/2+1}$ is composed of

1. n first-order features

   $$f_i(\mathbf{x}) = x_i$$

   with corresponding parameters collected in h. The h_i, $i = 1, \ldots, n$ control single-cell firing rates (in units of bins rather than Hz).

2. $n(n-1)/2$ second-order features

   $$f_{ij}(\mathbf{x}) = x_i x_j$$

   with parameters J_{ij}, $j, i = 1, \ldots, n$, $i < j$, controlling pairwise neuronal correlations, and

3. n+1 population-scale features

   $$f_k(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_i x_i = k \\ 0, & \text{otherwise} \end{cases}$$

   with parameters V_k, $k = 0, \ldots, n$. The vector V controls the overall number of spikes in each temporal bin.

Note that that there is some degeneracy between the parameter vectors V and both h and J —a global upwards shift of firing rates for example can be achieved both by adding a positive constant ϵ to each h_i, or by adding ϵk to each of the V_k. Similarly, adding a constant ϵ to every J_{ij} can be balanced by subtracting $\epsilon \frac{k(k-1)}{2}$ from each V_k. Since either manipulation of V is zero for $k = 0$, fixing $V_{k=0} = 0$ is not sufficient for getting rid of this parameter degeneracy. As we never interpreted the parameter-values themselves, but only the fit to data, we made no attempt to add additional constraints.

We can re-write the K-pairwise model into the general maximum entropy form by stacking the feature functions f_i, f_{ij}, and f_k into the vector-valued feature function f and doing the same with parameters h_i, J_{ij}, and V_k to obtain $\lambda = \{h, J, V\} \in$

$\mathbb{R}^{n(n+3)/2+1}$. The derivative of the log-likelihood with respect to any single parameter $\lambda_l, l = 1, \ldots, n(n+3)/2 + 1$ is given by

$$\begin{aligned}\frac{\delta}{\delta\lambda_l}\sum_{m=1}^{M}\log P(\mathbf{x}^{(m)}|\lambda) &= \frac{\delta}{\delta\lambda_l}\sum_{m=1}^{M}\left(\lambda^T f(\mathbf{x}^{(m)}) - \log Z_\lambda\right)\\ &= \sum_{m=1}^{M}\frac{\delta}{\delta\lambda_l}\lambda^T f(\mathbf{x}^{(m)}) - \frac{\delta}{\delta\lambda_l}M\log\sum_{\mathbf{x}}\exp\left(\lambda^T f(\mathbf{x})\right)\\ &= \sum_{m=1}^{M} f_l(\mathbf{x}^{(m)}) - M\frac{\sum_{\mathbf{x}}\lambda_l\exp\left(\lambda^T f(\mathbf{x})\right)}{\sum_{\mathbf{x}}\exp\left(\lambda^T F(\mathbf{x})\right)}\\ &= M\left(\frac{1}{M}\sum_{m=1}^{M} f_l(\mathbf{x}^{(m)}) - \mathrm{E}_\lambda[f_l(\mathbf{x})]\right) \qquad (14)\end{aligned}$$

As can be seen from equation (14), the gradient of the log-likelihood vanishes if and only if the data means match the expectations of $f(\mathbf{x})$ under the model.

To deal with data-sets of limited size, we maximised a regularised variant of the log-likelihood,

$$L(h, J, V|\sigma_h, \sigma_J, \Sigma) := \sum_{m=1}^{M}\log P(\mathbf{x}^{(m)}|h, J, V) - \frac{1}{\sigma_h}\|h\|_1 - \frac{1}{\sigma_J}\|J\|_1 - \frac{1}{2}V^T\Sigma^{-1}V \qquad (15)$$

$$\begin{aligned}\Sigma &= (\sigma_S S + \sigma_{\mathbb{I}}\mathbb{I}) - \frac{1}{\sigma_S + \sigma_{\mathbb{I}}}S_{0\bullet}{S_{0\bullet}}^T\\ S_{kk'} &= \exp\left(-\frac{(k-k')^2}{2\tau_S^2}\right)\\ S_{0k} &= \sigma_S\exp\left(-\frac{k^2}{2\tau_S^2}\right).\end{aligned}$$

Here, the matrix Σ implements a combined ridge and smoothing regression over V, with $(n+1)\times(n+1)$ identity matrix $\mathbb{I}$ and smoothing matrix S corresponding to a squared-exponential kernel [**?**]. We set $V_0 = 0$ and accounted for this by conditioning on V_0 and correspondingly subtracted $S_{0\bullet}(\sigma_S + \sigma_{\mathbb{I}})^{-1}{S_{0\bullet}}^T$ from Σ. We used $\sigma_h = \sigma_J = 10^4$, $\sigma_S = 10$, $\sigma_{\mathbb{I}} = 400$ and $\tau_S = 10$.

To fit maximum entropy models to large neural populations, one needs to

1. efficiently approximate the feature moments $\mathrm{E}_\lambda[f(\mathbf{x})]$ needed for the gradients of both eq. (13) and eq. (15), which for large populations ($n > 20$) can not be calculated exactly

2. find efficient methods for updating the parameters λ.

We introduce two modifications over previous approaches to fitting maximum entropy models to neural data [**?**] to improve computational efficiency:

1. We used pairwise Gibbs sampling and Rao-Blackwellisation to considerably improve estimation of the second-order feature moments $\mathrm{E}_\lambda[f_{ij}(\mathbf{x})]$

2. The authors of [51] described a trick for efficiently updating the parameters in pairwise binary maximum entropy models: If one restricted updates to coordinate-wise updates, then one can calculate the gain from updating a single variable in closed form, which makes it easy to select both the variable to update as well as the step-length in closed form. We show how this trick can be extended to allow

a joint update of all the population-count features V. In addition, the gain in log-likelihood is linear in the feature-moments, which makes it possible to compute it from a running average over the MCMC sample, and avoids having to store the entire sample in memory at any point.

We describe our contributions in the sections S1.1 and S1.2, respectively.

S1.1 Pairwise Gibbs sampling and Rao-Blackwellisation

Following previous work [16], we used MCMC sampling to approximate the expectations of the feature functions $f(\mathbf{x})$ under the K-pairwise model with parameters λ. These expected values $\mathrm{E}_\lambda[f(\mathbf{x})]$ are required to evaluate the gradients of the (penalised) log-likelihood, as well as the log-likelihood gains resulting from parameter updates. As the number of pairwise terms grows quadratically with population size n, most of the parameters of the model $P(\mathbf{x}|\lambda)$ for large n control pairwise moments $\mathrm{E}_\lambda[x_i x_j]$. To make the estimation of these pairwise interactions more efficient, we implemented a pairwise Gibbs sampler that for each update step of the Markov chain samples two variables x_i and x_j, $i \neq j$, $i, j \in 1, \ldots, n$. This furthermore allowed us to 'Rao-Blackwellise' the single-cell and pair-wise feature components $f_i(\mathbf{x}) = x_i$ and $f_{ij}(\mathbf{x}) = x_i x_j$ [**?**, 41, 42], i.e. to use the conditional probabilities $P(x_i = 1|x_{\sim i}, \lambda)$ and $P(x_i x_j = 1|x_{\sim\{i,j\}}, \lambda)$ for moment estimation, instead of the binary x_i and $x_i x_j$.

