Thermodynamics and signatures of criticality in a network of neurons

Gašper Tkačik[a], Thierry Mora[b], Olivier Marre[c], Dario Amodei[d,e], Stephanie E. Palmer[d,f], Michael J. Berry II[e,g], and William Bialek[d,h,1]

[a]Institute of Science and Technology Austria, A-3400 Klosterneuburg, Austria; [b]Laboratoire de Physique Statistique, CNRS, Université Pierre et Marie Curie (UPMC) and l'École Normale Supérieure, 75231 Paris Cedex 05, France; [c]Institut de la Vision, UMRS 968 UPMC, INSERM, CNRS U7210, F-75012 Paris, France; [d]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544; [e]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544; [f]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; [g]Department of Molecular Biology, Princeton University, Princeton, NJ 08544; and [h]Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, New York, NY 10016

Contributed by William Bialek, August 4, 2015 (sent for review July 22, 2014)

The activity of a neural network is defined by patterns of spiking and silence from the individual neurons. Because spikes are (relatively) sparse, patterns of activity with increasing numbers of spikes are less probable, but, with more spikes, the number of possible patterns increases. This tradeoff between probability and numerosity is mathematically equivalent to the relationship between entropy and energy in statistical physics. We construct this relationship for populations of up to $N = 160$ neurons in a small patch of the vertebrate retina, using a combination of direct and model-based analyses of experiments on the response of this network to naturalistic movies. We see signs of a thermodynamic limit, where the entropy per neuron approaches a smooth function of the energy per neuron as N increases. The form of this function corresponds to the distribution of activity being poised near an unusual kind of critical point. We suggest further tests of criticality, and give a brief discussion of its functional significance.

entropy | information | neural networks | Monte Carlo | correlation

Our perception of the world seems a coherent whole, yet it is built out of the activities of thousands or even millions of neurons, and the same is true for our memories, thoughts, and actions. It is difficult to understand the emergence of behavioral and phenomenal coherence unless the underlying neural activity also is coherent. Put simply, the activity of a brain—or even a small region of a brain devoted to a particular task—cannot be just the summed activity of many independent neurons. How do we describe this collective activity?

Statistical mechanics provides a language for connecting the interactions among microscopic degrees of freedom to the macroscopic behavior of matter. It provides a quantitative theory of how a rigid solid emerges from the interactions between atoms, how a magnet emerges from the interactions between electron spins, and so on (1, 2). These are all collective phenomena: There is no sense in which a small cluster of molecules is solid or liquid; rather, solid and liquid are statements about the joint behaviors of many, many molecules.

At the core of equilibrium statistical mechanics is the Boltzmann distribution, which describes the probability of finding a system in any one of its possible microscopic states. As we consider systems with larger and larger numbers of degrees of freedom, this probabilistic description converges onto a deterministic, thermodynamic description. In the emergence of thermodynamics from statistical mechanics, many microscopic details are lost, and many systems that differ in their microscopic constituents nonetheless exhibit quantitatively similar thermodynamic behavior. Perhaps the oldest example of this idea is the "law of corresponding states" (3).

The power of statistical mechanics to describe collective, emergent phenomena in the inanimate world led many people to hope that it might also provide a natural language for describing networks of neurons (4–6). However, if one takes the language of statistical mechanics seriously, then as we consider networks with larger and larger numbers of neurons, we should see the emergence of something like thermodynamics.

Theory

At first sight, the notion of a thermodynamics for neural networks seems hopeless. Thermodynamics is about temperature and heat, both of which are irrelevant to the dynamics of these complex, nonequilibrium systems. However, all of the thermodynamic variables that we can measure in an equilibrium system can be calculated from the Boltzmann distribution, and hence statements about thermodynamics are equivalent to statements about this underlying probability distribution. It is then only a small jump to realize that all probability distributions over N variables can have an associated thermodynamics in the $N \to \infty$ limit. This link between probability and thermodynamics is well-studied by mathematical physicists (7), and has been a useful guide to the analysis of experiments on dynamical systems (8, 9).

To be concrete, consider a system with N elements; each element is described by a state σ_i, and the state of the entire system is $\sigma \equiv \{\sigma_1, \sigma_2, \cdots, \sigma_N\}$. We are interested in the probability $P(\sigma)$ that we will find the system in any one of its possible states. It is natural to think not about the probability itself but about its logarithm,

$$E(\sigma) = -\ln P(\sigma). \quad [1]$$

In an equilibrium system, this is precisely the energy of each state (in units of $k_B T$), but we can define this energy for any probability distribution. As discussed in detail in *Supporting Information*, all

Significance

The activity of a brain—or even a small region of a brain devoted to a particular task—cannot be just the summed activity of many independent neurons. Here we use methods from statistical physics to describe the collective activity in the retina as it responds to complex inputs such as those encountered in the natural environment. We find that the distribution of messages that the retina sends to the brain is very special, mathematically equivalent to the behavior of a material near a critical point in its phase diagram.

Author contributions: G.T., T.M., O.M., D.A., S.E.P., M.J.B., and W.B. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper. This is a collaboration between theorists (G.T., T.M., S.E.P., and W.B.) and experimentalists (O.M., D.A., and M.J.B.). All authors contributed to all aspects of the work.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: wbialek@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1514188112/-/DCSupplemental.

of thermodynamics can be derived from the distribution of these energies. Specifically, what matters is how many states have $E(\sigma)$ close to a particular value E. We can count this number of states, $n(E)$, or more simply the number of states with energy less than E, $\mathcal{N}(E)$. Then we can define a microcanonical entropy $S(E) = \ln \mathcal{N}(E)$. If we imagine a family of systems in which the number of degrees of freedom N varies, then a thermodynamic limit will exist provided that both the entropy and the energy are proportional to N at large N. The existence of this limit is by no means guaranteed.

In most systems, including the networks that we study here, there are few states with high probability, and many more states with low probability. At large N, the competition between decreasing probability and increasing numerosity picks out a special value of $E = E^*$, which is the energy of the typical states that we actually see; E^* is the solution to

$$\frac{dS(E)}{dE} = 1. \qquad [2]$$

For most systems, the energy $E(\boldsymbol{\sigma})$ has only small fluctuations around E^*, $\langle(\delta E)^2\rangle/(E^*)^2 \approx 1/N$, and, in this sense, most of the states that we see have the same value of log probability per degree of freedom. However, hidden in the function $S(E)$ are all of the parameters describing the interactions among the N degrees of freedom in the system. At special values of these parameters, $[d^2S(E)/dE^2]_{E=E^*} \to 0$, and the variance of E diverges as N becomes large. This is a critical point, and it is mathematically equivalent to the divergence of the specific heat in an equilibrium system (10).