Rao-Blackwellisation provably reduces the variance of the resulting estimators, and empirically resulted in substantially faster convergence of the MCMC-estimated model firing rates $\mathrm{E}_\lambda[f_i(\mathbf{x})]$, second moments $\mathrm{E}_\lambda[f_{ij}(\mathbf{x})]$, and thus also of the covariances $\mathrm{cov}_\lambda(\mathbf{x}_i, \mathbf{x}_j|\lambda) = \mathrm{E}_\lambda[f_{ij}(\mathbf{x})] - \mathrm{E}_\lambda[f_i(\mathbf{x})]\mathrm{E}_\lambda[f_j(\mathbf{x})]$ (see supplementary figure S1). Unlike the binary variables x_i, $x_i x_j$ however, the conditional probabilities are real numbers from the interval $(0, 1)$ and cannot be stored in memory-efficient sparse matrices. We thus implemented a running average over conditional probabilities that discards the current chain element immediately after drawing the next one, while keeping track of the quantities

$$\mathrm{E}_\lambda[f_i(\mathbf{x})] \approx \frac{1}{\tilde{m}} \sum_{m=1}^{\tilde{m}} P(x_i^{(m)} = 1|x_{\sim\{i\}}^{(m)}, \lambda)$$

$$\mathrm{E}_\lambda[f_{ij}(\mathbf{x})] \approx \frac{1}{\tilde{m}} \sum_{m=1}^{\tilde{m}} P(x_i^{(m)} x_j^{(m)} = 1|x_{\sim\{i,j\}}^{(m)}, \lambda)$$

as $\tilde{m}$ increases from 1 to MCMC sample size $\tilde{M}$. We also kept track of the non-Rao-Blackwellised estimates

$$\mathrm{E}_\lambda[f_k(\mathbf{x})] \approx \frac{1}{\tilde{m}} \sum_{m=1}^{\tilde{m}} \delta\left(\sum_{i=1}^{n} x_i^{(m)}, k\right)$$

for the expectations of the population-level indicator feature functions $E_\lambda[f_k(\mathbf{x})] = P(K = k|\lambda)$, with Kronecker delta function $\delta(x, y) = 1$ if $x = y$, and $\delta(x, y) = 0$ otherwise.

We quantified the advantage of Rao-Blackwellising the Gibbs sampler with long Markov chains drawn from the K-pairwise maximum entropy model fits to populations of size $n = 100$ drawn from the simulated RGC data. For each investigated parameter fit, we ran two chains under different conditions: a first chain for which we Rao-Blackwellised the single-cell and pairwise feature moments, and a second chain for which we did not. These Markov chains were run for $\tilde{M} = 10^6$ sweeps and hence orders of magnitude longer than had occurred for the invidivual parameter updates within this study, which comprised 800 to 30000 sweeps, or 3.96×10^6 to 1.485×10^6 individual MCMC chain updates at $n = 100$. The long sample runs served to give an approximation for the "true" expected values of the target quantities of interest to us: firing rates $\mathrm{E}_\lambda[f_i(\mathbf{x})]$, covariances $\mathrm{cov}_\lambda(\mathbf{x}_i, \mathbf{x}_j)$ and population spike count distribution $P(K = k|\lambda)$.

We quantified the speed of convergence of the estimates to the "true" expected feature moments by the normalised MSE between sampler-derived feature moments after any given length $0 < \tilde{m} < \tilde{M}$ of the MCMC chain and the results we got after the full chain length. After the full $\tilde{M} = 10^6$ sweeps, the Rao-Blackwellised and non-Rao-Blackwellised estimates on average differed by $1.7 \times 10^{-4}\%$, 0.013% and $4 \times 10^{-6}\%$ normalised MSE for firing rates, covariances and population spike count distributions, respectively. We computed the distance to "truth" for each condition as the normalised MSE to the $\mathrm{E}_\lambda[f(\mathbf{x})]$ averaged over both conditions. We obtained MCMC estimates for the feature moments of the K-pairwise maximum entropy models fits to 10 subsampled populations of $n = 100$ neurons each drawn from our retina simulation. Supplementary figure S1a displays the results for the two conditions, Rao-Blackwellised vs. non-Rao-Blackwellised, for each of the 10 investigated fits.

MSEs of firing rates for single-cell features $\mathrm{E}_\lambda[\mathbf{x}_i]$ did not benefit from Rao-Blackwellisation. This is expected, as each x_i is sampled $n - 1$ times per sweep and thus the moments are already well estimated relative to the second-order features. For covariances $\mathrm{cov}_\lambda(\mathbf{x}_i, \mathbf{x}_j)$, normalised MSEs showed clear improvement under Rao-Blackwellisation, visible as an approximately constant offset between the avarages over all 10 parameter fits in the loglog-domain as seen in figure S1b. The normalised MSE on average was 3.19 times higher for non-Rao-Blackwellised (given by the downwards offset of the normalised MSEs of the Rao-Blackwellised estimates). The fraction of samples needed from Rao-Blackwellised runs to achieve the same normalised MSE on the pariwise moments than non-Rao-Blackwellised runs (given by the leftward offset of the normalised MSEs of the Rao-Blackwellised) overall was 32.02%. The fraction ranged from 34.93% at 800 sweeps to 31.74% at 30000 sweeps. The ratio of normalised MSEs was similarly stable, being 2.96 times higher at 800 sweeps and 3.27 times higher at 30000 sweeps for non-Rao-Blackwellised samples than for Rao-Blackwellised ones.

S1.2 Exploiting the structure of the K-pairwise feature functions allows blockwise parameter updates.

As described in the previous section, we can use MCMC to obtain the expected values of the feature function $\mathrm{E}_\lambda[f(\mathbf{x})]$ that are needed to to optimise the model parameters λ. To find the parameter setting $\hat{\lambda}$ which maximise the log-likelihood over the given data vectors $\mathbf{x}^{(m)}$, $m = 1, \ldots, M$, we follow an iterative update scheme introduced previously [51], and extend it to the K-pairwise model. The update scheme optimises parameter changes $\lambda^{new} - \lambda^{old}$ relative to a current parameter estimate λ^{old}, rather than the parameters λ directly. The benefit of this scheme over standard gradient ascent on the regularised log-ligkelihood as in eq. (14) is that we can give closed-form solutions for optimal values of a single component λ_l when temporarily holding all other components $\lambda_{\sim l}$ fixed.

Changing the current parameter estimate λ^{old} to λ^{new} leads to a change in log-likelihood of

$$\begin{aligned}\Delta L(\lambda^{new}, \lambda^{old}) &= \frac{1}{M}\sum_{m=1}^{M} \log P(\mathbf{x}^{(m)}|\lambda^{new}) - \frac{1}{M}\sum_{m=1}^{M} \log P(\mathbf{x}^{(m)}|\lambda^{old}) \\ &= (\lambda^{new} - \lambda^{old})^T \left(\frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x}^{(m)})\right) - \mathrm{E}_{\lambda^{old}}\left[\exp\left((\lambda^{new} - \lambda^{old})^T f(\mathbf{x})\right)\right]\end{aligned} \tag{16}$$

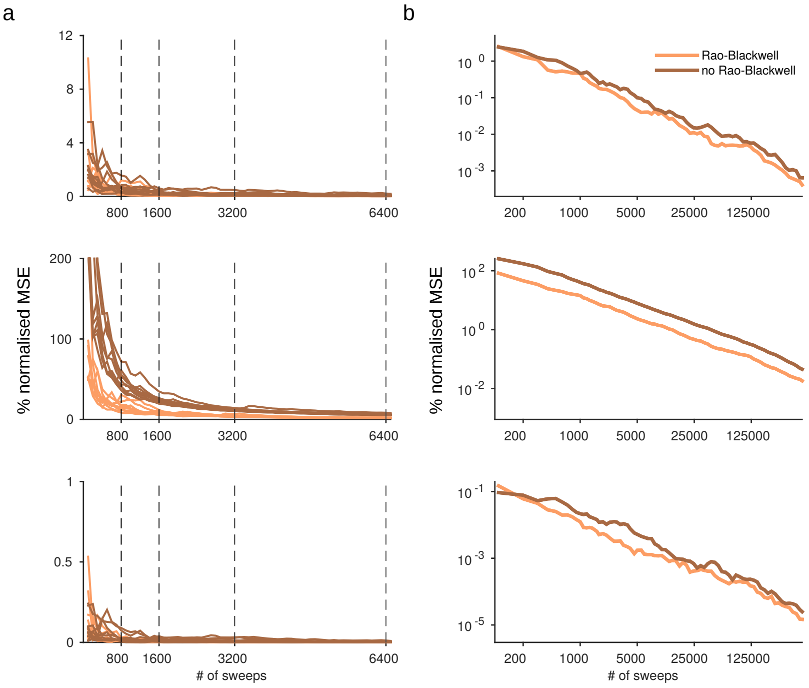

Figure S1. Impact of Rao-Blackwellisation a) Comparison of normalised MSE between Rao-Blackwellised and non-Rao-Blackwellised Gibbs sampling, as a function of MCMC chain length, on the 10 subpopulations of size $n = 100$ used in the paper. Top: means, i.e. first-order moments $\mathrm{E}_\lambda[\mathbf{x}_i]$, Center: covariances $\mathrm{cov}_\lambda(\mathbf{x}_i, \mathbf{x}_j)$, Bottom: population-spike count features. No Rao-Blackwellization was used for population-spike count features $P(K = k|\lambda)$. Vertical lines and horizontal axis ticks mark Markov chain lengths used for computing the 1st, 1001st 2001st, ... updates of parameter entries λ_l during training the K-pairwise models to data. All MSEs in this figure are computed as errors between estimated firing rates / covariances / $P(K)$ at given chain length versus the average of the estimates obtained after 10^6 sweeps. **b)** Behavior of MSEs for large MCMC chain lengths. Traces are averages over the 10 traces from panel a.

The only relevant expectations are w.r.t. the data distribution and $P(\mathbf{x}|\lambda^{old})$, i.e. the current parameter estimate. The term $\mathrm{E}_{\lambda^{old}}[\exp\left((\lambda^{new} - \lambda^{old})^T f(\mathbf{x})\right)]$ can be simplified when restricting the update vector $\lambda^{new} - \lambda^{old}$ to be non-zero only in selected components. In the simplest case, only a single component λ_l is updated. In this case, the fact that all components of the K-pairwise feature function $f(\mathbf{x})$ are binary, allows to move the exponent out of the expected value, a trick used by [51]:

The resulting single-coordinate updates only require the feature moments $\mathrm{E}_{\lambda^{old}}[f_l(\mathbf{x})]$:

$$\begin{aligned}\mathrm{E}_{\lambda^{old}}[\exp\left((\lambda^{new} - \lambda^{old})^T f(\mathbf{x})\right)] &= \mathrm{E}_{\lambda^{old}}[\exp((\lambda_l^{new} - \lambda_l^{old}) f_l(\mathbf{x})] \\ &= \mathrm{E}_{\lambda_{old}}[1 + (\exp(\lambda_l^{new} - \lambda_l^{old}) - 1) f_l(\mathbf{x})] \\ &= 1 + (\exp(\lambda_l^{new} - \lambda_l^{old}) - 1)\mathrm{E}_{\lambda_{old}}[f_l(\mathbf{x})]\end{aligned}$$

Equation (16) can now be solved analytically for the single free component λ_l^{new} that maximises the change in log-likelihood. A closed-form optimal solution is still possible when adding an $l1$-penalty to the log-likelihood [51]. We use this $l1$-regularised variant to calculate the possible gain in penalized log-likelihood for each possible update of the single-cell (h_i) and pairwise (J_{ij}) feature moments $\mathrm{E}[x_i]$ and $\mathrm{E}[x_i x_j]$, and then perform the update which yield the largest gain.

If we instead allow more than a single component l of the update $\lambda_l^{new} - \lambda_l^{old}$ to be non-zero, we in general would have to deal with the term

$$\mathrm{E}_{\lambda_{old}}\left[\prod_{l \in J}[1 + (\exp(\lambda_l^{new} - \lambda_l^{old}) - 1) f_l(\mathbf{x})]\right]$$

which requires the higher-order moments $\mathrm{E}_{\lambda_{old}}\left[\prod_{l \in I} f_l(\mathbf{x})\right]$ for all $I \subseteq J$ and $J \subseteq \{1, \ldots, n\}$ being the index set of components that are not set to zero.