These observations focus our attention on the "density of states" $\mathcal{N}(E)$. Rather than asking how often we see specific combinations of spiking and silence in the network, we ask how many states there are with a particular probability.

Experimental Example

The vertebrate retina offers a unique system in which the activity of most of the neurons comprising a local circuit can be monitored simultaneously using multielectrode array recordings. As described more fully in ref. 11, we stimulated salamander retina with naturalistic grayscale movies of fish swimming in a tank (Fig. 1*A*), while recording from 100 to 200 retinal ganglion cells (RGCs); additional experiments used artificial stimulus ensembles, as described in *Supporting Information*. Sorting the raw data (12), we identified spikes from 160 neurons whose activity passed our quality checks and was stable for the whole ~ 2 h duration of the experiment; a segment of the data is shown in Fig. 1*B*. These experiments monitored a substantial fraction of the RGCs in the area of the retina from which we record, capturing the behavior of an almost complete local population responsible for encoding a small patch of the visual world. The experiment collected a total of $\sim 2 \times 10^6$ spikes, and time was discretized in bins of duration $\Delta\tau = 20$ ms; all of the results discussed below are substantially the same at $\Delta\tau = 10$ ms and $\Delta\tau = 40$ ms (Fig. S1). For each neuron i, $\sigma_i = 1$ in a bin denotes that the neuron emitted at least one spike, and $\sigma_i = 0$ denotes that it was silent.

Counting States

Conceptually, estimating the function $\mathcal{N}(E)$ and hence the entropy vs. energy is easy: We count how often each state occurs, thus estimating its probability, and then count how many states have (log) probabilities in a given range. In Fig. 1 *C* and *D*, we show the first steps in this process. We identify the unique patterns of activity—combinations of spiking and silence across all 160 neurons—that occur in the experiment, and then count how many times each of these patterns occurs.

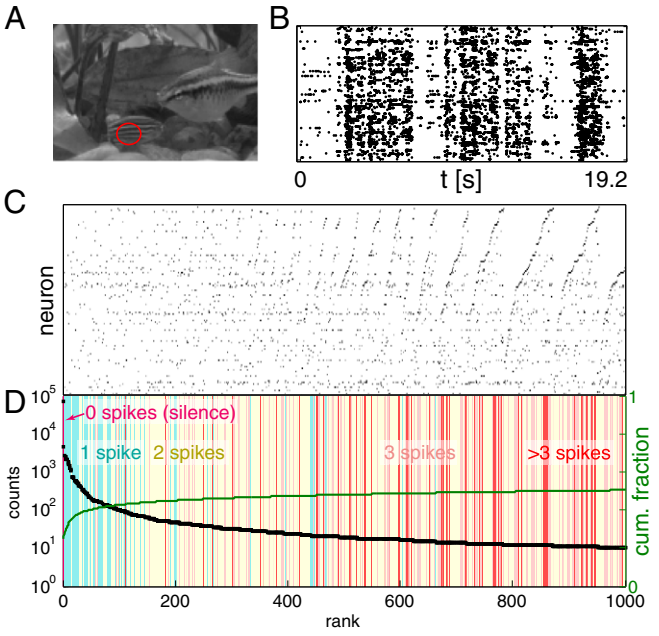

Fig. 1. Counting states in the response of RGCs. (*A*) A single frame from the naturalistic movie; red ellipse indicates the approximate extent of a receptive field center for a typical RGC. (*B*) Responses of $N = 160$ neurons to a 19.2-s naturalistic movie clip; dots indicate the times of action potentials from each neuron. In subsequent analyses, these events are discretized into binary (spike/silence) variables in time slices of $\Delta\tau = 20$ ms. (*C*) The 1,000 most common binary patterns of activity across $N = 160$ neurons, in order of their frequency. (*D*) Number of occurrences of each pattern (black, left axis), and the cumulative weight of the patterns in the empirical probability distribution (green, right axis), with labels for the total number of spikes in each pattern.

Even without trying to compute $S(E)$, the results of Fig. 1*D* are surprising. With N neurons that can either spike or remain silent, there are 2^N possible states. Not all these states can be visited equally often, because spikes are less common than silences, but even taking account of this bias, and trying to capture the correlations among neurons, our best estimate of the entropy for the patterns of activity we observe is $s \approx 0.15$ bits/neuron (see below). With $N = 160$ cells, this means that the patterns of activity are spread over $2^{Ns} \approx 1.67 \times 10^7$ possibilities, 100 times larger than the number of samples that we collected during our experiment. Indeed, most of the states that we saw in the full population occurred only once. However, roughly one thousand states occurred with sufficient frequency that we can make a reasonable estimate of their probability just by counting across ~2 h. Thus, the probability distribution $P(\sigma)$ is extremely inhomogeneous.

To probe more deeply into the tail of low-probability events, we can construct models of the distribution of states, and we have done this using the maximum entropy method (11, 13): We take from experiment certain average behaviors of the network, and then search for models that match these data but otherwise have as little structure as possible. This works if matching a relatively small number of features produces a model that predicts many other aspects of the data.