The population spike count features $f_k(\mathbf{x})$, however, are mutually exclusive (only one of the n+1 features can be non-zero at any time), and therefore we can all parameters of V jointly, and still pull the expectation term outside of the expectation. For the population-spike count features $f_k(\mathbf{x})$, hereafter collectively called $f^V(x) \in \{0,1\}^{n+1}$, all such terms of order $||I|| > 1$ are zero due to the sparsity of $f^V(x)$. When restricting the current parameter update of λ to only update components corresponding to V, we have

$$\Delta L(V^{new}, V^{old}) = (V^{new} - V^{old})^T \left(\frac{1}{M}\sum_{m=1}^{m} f^V(\mathbf{x}^{(m)})\right) - \mathrm{E}_{\lambda^{old}}[\exp((V^{new} - V^{old})^T f^V(\mathbf{x}))]$$

and

$$\begin{aligned}
\mathrm{E}_{\lambda^{old}}[\exp\left((V^{new} - V^{old})^T f^K(\mathbf{x})\right)] &= \sum_{\mathbf{x}} \exp\left((V^{new} - V^{old})^T f^K(\mathbf{x})\right) P(\mathbf{x}|\lambda^{old}) \\
&= \sum_{k=0}^{n} \sum_{\mathbf{x}:\sum_i x_i = k} \exp\left((V_k^{new} - V_k^{old} f_k^K(\mathbf{x})\right) P(\mathbf{x}|\lambda^{old}) \\
&= \sum_{k=0}^{n} \exp\left((V_k^{new} - V_k^{old}) f_k^K(\mathbf{x})\right) \sum_{\mathbf{x}:\sum_i x_i = k} P(\mathbf{x}|\lambda^{old}) \\
&= \sum_{k=0}^{n} \exp\left((V_k^{new} - V_k^{old}) f_k^K(\mathbf{x})\right) P(k|\lambda^{old})
\end{aligned}$$

We obtained estimates of the values of $P(k|\lambda^{old}) = \mathrm{E}_{\lambda^{old}}[f_k(\mathbf{x})]$ from the MCMC sample using the indicator functions $f_k(\mathbf{x})$, and optimising w.r.t. V_k^{new}, $k \in \{1,\ldots,n\}$ using gradient-based methods [**?**].

In summary, our update-scheme for maximising the log-likelihood proceeds as follows: For a given parameter vector λ^{old}, we first estimate the expectation of the feature functions $f_i(\mathbf{x})$, $f_{ij}(\mathbf{x})$ and $f_k(\mathbf{x})$ using a running average over an MCMC sampling and Rao-Blackwellization. We then calculate, for each possible single-neuron parameter h_i and each possibly pairwise term J_{ij} the gain in penalised log-likelihood that we would get from updating it, using methods as described above and derived in [51]. We additionally compute the gain in penalised log-likelihood that would result from optimising all n of the free V parameters jointly, using a convex optimization. Finally, we choose the update that brings the largest gain, and either update a single h_i, a single J_{ij}, or all V parameters. Subsequently, we again estimate the new feature functions using MCMC sampling given the current estimate of $\lambda^{old} \leftarrow \lambda^{new}$ before we update again. We initialised the algorithm assuming independent neurons (i.e. setting each h_i using the firing rate of each neuron, and leaving J and V zero). The algorithm then typically first updated all V parameters, before proceeding to jump between different J, h and V updates.

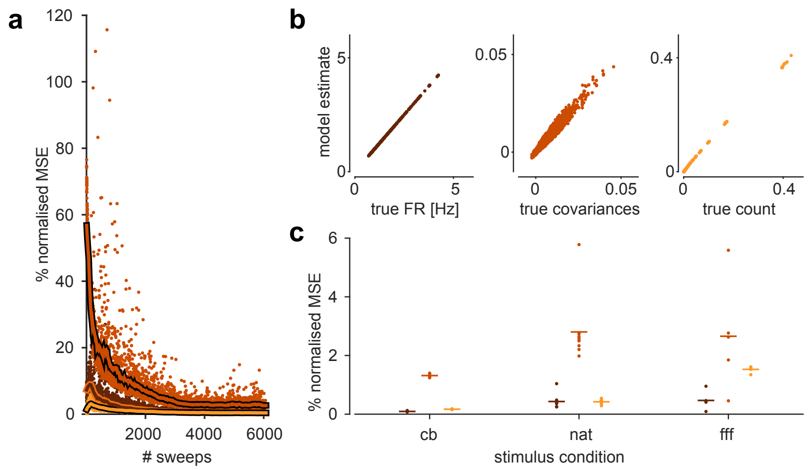

Figure S2. Quality of fits for K-pairwise maximum entropy model across multiple populations and stimulus conditions a) Normalised MSEs for firing rates, covariances and $P(K)$ during parameter learning. Error values collapses across 10 subpopulations at $n = 100$, fit to simulated activity in response to natural images, one point for each displayed iteration and each subpopulation. Lines are moving averages (smoothing kernel width = 150 param. updates). **b)** Quality of fit after parameter learning. Data vs. model estimates for firing rates, covariances and $P(K)$, collapsed over all 10 subpoplations with size $n = 100$. **c)** Quality of fit for different stimulus types. Normalised MSEs after maximum entropy model fitting shown for 10 subpopulations for natural images (nat) and 5 subpopulations each for checkerboard (cb) and full-field flicker (fff). All subpopulations of size $n = 100$. Vertical bars give averages. Colours as in **a), b)**.

S2 Supplementary Text: Specific heat in simple models

We refer to a maxmimum entropy model as 'flat' if it is fully specified by the population spike count distribution $P(\sum_{i=1}^{n} x_i = k)$, i.e. the model class studied in [21, 29, 30]. In this model class, all neurons have the same firing rate μ and pairwise correlation ρ. As neuron identities become interchangeable, all $\binom{n}{k}$ possible patterns $\mathbf{x}$ with $\sum_{i=1}^{n} = k$ are assigned the same probability $P(k) = P(\mathbf{x})\binom{n}{k}$. In flat models, all relevant population properties can be computed from summing over $n + 1$ different spike counts, and one never has to (explicitly) sum over the entire 2^n possible spike patterns.

S2.1 A non-critical special case: Independent neurons

A special case of a flat model is an independent model in which all neurons have the same firing rates and zero correlations. Assuming independent spiking for each of the n neurons and a shared probability $q \in [0, 1]$ to fire in a time bin, the distribution of population spike counts $K = \sum_{i=1}^{n} x_i$ is given by a binomial distribution,

$$P(\mathbf{x}|q) = q^k(1-q)^{n-k}$$
$$P(k|q) = \binom{n}{k} q^k(1-q)^{n-k}$$

To compute specific heat capacities for the underlying neural population of size n, we can rewrite the binomial distribution in maximum entropy form

$$
\begin{aligned}
P(\mathbf{x}|V) &= \frac{1}{Z(V)} \exp\left(V_k\right) \\
P(k|V) &= \frac{1}{Z(V)} \binom{n}{k} \exp\left(V_k\right)
\end{aligned}
$$

Re-introducing parameters V_k, $k = 0, \ldots, n$, we find

$$
\begin{aligned}
V_k &= \log P(k|q) - \log \binom{n}{k} + \log Z(V) \\
&= k \log(q) + (n-k) \log(1-q))
\end{aligned}
$$

and for the heat capacity, we get

$$
\begin{aligned}
\mathrm{Var}[\log P(x|V)] &= \mathrm{Var}[k \log(q) + (n-k) \log(1-q)] \\
&= (\log(q) - \log(1-q))^2 \ \mathrm{Var}[k]
\end{aligned}
$$

The binomial variance is $\mathrm{Var}[k] = nq(1-q)$. We plug this in and see that at unit temperature $T = 1$, the specific heat is given by

$$
c(T=1) = \frac{1}{nT^2} \mathrm{Var}[\log P(\mathbf{x}|V)] = q(1-q)(\log(q) - \log(1-q))^2 \tag{17}
$$

which is independent of population size n.

When explicitly introducing temperatures other than $T = 1$, we add a factor $\frac{1}{T} = \beta$ that scales the parameters V and renormalise, yielding

$$
P(k|V,T) = \frac{1}{Z(\beta V)} \binom{n}{k} \exp(\beta V_k)
$$

where V_k, $k = 0$, ..., n is defined w.r.t. q as above. This is the same functional form as was given for the binomial distribution at $T = 1$, with only parameters V being replaced by βV. We can also go back to the standard binomial parametrisation with $q_\beta = \frac{q^\beta}{q^\beta + (1-q)^\beta}$ and obtain

$$
P(k|V,T) = \binom{n}{k} q_\beta^k (1-q_\beta)^{(n-k)}
$$

Changing the temperature $T = \frac{1}{\beta}$ retains the binomial form of the population model, and we can generalise the expression for the specific heat (17) of the independent flat model for any temperature T to be

$$
c(T) = \frac{1}{T^2} q_\beta (1-q_\beta)(\log(q_\beta) - \log(1-q_\beta))^2
$$

which again is independent of the population size n. The independent flat model is a case that does not show divergent specific heat, and for which the peak of the heat is not necessarily at unit temperature. Next, we will derive why this makes the binomial model one of only two non-critical special cases.

S2.2 Aside: Asymptotic entropy in flat models

To calculate the variance of log-probabilities, we first need the mean log-probability, i.e. the (negative) entropy.

Entropy: Recalling that $P(k) = P(\mathbf{x})\binom{n}{k}$, the entropy of the flat model for general $P(k)$ can be written as

$$\begin{aligned}H_n &= -\sum_x P(\mathbf{x}) \log P(\mathbf{x}) \\ &= -\sum_k \sum_{\mathbf{x}:\sum_i x_i = k} P(\mathbf{x}) \log P(\mathbf{x}) \\ &= -\sum_k P(k) \left(\log P(k) - \log \binom{n}{k} \right)\end{aligned}$$

Thus, the entropy of a flat model is

$$H_n = -\sum_k P(k) \left(\log P(k) - \log \binom{n}{k} \right)$$

Asymptotic entropy: We assume that $P(k)$ has a limiting distribution $f(r)$, where $r \in [0, 1]$ is the probability density of a proportion of r neurons spiking simultaneously. Therefore, for large n

$$\begin{aligned}H_n &= -\sum_k P(k) \left(\log P(k) - \log \binom{n}{k} \right) \\ &\approx -\sum_k \frac{1}{n} f\left(\frac{k}{n}\right) \left(\log P(k) - \log \binom{n}{k} \right) \\ &\approx -\int_0^1 f(r) \left(\log \frac{f(r)}{n} - \log \binom{n}{nr} \right) dr \\ &= -\log(n) \int_0^1 f(r) \log f(r) dr + n \int_0^1 f(r)\eta(r) dr\end{aligned}$$

Here, we used the fact that, for large n,

$$\log \binom{n}{nr} \approx n \left(-r \log r - (1-r) \log(1-r) \right) =: n\eta(r) \tag{18}$$

As the first term only grows with $\log(n)$, and the second with n, we get that the entropy of a flat model, for large n, is given by

$$H_n = n \int_0^1 f(r)\eta(r) dr =: nh \tag{19}$$

S2.3 Asymptotic specific heat in flat models at unit temperature

Next, we calculate the specific heat, first exactly and then for large n, and finally for weakly correlated models:

First, the specific heat is given by

$$\begin{aligned}c(T=1) = \frac{1}{n}\mathrm{Var}[\log P(\mathbf{x})] &= \frac{1}{n} \sum_x P(\mathbf{x}) \left(\log P(\mathbf{x}) - \mathrm{E}[\log P(\mathbf{x})] \right)^2 \\ &= \frac{1}{n} \sum_k P(k) \left(\log P(k) - \log \binom{n}{k} - \mathrm{E}[\log P(\mathbf{x})] \right)^2\end{aligned}$$

Using $\mathrm{E}[\log P(\mathbf{x})] = -H_n$, we get that

$$c(T=1) = \frac{1}{n}\sum_k P(k)\left(\log P(k) - \log\binom{n}{k} + H_n\right)^2 \text{ or}$$
$$= \frac{1}{n}\sum_k P(k)\left(\log P(k) - \log\binom{n}{k}\right) - \frac{1}{n}H_n^2$$

For large n, we have that $P(k) \approx \frac{1}{n} f\left(\frac{k}{n}\right)$. We get that

$$\begin{aligned}
c(T=1) &= \frac{1}{n}\sum_k P(k)\left(\log P(k) - \log\binom{n}{k} + H_n\right)^2 \\
&\approx \frac{1}{n}\sum_k \frac{1}{n} f\left(\frac{k}{n}\right)\left(\log\left(\frac{1}{n} f\left(\frac{k}{n}\right)\right) - \log\binom{n}{k} + H_n\right)^2 \\
&\approx \frac{1}{n}\int_0^1 f(r)\left(\log f(r) - \log n - \log\binom{n}{nr} + H_n\right)^2 dr \\
&\approx \frac{1}{n}\int_0^1 f(r)\left(\log f(r) - \log n - n\eta(r) + nh_n\right)^2 dr \\
&= \frac{1}{n}\int_0^1 f(r)\Big(\left(\log f(r) - \log n\right)^2 + n^2\left(\eta(r) - h_N\right)^2 + 2n\left(\log f(r) - \log n\right)\left(h_n - \eta(r)\right)\Big) dr \\
&= \frac{1}{n}\int_0^1 f(r)\left(\log^2 f(r) + \log f(r)\left(2n\left(h_n - \eta(r)\right) - 2\log n\right)\right) dr \\
&+ \frac{1}{n}\int_0^1 f(r)\left(\log^2 n - n^2\left(h_n - \eta(r)\right)\right) dr
\end{aligned}$$

For large n, this integral is dominated by the term in n^2, and thus the specific heat is asymptotically given by

$$c(T=1) = n\int_0^1 f(r)\left(\eta(r) - h\right)^2 dr \qquad (20)$$

Therefore, in general, the specific heat grows linearly, and hence diverges (see Fig S3). The only exception to this are models for which $\eta(r) - h_n = 0$ for almost all r. This happens if $f(r)$ is a delta-distribution, $f(r) = \delta(r-\mu)$, in which case $h_n = \eta(\mu)$ and therefore the integral vanishes. This occurs whenever the pairwise correlations do not grow proportionally with n^2, as then the variance of the population spike count collapses in the limit. One such special case is the binomial distribution over k, as already demonstrated above using a more direct approach. There is a second special case, namely if $f(r)$ is a combination of two δ-peaks at μ and $1-\mu$ (See [21] for details)– this special case corresponds to a flat Ising model.