The maximum entropy approach to networks of neurons has been explored, in several different systems, for nearly a decade (14–23), and there have been parallel efforts to use this approach in other biological contexts (24–35). Recently, we have used the maximum entropy method to build models for the activity of up to $N = 120$ neurons in the experiments described above (11); see Fig. 2. We take from experiment the mean probability of each neuron generating a spike ($\langle\sigma_i\rangle$), the correlations between spiking in pairs of neurons

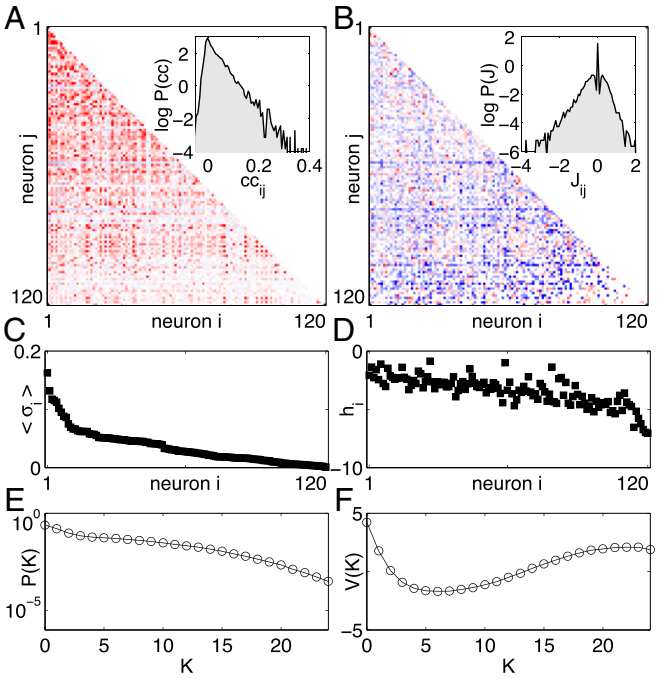

Fig. 2. Maximum entropy models for retinal activity in response to natural movies (11). (*A*) The correlation coefficients between pairs of neurons (red, positive; blue, negative) for a 120-neuron subnetwork. *Inset* shows the distribution of the correlation coefficients over the population. (*B*) The pairwise coupling matrix of the inferred model, J_{ij} from Eq. **4**. *Inset* shows the distribution of these pairwise couplings across all pairs ij. (*C*) The average probability of spiking per time bin for all neurons (sorted). (*D*) The corresponding bias terms h_i in Eq. **4**. (*E*) The probability $P(K)$ that K out of the N neurons spike in the same time bin. (*F*) The corresponding global potential $V(K)$ in Eq. **4**. Notice that *A*, *C*, and *E* describe the statistical properties observed for these neurons, whereas *B*, *D*, and *F* describe parameters of the maximum entropy model that reproduces these data within experimental errors.

($\langle\sigma_i\sigma_j\rangle$), and the probability that K out of the N neurons spike in the same small window of time [$P(K)$]. Mathematically, the maximum entropy models consistent with these data have the form

$$P(\{\sigma_i\}) = \frac{1}{Z}\exp[-E(\{\sigma_i\})], \quad [3]$$

$$E(\{\sigma_i\}) = -\sum_{i=1}^{N} h_i\sigma_i - \frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j - V(K), \quad [4]$$

where $K = \sum_{i=1}^{N}\sigma_i$ counts the number of neurons that spike simultaneously, and Z is set to ensure normalization. All of the parameters $\{h_i, J_{ij}, V(K)\}$ are determined by the measured averages $\{\langle\sigma_i\rangle, \langle\sigma_i\sigma_j\rangle, P(K)\}$.

This model accurately predicts the correlations among triplets of neurons (figure 7 in ref. 11), and how the probability of spiking in individual neurons depends on activity in the rest of the population (figure 9 in ref. 11). One can even predict the time-dependent response of single cells from the behavior of the population, without reference to the visual stimulus (figure 15 in ref. 11). Most important for our present discussion, the distribution of the energy $E(\{\sigma_i\})$ across the observed patterns of activity agrees with the distribution predicted by the model, deep into the tail of patterns that occur only once in the 2-h-long experiment (figure 8 in ref. 11). This distribution is closely related to the plot of entropy vs. energy that we would like to construct, and so the agreement with experiment gives us confidence.

The direct counting of states (Fig. 1) and the maximum entropy models (Fig. 2) give us two complementary ways of estimating the function $\mathcal{N}(E)$ and hence the entropy vs. energy in the same data set. Results are in Fig. 3; see also *Supporting Information*.

As emphasized above, the plot of entropy vs. energy contains all of the thermodynamic behavior of a system, and this has a meaning for any probability distribution, even if we are not considering a system at thermal equilibrium. Thus, Fig. 3 is as close as we can get to constructing the thermodynamics of this network. With the direct counting of states, we see less and less of the plot at larger N, but the part we can see is approaching a limit as $N \to \infty$, and this is confirmed by the results from the maximum entropy models. This, by itself, is a significant result. If we write down a model like Eq. **4**, then, in a purely theoretical discussion, we can scale the couplings between neurons J_{ij} with N to guarantee the existence of a thermodynamic limit (5), but with J_{ij} constructed from real data, we can't impose this scaling ourselves—either it emerges from the data or it doesn't. We can make the emergence of the thermodynamic limit more precise by noting that, at a fixed value of S/N, the value of E/N extrapolates to a well-defined limit in a plot vs. $1/N$, as in Fig. 3*A*, *Inset*. The results of this extrapolation are strikingly simple: The entropy is equal to the energy, within (small) error bars.

Interpreting the Entropy vs. Energy Plot

If the plot of entropy vs. energy is a straight line with unit slope, then Eq. **2** is solved not by one single value of E but by a whole range. Not only do we have $d^2S/dE^2 = 0$, as at an ordinary critical point, but all higher-order derivatives also are zero. Thus, the results in Fig. 3 suggest that the joint distribution of activity across neurons in this network is poised at a very unusual critical point.

We expect that states of lower probability (e.g., those in which more cells spike) are more numerous (because there are more ways to arrange K spikes among N cells as K increases from very low values). However, the usual result is that this trade-off—which is precisely the trade-off between energy and entropy in thermodynamics—selects typical states that all have roughly the same probability. The statement that $S(E) = E$, as suggested in Fig. 3, is the statement that states which are ten times less probable are exactly 10 times more numerous, and so there is no typical value of the probability.

The vanishing of d^2S/dE^2 corresponds, in an equilibrium system, to the divergence of the specific heat. Although the neurons obviously are not an equilibrium system, the model in Eqs. **3** and **4** is mathematically identical to the Boltzmann distribution. Thus, we can take this model seriously as a statistical mechanics problem, and compute the specific heat in the usual way. Further, we can change the effective temperature by considering a one-parameter family of models,

$$P(\{\sigma_i\}; T) = \frac{1}{Z(T)}\exp\left[-\frac{1}{T}E(\{\sigma_i\})\right], \quad [5]$$

with $E(\{\sigma_i\})$ as before (Eq. **4**). Changing T is just a way of probing one direction in the parameter space of possible models, and is not a physical temperature; the goal is to see whether there is anything special about the model (at $T = 1$) that describes the real system.