S2.4 In flat models, specific heat does not diverge for temperatures which are not equal to 1:

Above we showed that at unit temperature, the specific heat for flat models (almost) always diverges. Now, we show that this is NOT true for any other temperature. This explains that, for any $f(r)$, we will find that the unit temperature is 'special'.

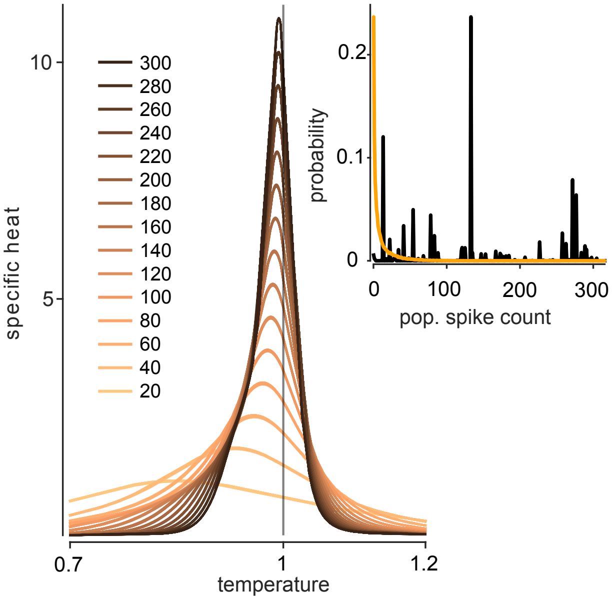

Figure S3. Diverging specific heat for a non-natural spike-count distribution The values of the population spike count distribution $P(K)$ obtained from the retinal simulation with $N = 316$ in response to natural image stimulation (orange, inset) were shuffled (black trace, inset) across K, to yield a 'pathological' $P(K)$. We simulated data for this $P(K)$ from a flat model, and subsampled subpopulations of size $n = 20, \ldots, 300$. The specific heat traces computed from this data also diverges and has a peak at unit temperature

First, we calculate the spike-count distribution at any inverse temperature β:

$$P_\beta(\mathbf{x}) = \frac{1}{Z_\beta} P(\mathbf{x})^\beta$$

$$P_\beta(k) = \frac{1}{Z_\beta} \binom{n}{k}^{1-\beta} P(k)^\beta$$

For large n,

$$\begin{aligned} f_\beta(r) &\approx n P_\beta(rn) \\ &= \frac{n}{Z_\beta} \binom{n}{nr}^{1-\beta} P(rn)^\beta \\ &\approx \frac{n}{Z_\beta} \exp\left(n(1-\beta)\eta(r)\right) P^\beta(rn) \end{aligned}$$

For large populations, this expression is dominated by the exponential term $\exp\left(n(1-\beta)\eta(r)\right)$. For $\beta < 1$, the exponential term is in turn dominated by the mode of $\eta(r)$, which is at $r = \frac{1}{2}$. Thus, for $\beta < 1$, $f_\beta(r) = \delta(r - \frac{1}{2})$, a delta-peak at $r = \frac{1}{2}$.

Conversely, for $\beta > 1$, the argument of the exponential has its peaks at $r = 0$ and $r = 1$, and therefore $f_\beta(r) = \frac{1}{2}\delta(r-1) + \frac{1}{2}\delta(r-0)$. In this case, we also have that the integral in the specific heat vanishes, and that the specific heat does not diverge.

S3 Specific heat divergence rate in flat models as function of correlation strength

In the next two sections, we will derive analytic expressions to predict the specific heat divergence rate in flat models as a function of the correlation strength within the population. Starting out from eq. (20), we will use two different approximations to $f(r)$ that will each yield results that allow us to better understand the behavior of the specific heat at unit temperature $c(T = 1)$ in flat models.

S3.1 Asymptotic entropy and specific heat in weakly correlated flat models:

Next, we examine entropy and specific heat in models with weak correlations. If the model is weakly correlated and its mode is not at 0 or 1 we can assume it to be approximately Gaussian with mean μ and variance σ^2,

$$f(r) = \frac{1}{Z} \exp\left(-\frac{1}{2\sigma^2}(r-\mu)^2\right).$$

We first calculate the entropy: We expand $\eta(r)$ to second order around μ,

$$\begin{aligned}
\eta(r) &= \eta(\mu+\delta) = \eta(\mu) + \eta'(\mu)\delta + \frac{\delta^2}{2}\eta''(\mu) + ..., \text{where} \\
\eta'(r) &= \log\left(\frac{1-r}{r}\right) \\
\eta''(r) &= \frac{-1}{r(1-r)}, \text{ so} \\
\eta(\mu+\delta) &= \eta(\mu) + \delta\log\left(\frac{1-\mu}{\mu}\right) - \frac{\delta^2}{2\mu(1-\mu)} + ... \\
&=: \alpha + \delta\beta + \delta^2\gamma
\end{aligned}$$

Thus, the asymptotic entropy-rate is given by

$$\begin{aligned}
h &= \int f(r)\eta(r)dr \\
&= \alpha + 0\beta + \gamma\sigma^2 \\
&= \eta(\mu) - \frac{1}{2\mu(1-\mu)}\sigma^2
\end{aligned}$$

We further investigate the variance, again neglecting all terms which are of higher order than 2, obtaining

$$\begin{aligned}
(\eta(\mu+\delta) - h)^2 &= \left((\alpha - h) + \beta\delta + \gamma\delta^2\right)^2 \\
&= (\alpha-h)^2 + \delta^2\beta^2 + 2(\alpha-h)\beta\delta + 2(\alpha-h)\gamma\delta^2 + 2(\alpha-h)\gamma\delta^2 + \dots \\
&= (\alpha-h)^2 + \delta\left(2(\alpha-h)\beta\right) + \delta^2\left(\beta^2 + 2(\alpha-h)\gamma\right) + \dots
\end{aligned}$$