Results for the heat capacity of our model vs. T are shown in Fig. 4. There is a dramatic peak, and, as we look at larger groups of neurons, the peak grows and moves closer to $T = 1$, which is the model of the actual network. Importantly, the heat capacity grows even when we normalize by N, so that the specific heat, or heat capacity per neuron, is growing with N, as expected at a critical point, and these signatures are clearer in models that provide a more accurate description of the population activity; for details, see *Supporting Information*.

The temperature is only one axis in parameter space, and, along this direction, there are variations in both the correlations among neurons and their mean spike rates. As an alternative, we consider

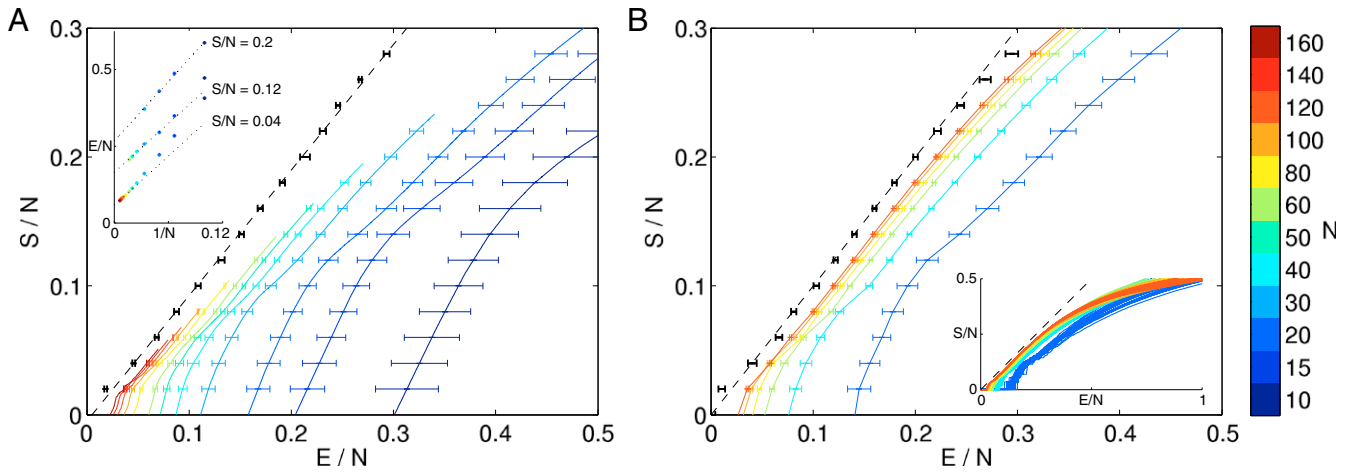

Fig. 3. Entropy vs. energy. (*A*) Computed directly from the data. Different colors show results for different numbers of neurons; in each case, we choose 1,000 groups of size N at random, and points are means with SDs over groups. *Inset* shows extrapolations of the energy per neuron at fixed entropy per neuron, summarized as black points with error bars in the main plot. Dashed line is the best linear fit to the extrapolated points, $S/N = (0.974 \pm 0.021)(E/N) + (-0.005 \pm 0.003)$. (*B*) Computed from the maximum entropy models. We choose $N = 20,\ 40, \cdots,\ 120$ neurons out of the population from which we record, and, for each of these subnetworks, we construct maximum entropy models as in Eq. **4**; for details of the entropy calculation, see *Supporting Information*. *Inset* shows results for many subnetworks, and the main plot shows means and SDs across the different groups of N neurons. Black points with error bars are an extrapolation to infinite N, as in *A*, and the dashed line is $S = E$.

a family of models in which the strength of correlations changes but spike rates are fixed. To do this, we replace the energy function in Eq. **4** with

$$E_\alpha(\{\sigma_i\}) = -\sum_{i=1}^{N} h_i'(\alpha)\sigma_i - \alpha\left[\frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j + V(K)\right], \qquad [6]$$

where α controls the strength of correlations, and we adjust all of the $h_i'(\alpha)$ to hold mean spike rates to their observed values.

At $\alpha = 0$, the model describes a population of independent neurons, so that the correlation coefficients are all $C = 0$ (Fig. 5*A*, *Left*). For $\alpha > 1$, the distribution of correlations broadens such that at $\alpha = 2$, some pairs are very strongly correlated (Fig. 5*A*, *Right*). This is reflected in a distribution of states that cluster around a small number of prototypical patterns, much as in the Hopfield network (4, 5). The entropy vs. energy plot, shown in Fig. 5*B*, singles out the ensemble at $\alpha = 1$ (Fig. 5*A*, *Middle*): Going toward independence (smaller α) gives rise to a concave bump at low energies, whereas $\alpha > 1$ ensembles deviate away from the equality line more at high energies. Correspondingly, we see in Fig. 5*D* that there is a peak in the specific heat of the model ensemble near $\alpha = 1$. As we look at larger and larger networks, this peak rises and moves toward $\alpha = 1$, which describes the real system.

The evidence for criticality that we find here is consistent with extrapolations from the analysis of smaller populations (15, 18). Those predictions were based on the assumption that spike probabilities and pairwise correlations in the smaller populations are drawn from the same distribution as in the full system, and that these distributions are sufficient to determine the thermodynamic behavior (36). Signs of criticality also are observable in simpler models, which match only the distribution of summed activity in the network (22), but less accurate models have weaker signatures of criticality (Fig. S2).

Couldn't It Just Be...?

In equilibrium thermodynamics, poising a system at a critical point involves careful adjustment of temperature and other parameters. Finding that the retina seems to be poised near criticality should thus be treated with some skepticism. Here we consider some ways in which we could be misled into thinking that the system is critical when it is not (see also *Supporting Information*).

Part of our analysis is based on the use of maximum entropy models, and one could worry that the inference of these models is unreliable for finite data sets (37, 38). Expanding on the discussion of this problem in ref. 11, we find a clear peak in the specific heat when we learn models for $N = 100$ neurons from even one-tenth of our data, and the variance across fractions of the data are only a few percent (Fig. S3).