Integrating this expression over $f(r)$, and dropping all terms in σ which are of order higher than 2, we get

$$\begin{aligned}\int f(r)(\eta(r)-h)^2 &= (\alpha-h)^2+\sigma^2\left(\beta^2+2(\alpha-h)\gamma\right)\\ &= \frac{\sigma^4}{\mu^2(1-\mu)^2}+\sigma^2\left(\log^2\left(\frac{1-\mu}{\mu}\right)-\frac{\sigma^2}{\mu^2(1-\mu)^2}\right)\\ &\approx \sigma^2\log^2\left(\frac{1-\mu}{\mu}\right)\end{aligned}$$

In summary, we arrive at

$$\begin{aligned}&c(T=1)=n\sigma^2\log^2\left(\frac{1-\mu}{\mu}\right) \text{ and, for small } \mu,\\ &c(T=1)=n\sigma^2\log^2(\mu)\end{aligned}$$

In other words, a population of a given size n at fixed firing rate μ that has a high specific heat is simply a population which is very correlated. Inspecting the equations above, we see that the final results do not critically depend on the Gaussian assumption—the only requirement for the calculation to be accurate is that the distribution is reasonably peaked around its mean.

S3.2 Asymptotic specific heat in the beta-binomial population model

For the beta-binomial model, we assume $f(r)$ to be given by a beta distribution, i.e.

$$f(r)=\frac{1}{B(\alpha,\beta)}r^{\alpha-1}(1-r)^{\beta-1}.$$

Such $f(r)$ arise for large populations when the population spike count k is described by a beta-binomial distribution, and the choice for the beta distribution as a model for $f(r)$ was motivated by the successful application of beta-binomial models $P(k|\alpha,\beta)$ to our simulated RGC activity (see Fig. S4).

For beta-distributed r, we have

$$\begin{aligned}\mathrm{E}[r]&=\frac{\alpha}{\alpha+\beta},\\ \mathrm{Var}[r]&=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},\\ \mathrm{E}[\log r]&=\gamma(\alpha)-\gamma(\alpha+\beta),\end{aligned}$$

where γ denotes the digamma function.

The entropy can be calculated using known results on the expectation of the log,

$$h=\gamma(\alpha+\beta+1)-\frac{\alpha}{\alpha+\beta}\gamma(\alpha+1)-\frac{\beta}{\alpha+\beta}\gamma(\beta+1)$$

For the specific heat at unit temperature according to equation (20), we however also require the expected values

$$\mathrm{E}[r^2\log^2 r],\mathrm{E}[(1-r)^2\log^2(1-r)],\mathrm{E}[r(1-r)\log r\log(1-r)]$$

i.e.

$$\mathrm{E}[r^k(1-r)^l\log^m r\log^n(1-r)]=\int_0^1 f(r)\{r^k(1-r)^l\log^m r\log^n(1-r)\}dr \qquad (21)$$

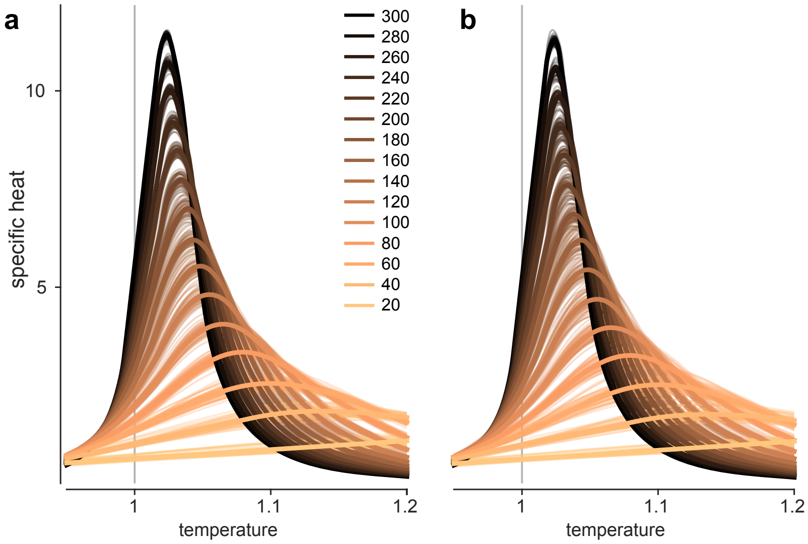

Figure S4. (No) Influence of beta-binomial approximation on heat capacity Specific heat capacities computed from population spike count distributions $P(K)$. Spike count distributions for population sizes $n = 20, \ldots, 300$ were obtained from 50 uniformly drawn subpopulations each. Simulated retinal activity was taken from of the retina simulation with in total $N = 316$ RGCs that responded to natural image stimulation. Resulting specific heat traces computed from beta-binomial approximations to the spike count distributions (left) and from raw $P(K)$ (right) do not display visible differences.

under beta-binomial distribution $f(r)$, where $k, l, m, n \in \{0, 1, 2\}$.

We begin the derivation of these terms by observing that

$$\begin{aligned}
u(m,n)(r, \alpha+k, \beta+l) &= \log(r)^m r^{(\alpha+k-1)} \log(1-r)^n (1-r)^{(\beta+l-1)} \\
\frac{\delta}{\delta\alpha} u_{(m,n)}(r, \alpha+k, \beta+l) &= \log(r)^{m+1} r^{(\alpha+k-1)} \log(1-r)^n (1-r)^{(\beta+l-1)} \\
&= u_{(m+1,n)}(r, \alpha+k, \beta+l) \\
\frac{\delta}{\delta\beta} u_{(m,n)}(r, \alpha+k, \beta+l) &= \log(r)^m r^{(\alpha+k-1)} \log(1-r)^{n+1} (1-r)^{(\beta+l-1)} \\
&= u_{(m,n+1)}(r, \alpha+k, \beta+l)
\end{aligned}$$

for any $k, l \in \mathbb{N}$. Note that the exponents k, l are readily absorbed into new effective beta distribution parameters $\alpha' = \alpha + k$, $\beta' = \beta + l$.

The triplets $(u_{(m,n)}, u_{(m+1,n)}, u_{(m,n+1)})$ for any $m, n \in \mathbb{N}$ recursively express the integrands of (21) as continuous derivatives, which allows us to repeatedly apply Leibniz' rule to the integral. We first deal with $\mathrm{E}[r^k \log^m r]$, where $m = k = 2$, $n = l = 0$, $\alpha' = \alpha + 2$, $\beta' = \beta$, which is the first of the three expected values we need to compute the specific heat at unit temperature:

$$\begin{aligned}
\mathrm{Beta}(\alpha,\beta)\mathrm{E}[r^2\log^2 r] &= \int_0^1 r^{\alpha-1}(1-r)^{\beta-1}\log^2(r)r^2dr \\
&= \int_0^1 \frac{\delta^2}{\delta\alpha^2}\{r^{\alpha+1}(1-r)^{\beta-1}\}dr \\
&= \int_0^1 \frac{\delta}{\delta\alpha}\{\frac{\delta}{\delta\alpha}\{r^{\alpha+1}(1-r)^{\beta-1}\}\}dr \\
&= \frac{\delta}{\delta\alpha}\int_0^1 \frac{\delta}{\delta\alpha}\{r^{\alpha+1}(1-r)^{\beta-1}\}dr \\
&= \frac{\delta^2}{\delta\alpha^2}\int_0^1 r^{\alpha+1}(1-r)^{\beta-1}dr \\
&= \frac{\delta^2}{\delta\alpha^2}\mathrm{Beta}(\alpha+2,\beta)
\end{aligned}$$

The first two derivatives of $\mathrm{Beta}(\alpha',\beta')$ w.r.t. α are given by

$$\frac{\delta}{\delta\alpha}\mathrm{Beta}(\alpha',\beta') = \mathrm{Beta}(\alpha',\beta')(\psi_0(\alpha')-\psi_0(\alpha'+\beta')) \text{ and}$$
$$\frac{\delta^2}{\delta\alpha^2}\mathrm{Beta}(\alpha',\beta') = \mathrm{Beta}(\alpha',\beta')\left((\psi_0(\alpha')-\psi_0(\alpha'+\beta'))^2+\psi_1(\alpha')-\psi_1(\alpha'+\beta')\right).$$

We obtain the m-th derivative also for $m > 2$ using an iterative rule. The beta-binomial normaliser $\mathrm{Beta}(\alpha',\beta')$ furthermore cancels out with the denominator $\mathrm{Beta}(\alpha,\beta)$ of the original beta distribution through

$$\mathrm{Beta}(\alpha+k,\beta+l) = \frac{\prod_{i=0}^{k-1}(\alpha+i)\prod_{j=0}^{l-1}(\beta+j)}{\prod_{i=0}^{k+l-1}(\alpha+\beta+i)}\mathrm{Beta}(\alpha,\beta)$$

Combining the previous results gives

$$\begin{aligned}
\mathrm{E}[r^2\log^2 r] &= \frac{1}{\mathrm{Beta}(\alpha,\beta)}\int_0^1 r^{\alpha-1}(1-r)^{\beta-1}log^2(r)r^2dr \qquad (22)\\
&= \frac{1}{\mathrm{Beta}(\alpha,\beta)}\frac{\delta^2}{\delta\alpha^2}\mathrm{Beta}(\alpha+2,\beta) \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\frac{1}{\mathrm{Beta}(\alpha+2,\beta)}\frac{\delta^2}{\delta\alpha^2}\mathrm{Beta}(\alpha+2,\beta) \\
&= \frac{\alpha(\alpha+1)\left((\psi_0(\alpha+2)-\psi_0(\alpha+\beta+2))^2+\psi_1(\alpha+2)-\psi_1(\alpha+\beta+2)\right)}{(\alpha+\beta)(\alpha+\beta+1)}.
\end{aligned}$$

For $m=2$, $k=1$, $n,l=0$ the result

$$E[r\log^2 r] = \frac{\alpha}{\alpha+\beta}[(\psi_0(\alpha+1)-\psi_0(\alpha+\beta+1))^2+\psi_1(\alpha+1)-\psi_1(\alpha+\beta+1)]$$

is identical to the one from [**?**] in the appendix A.3, eq. (28).

We have $\mathrm{Beta}(\alpha,\beta)=\mathrm{Beta}(\beta,\alpha)$, i.e. the above equations hold symmetrically for α and β interchanged, and n,l instead of m,k. This gives us the second required term to compute the specific heat at unit temperature,

$$\begin{aligned}
&\mathrm{E}[(1-r)^2\log^2(1-r)] \qquad (23)\\
&= \frac{\beta(\beta+1)\left((\psi_0(\beta+2)-\psi_0(\alpha+\beta+2))^2+\psi_1(\beta+2)-\psi_1(\alpha+\beta+2)\right)}{(\alpha+\beta)(\alpha+\beta+1)}
\end{aligned}$$

Including derivatives w.r.t. both α and β, we more generally arrive at

$$\mathrm{E}[\log(r)^m r^k \log(1-r)^n (1-r)^l] = \frac{\prod_{i=0}^{k-1}(\alpha+i)\prod_{j=0}^{l-1}(\beta+j)}{\prod_{i=0}^{k+l-1}(\alpha+\beta+i)} g_{(m,n)}(\alpha+k, \beta+l).$$

We get recursive formulas for $g_{(m,n)}$, starting at $g_{(0,0)}(\alpha, \beta) = 1$:

$$\begin{aligned} g_{(m+1,n)}(\alpha,\beta) &= (\psi_0(\alpha) - \psi_0(\alpha+\beta))\, g_{(m,n)}(\alpha,\beta) + \frac{\delta}{\delta\alpha} g_{(m,n)}(\alpha+\beta) \\ g_{(m,n+1)}(\alpha,\beta) &= (\psi_0(\beta) - \psi_0(\alpha+\beta))\, g_{(m,n)}(\alpha,\beta) + \frac{\delta}{\delta\beta} g_{(m,n)}(\alpha+\beta). \end{aligned}$$

To compute $c(T=1)$, we still require the case of $m = k = n = l = 1$ given by

$$\mathrm{E}[r(1-r)\log(r)\log(1-r)] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} g_{(1,1)}(\alpha+1, \beta+1) \tag{24}$$

with

$$\begin{aligned} g_{(1,1)}(\alpha+k,\beta+l) = \psi_0(\alpha+k)\psi_0(\beta+l) - \psi_0(\alpha+\beta+k+l)\left(\psi_0(\alpha+k) + \psi_0(\beta+l)\right) \\ + \psi_0(\alpha+\beta+k+l)^2 - \psi_1(\alpha+\beta+k+l). \end{aligned} \tag{25}$$

Combining the results of equations (22), (23), (24), (25) with eq. (20), we arrive at

$$\begin{aligned} \frac{c(T=1)}{n} &= \int_0^1 f(r)\,(\eta(r) - h)^2\, dr \\ &= \frac{\alpha(\alpha+1)\psi_1(\alpha+1) + \beta(\beta+1)\psi_1(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\ &+ \frac{\alpha\beta\left(\psi_0(\alpha+1) - \psi_0(\beta+1)\right)^2}{(\alpha+\beta)^2(\alpha+\beta+1)} - \psi_1(\alpha+\beta+1). \end{aligned}$$