Although the inference of maximum entropy models is accurate, less interesting models might mimic the signatures of criticality. In particular, it has been suggested that independent neurons with a broad distribution of spike rates could generate a distribution of N neuron activity patterns $\{\sigma_i\}$ that mimics some aspects of critical

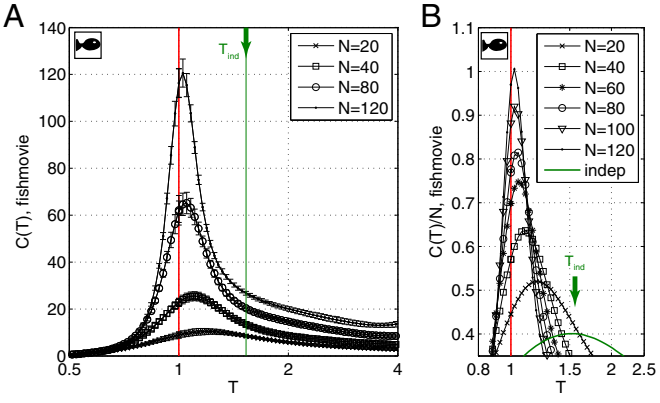

Fig. 4. Heat capacity in maximum entropy models of neurons responding to naturalistic stimuli. (*A*) Heat capacity, $C(T)$, computed for subnetworks of $N = 20$, 40, 80, and 120 neurons. Points are averages, and error bars are SDs, across 30 choices of subnetwork at each N. For details of the computations, see *Supporting Information*. (*B*) Zoom–in of the peak of $C(T)$, shown as an intensive quantity, $C(T)/N$. The peak moves toward $T = 1$ with increasing N, and the growth of $C(T)$ near $T = 1$ is faster than linear with N. Green arrow points to results for a model of independent neurons whose spike probabilities (firing rates) match the data exactly; the heat capacity is exactly extensive.

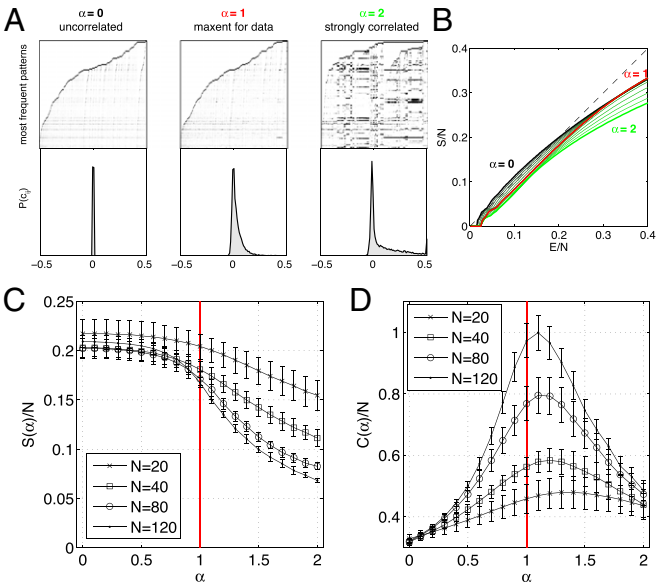

Fig. 5. Changing correlations at fixed spike rates. (*A*) Three maximum entropy (maxent) models for a 120-neuron network, where correlations have been eliminated (*Left*, $\alpha = 0$), left at the strength found in data (*Middle*, $\alpha = 1$), or scaled up (*Right*, $\alpha = 2$). (*Top*) The 10,000 most frequent patterns (black, spike; white, silence) in each model. (*Bottom*) The distribution of pairwise correlation coefficients. (*B*) Entropy vs. energy for the networks in *A*. (*C*) Entropy per neuron as a function of α, for different subnetwork sizes N. (*D*) Heat capacity per neuron exhibits a peak close to $\alpha = 1$. Error bars are SDs over 10 subnetworks for each N and α.

behavior (39). However, in an independent model built from the actual spike rates of the neurons, the probability of seeing the same state twice would be less than one part in a billion, dramatically inconsistent with the measured $P_c \approx 0.04$. Such independent models also cannot account for the faster than linear growth of the heat capacity with N (Fig. 4), which is an essential feature of the data and its support for criticality.

In maximum entropy models, the probability distribution over patterns of neural activity is described in terms of interactions between neurons, such as the terms J_{ij} in Eq. **4**; an alternative view is that the correlations result from the response of the neurons to fluctuating external signals. Testing this idea has a difficulty that has nothing to do with neurons: In equilibrium statistical mechanics, models in which spins (or other degrees of freedom) interact with one another are mathematically equivalent to a collection of spins responding independently to fluctuating fields (see *Supporting Information* for details). Thus, correlations always are interpretable as independent responses to unmeasured fluctuations, and, for neurons, there are many possibilities, including sensory inputs. However, the behavior that we see cannot be simply inherited from correlations in the visual stimulus, because we find signatures of criticality in response to movies with very different correlation structures (Fig. S4). Further, the pattern of correlations among neurons is not simply explained in terms of overlaps among receptive fields (Fig. S5), and, at fixed moments in the stimulus movie, neurons with nonzero spike probabilities have correlations across stimulus repetitions that can be even stronger than across the experiment as a whole (Fig. S6).

When we rewrite a model of interacting spins as independent spins responding to fluctuating fields, criticality usually requires that the distribution of fluctuations be very special, e.g., with the variance tuned to a particular value. In this sense, saying that correlations result from fluctuating inputs doesn't explain our observations. Recently, it has been suggested that sufficiently broad distributions of fluctuations lead generically to critical phenomenology (40). As explained in *Supporting Information*, mean field models have the property that the variance of the effective fields becomes large at the critical point, but more general models do not, and the correlations we observe do not have the form expected from a mean field model. The fact that quantitative changes in the strength of correlations would drive the system away from criticality (Fig. 5*D*) suggests that the distribution of equivalent fluctuating fields must be tuned, rather than merely having sufficiently large fluctuations.

Discussion

The traditional formulation of the neural coding problem makes an analogy to a dictionary, asking for the meaning of each neural response in terms of events in the outside world (41). However, before we can build a dictionary, we need to know the lexicon, and, for large populations of neurons, this already is a difficult problem: With 160 neurons, the number of possible responses is larger than the number of words in the vocabulary of a well-educated English speaker, and is more comparable to the number of possible short phrases or sentences. In the same way that the distribution of letters in words embodies spelling rules (28), and the distribution of words in sentences encodes aspects of grammar (42) and semantic categories (43), we expect the distribution of activity across neurons to reveal structures of biological significance.

In the small patch of the retina that we consider, no two cells have truly identical input/output characteristics (44). Nonetheless, if we count how many combinations of spiking and silence have a given probability in groups of $N > 20$ cells, this relationship is reproducible from group to group, and simplifies at larger N. This relationship between probability and numerosity of states is mathematically identical to the relationship between energy and entropy in statistical physics, and the simplification with increasing N suggests that we are seeing signs of a thermodynamic limit.

If we can identify the thermodynamic limit, we can try to place the network in a phase diagram of possible networks. Critical surfaces that separate different phases often are associated with a balance between probability and numerosity: States that are a factor F times less probable also are a factor F times more numerous. At conventional critical points, this balance occurs only in a small neighborhood of the typical probability, but, in the network of RGCs, it extends across a wide range of probabilities (Fig. 3). In model networks with slightly stronger or weaker correlations among pairs of neurons, this balance breaks down (Fig. 5), and less accurate models have weaker signatures of criticality (Fig. S2).

The strength of correlations depends on the structure of visual inputs, on the connectivity of the neural circuit, and on the state of adaptation in the system. The fact that we see signatures of criticality in response to very different movies, but not in model networks with stronger or weaker correlations, suggests that adaptation is tuning the system toward criticality. A sudden change of visual input statistics should thus drive the network to a noncritical state, and, during the course of adaptation, the distribution of activity should relax back to the critical surface. This can be tested directly.

Is criticality functional? The extreme inhomogeneity of the probability distribution over states makes it possible to have an instantaneously readable code for events that have a large dynamic range of likelihoods or surprise, and this may be well-suited to the the natural environment; it is not, however, an efficient code in the usual sense. Systems near critical points are maximally responsive to certain external signals, and this sensitivity also may be functionally useful. Most of the systems that exhibit criticality in the thermodynamic sense also exhibit a wide range of time scales in their dynamics, so that criticality may provide a general strategy for neural systems to bridge the gap between the microscopic time scale of spikes and the macroscopic

time scales of behavior. Critical states are extremal in all these different senses, and more; it may be difficult to decide which is relevant for the organism.

Related signatures of criticality have been detected in ensembles of amino acid sequences for protein families (29), in flocks of birds (33) and swarms of insects (45), and in the network of genes controlling morphogenesis in the early fly embryo (46); there is also evidence that cell membranes have lipid compositions tuned to a true thermodynamic critical point (47). Different, dynamical notions of criticality have been explored in neural (48, 49) and genetic (50, 51) networks, and in the active mechanics of the inner ear (52–54); recent work connects dynamical and statistical criticality, with the retina as an example (55). These results hint at a general principle, but there is room for skepticism. A new generation of experiments should provide decisive tests of these ideas.

Materials and Methods

Experiments were performed on the larval tiger salamander, *Ambystoma tigrinum tigrinum*, in accordance with institutional animal care standards at Princeton University.

ACKNOWLEDGMENTS. We thank A. Cavagna, D. S. Fisher, I. Giardina, M. Ioffe, S. C. F. van Opheusden, E. Schneidman, D. J. Schwab, J. P. Sethna, and A. M. Walczak for helpful discussions and comments on the manuscript. Research was supported in part by National Science Foundation Grants PHY-1305525, PHY-1451171, and CCF-0939370, by National Institutes of Health Grant R01 EY14196, and by Austrian Science Foundation Grant FWF P25651. Additional support was provided by the Fannie and John Hertz Foundation, by the Swartz Foundation, by the W. M. Keck Foundation, and by the Simons Foundation.

1. Anderson PW (1972) More is different. *Science* 177(4047):393–396.
2. Sethna JP (2006) *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford Univ Press, Oxford).
3. Guggenheim EA (1945) The principle of corresponding states. *J Chem Phys* 13(7): 253–261.
4. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79(8):2554–2558.
5. Amit DJ (1989) *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge Univ Press, Cambridge, UK).
6. Hertz J, Krogh A, Palmer RG (1991) *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA).
7. Ruelle D (1978) *Thermodynamic Formalism: The Mathematical Structures of Classical Equilibrium Statistical Mechanics* (Addison-Wesley, Reading, MA).
8. Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: The characterization of strange sets. *Phys Rev A* 33(2): 1141–1151.
9. Feigenbaum MJ, Jensen MH, Procaccia I (1986) Time ordering and the thermodynamics of strange sets: Theory and experimental tests. *Phys Rev Lett* 57(13):1503–1506.
10. Schnabel S, Seaton DT, Landau DP, Bachmann M (2011) Microcanonical entropy inflection points: Key to systematic understanding of transitions in finite systems. *Phys Rev E Stat Nonlin Soft Matter Phys* 84(1 Pt 1):011127.
11. Tkačik G, et al. (2014) Searching for collective behavior in a large network of sensory neurons. *PLoS Comput Biol* 10(1):e1003408.
12. Marre O, et al. (2012) Mapping a complete neural population in the retina. *J Neurosci* 32(43):14859–14873.
13. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106(4): 620–630.
14. Schneidman E, Berry MJ, 2nd, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087): 1007–1012.
15. Tkačik G, Schneidman E, Berry MJ, II, Bialek W (2006) Ising models for networks of real neurons. arXiv:q-bio/0611072.
16. Shlens J, et al. (2006) The structure of multi-neuron firing patterns in primate retina. *J Neurosci* 26(32):8254–8266.
17. Tang A, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J Neurosci* 28(2):505–518.
18. Tkačik G, Schneidman E, Berry MJ, II, Bialek W (2009) Spin glass models for a network of real neurons. arXiv:0912.5409.
19. Shlens J, et al. (2009) The structure of large-scale synchronized firing in primate retina. *J Neurosci* 29(15):5022–5031.
20. Ohiorhenuan IE, et al. (2010) Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* 466(7306):617–621.
21. Ganmor E, Segev R, Schneidman E (2011) Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc Natl Acad Sci USA* 108(23):9679–9684.
22. Tkačik G, et al. (2013) The simplest maximum entropy model for collective behavior in a neural network. *J Stat Mech* 2013:P03011.
23. Granot-Atedgi E, Tkačik G, Segev R, Schneidman E (2013) Stimulus-dependent maximum entropy models of neural population codes. *PLOS Comput Biol* 9(3):e1002922.
24. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc Natl Acad Sci USA* 103(50):19033–19038.
25. Tkačik G (2007) Information flow in biological networks. dissertation (Princeton University, Princeton, NJ).
26. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions. arXiv:0712.4397.
27. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA* 106(1):67–72.
28. Stephens GJ, Bialek W (2010) Statistical mechanics of letters in words. *Phys Rev E Stat Nonlin Soft Matter Phys* 81(6 Pt 2):066119.
29. Mora T, Walczak AM, Bialek W, Callan CG, Jr (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107(12):5405–5410.
30. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6(12):e28766.
31. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109(26):10340–10345.
32. Bialek W, et al. (2012) Statistical mechanics for natural flocks of birds. *Proc Natl Acad Sci USA* 109(13):4786–4791.
33. Bialek W, et al. (2014) Social interactions dominate speed control in poising natural flocks near criticality. *Proc Natl Acad Sci USA* 111(20):7212–7217.
34. Santolini M, Mora T, Hakim V (2014) A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites. *PLoS One* 9(6): e99015.
35. Mora T, Bialek W (2011) Are biological systems poised at criticality? *J Stat Phys* 144(2): 268–302.
36. Castellana M, Bialek W (2014) Inverse spin glass and related maximum entropy problems. *Phys Rev Lett* 113(11):117204.
37. Mastromatteo I, Marsili M (2011) On the criticality of inferred models. *J Stat Mech* 2011:P10012.
38. Marsili M, Mastromatteo I, Roudi Y (2013) On sampling and modeling complex systems. *J Stat Mech* 2013:P09003.
39. van Opheusden SCF (2013) Critical states in retinal population codes, Masters thesis (Universiteit Leiden, Leiden, The Netherlands).
40. Schwab DJ, Nemenman I, Mehta P (2014) Zipf's law and criticality in multivariate data without fine-tuning. *Phys Rev Lett* 113(6):068102.
41. Rieke F, Warland D, de Ruyter van Steveninck R, Bialek Spikes W (1997) *Exploring the Neural Code* (MIT Press, Cambridge, MA).
42. Pereira F (2000) Formal grammar and information theory: Together again? *Philos Trans R Soc Lond A* 358:1239–1253.
43. Pereira FC, Tishby N, Lee L (1993) Distributional clustering of English words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, ed Schubert LK (Assoc Comput Linguist, Stroudsburg, PA), pp 183–190.
44. Schneidman E, Bialek W, Berry MJ, II (2003) An information theoretic approach to the functional classification of neurons. *Advances in Neural Information Processing 15*, eds Becker S, Thrun S, Obermayer K (MIT Press, Cambridge, MA), pp 197–204.
45. Attanasi A, et al. (2014) Finite-size scaling as a way to probe near-criticality in natural swarms. *Phys Rev Lett* 113(23):238102.
46. Krotov D, Dubuis JO, Gregor T, Bialek W (2014) Morphogenesis at criticality. *Proc Natl Acad Sci USA* 111(10):3683–3688.
47. Honerkamp-Smith AR, Veatch SL, Keller SL (2009) An introduction to critical points for biophysicists; observations of compositional heterogeneity in lipid membranes. *Biochim Biophys Acta* 1788(1):53–63.
48. Beggs JM, Plenz D (2003) Neuronal avalanches in neocortical circuits. *J Neurosci* 23(35):11167–11177.
49. Friedman N, et al. (2012) Universal critical dynamics in high resolution neuronal avalanche data. *Phys Rev Lett* 108(20):208102.
50. Nykter M, et al. (2008) Gene expression dynamics in the macrophage exhibit criticality. *Proc Natl Acad Sci USA* 105(6):1897–1900.
51. Balleza E, et al. (2008) Critical dynamics in genetic regulatory networks: Examples from four kingdoms. *PLoS One* 3(6):e2456.
52. Eguíluz VM, Ospeck M, Choe Y, Hudspeth AJ, Magnasco MO (2000) Essential nonlinearities in hearing. *Phys Rev Lett* 84(22):5232–5235.
53. Camalet S, Duke T, Jülicher F, Prost J (2000) Auditory sensitivity provided by self-tuned critical oscillations of hair cells. *Proc Natl Acad Sci USA* 97(7):3183–3188.
54. Ospeck M, Eguíluz VM, Magnasco MO (2001) Evidence of a Hopf bifurcation in frog hair cells. *Biophys J* 80(6):2597–2607.
55. Mora T, Deny S, Marre O (2015) Dynamical criticality in the collective activity of a population of retinal neurons. *Phys Rev Lett* 114(7):078105.

Supporting Information

Tkačik et al. 10.1073/pnas.1514188112

Thermodynamics and Probability Distributions

The fundamental variables of thermodynamics are energy, temperature, and entropy. For the states taken on by a network of neurons, energy and temperature are meaningless, so it is difficult to see how we can construct a thermodynamics for these systems. However, in statistical mechanics, all thermodynamic quantities are derivable from the Boltzmann distribution, the probability that the system will be found in any particular state. Thus, all thermodynamic statements are equivalent to statements about this underlying probability distribution, and, in this sense, we should be able to construct thermodynamics for a much broader range of probability distributions that describe a large number of variables.

The idea that probability distributions over N variables can have an associated thermodynamics in the $N \to \infty$ limit is powerful but perhaps not so widely used. This connection is well-studied by mathematical physicists (1) and has been a guide to the analysis of experiments on dynamical systems (2, 3). We have used these ideas to construct a thermodynamics of natural images (4) and have emphasized the connection of thermodynamic criticality to Zipf's law (5). Here we give a somewhat pedagogical discussion, in the hope of making the results accessible to a broader audience.

The Boltzmann Distribution. We start by recalling that, for a system in thermal equilibrium at temperature T, the probability of finding the system in state s is given by

$$P_s = \frac{1}{Z} e^{-E_s/k_B T}, \qquad \text{[S1]}$$

where E_s is the energy of the state, and Boltzmann's constant k_B converts between conventional units of temperature and energy. The partition function Z serves to normalize the distribution, which requires

$$Z = \sum_s e^{-E_s/k_B T}, \qquad \text{[S2]}$$

but, in fact, this normalization constant encodes many physical quantities. The logarithm of the partition function is proportional to the free energy of the system, the derivative of the free energy with respect to the volume occupied by the system is the pressure, the derivative with respect to the strength of an applied magnetic field is the magnetization, and so on.

The state of a system is defined by the joint configuration of all its parts. Thus, in a classical gas or liquid, s is defined by the positions and velocities of all of the constituent atoms. Different gases or liquids differ not because these variables are different but because the energy E_s is a different function of these N underlying variables. However, thermodynamics doesn't make reference to all these details. Which aspects of the underlying microscopic rules actually matter for predicting the free energy and its derivatives?

We can write the sum over all states as a sum first over states that have the same energy and then as a sum over energies. We do this by introducing an integral over a delta function into the sum,

$$Z = \sum_s e^{-E_s/k_B T} = \sum_s \left[\int dE\, \delta(E - E_s) \right] e^{-E_s/k_B T} \qquad \text{[S3]}$$

$$= \int dE \sum_s \delta(E - E_s) e^{-E_s/k_B T} \qquad \text{[S4]}$$

$$= \int dE\, e^{-E/k_B T} \left[\sum_s \delta(E - E_s) \right]. \qquad \text{[S5]}$$

We see that the way in which the energy depends on each state appears only in the brackets, a function $n(E)$ that counts how many states have a particular energy,

$$n(E) = \sum_s \delta(E - E_s). \qquad \text{[S6]}$$

Looking ahead to the analysis of real data, it will be convenient to rearrange Eq. **S5** slightly. Instead of counting the number of states that have energy E, we can count the number of states with energy less than E,

$$\mathcal{N}(E) = \sum_s \Theta(E - E_s), \qquad \text{[S7]}$$

where the step function is defined by

$$\Theta(x > 0) = 1 \qquad \text{[S8]}$$

$$\Theta(x < 0) = 0. \qquad \text{[S9]}$$

However, the step function is the integral of the delta function, which means that we can integrate by parts in Eq. **S5** to give

$$Z = \frac{1}{k_B T} \int dE\, e^{-E/k_B T} \mathcal{N}(E). \qquad \text{[S10]}$$

If we think about N variables, each of which can take on only two states, the total number of states is 2^N. More generally, we expect that the number of possible states in a system with N variables is exponentially large, so it is natural to think not about the number of states $\mathcal{N}(E)$ but about its logarithm,

$$S(E) = \ln \mathcal{N}(E), \qquad \text{[S11]}$$

which is called the entropy.

As a technical aside, we can either define the entropy in terms of the density of states with energy close to E, what we have called $n(E)$, or use the number of states with energy less than E, what we have called $\mathcal{N}(E)$. When the number of degrees of freedom is small, these are both badly behaved functions—$n(E)$ is singular, and $\mathcal{N}(E)$ has visible steps. However, as N becomes large, both functions become smooth, and we can do all of the usual operations of differentiation or integration by parts without worries. Importantly, when it comes time to analyze experimental data, using $\mathcal{N}(E)$ allows us to avoid making bins along the E axis.

Substituting from Eq. **S11** into Eq. **S10**, the partition function can be written as an integral determined only by the function $S(E)$, entropy vs. energy,

$$Z = \frac{1}{k_B T} \int dE \, \exp\left[-\frac{E}{k_B T} + S(E) \right]. \quad \textbf{[S12]}$$

One of the key ideas in thermodynamics is that certain variables are "extensive," that is, proportional to the number of particles or variables in the system, whereas other variables are "intensive," independent of the system size. Temperature is an intensive variable, whereas energy and entropy are extensive variables. It is then natural to think about the energy per particle, $\epsilon = E/N$, and the entropy per particle, $S(E)/N = s(\epsilon)$. In the limit of large N, we expect $s(\epsilon)$ to become a smooth function. Substituting into Eq. **S12**, the partition function can be written as

$$Z = \frac{N}{k_B T} \int d\epsilon \, e^{-Nf(\epsilon)/k_B T} \quad \textbf{[S13]}$$

$$f(\epsilon) = \epsilon - k_B T s(\epsilon). \quad \textbf{[S14]}$$

We note that $f(\epsilon)$ is the difference between energy and entropy, scaled by the temperature, and is called the free energy.

Whenever we have an integral of the form in Eq. **S13**, at large N, we expect that it will be dominated by values of ϵ close to the minimum of $f(\epsilon)$. This minimum ϵ_* is the solution to the equation

$$\frac{df(\epsilon)}{d\epsilon} = 0 \Rightarrow \frac{1}{k_B T} = \frac{ds(\epsilon)}{d\epsilon}, \quad \textbf{[S15]}$$

which we can also think of as defining the temperature. Notice that T being positive requires that the system have $ds(\epsilon)/d\epsilon > 0$, which means there are more states with higher energies.

If we expand $f(\epsilon)$ in the neighborhood of ϵ_*, we have

$$f(\epsilon) = f(\epsilon_*) - \frac{k_B T}{2} \frac{d^2 s(\epsilon)}{d\epsilon^2}\bigg|_{\epsilon_*} (\epsilon - \epsilon_*)^2 + \cdots, \quad \textbf{[S16]}$$

which gives

$$Z \approx \frac{N}{k_B T} e^{-Nf(\epsilon_*)/k_B T} \int d\epsilon \, e^{-A(\epsilon - \epsilon_*)^2} \quad \textbf{[S17]}$$

$$A = \frac{N}{2} \left[-\frac{d^2 s(\epsilon)}{d\epsilon^2}\bigg|_{\epsilon_*} \right]. \quad \textbf{[S18]}$$

This looks as if the energy per particle is drawn from an approximately Gaussian distribution, with mean ϵ_* and variance

$$\left\langle (\delta\epsilon)^2 \right\rangle = \frac{1}{N} \left[-\frac{d^2 s(\epsilon)}{d\epsilon^2}\bigg|_{\epsilon_*} \right]^{-1}, \quad \textbf{[S19]}$$

and, indeed, this can be shown more directly from the Boltzmann distribution.

With the interpretation of ϵ_* as the mean energy per particle, we can use Eq. **S15** to calculate how this energy changes when we change the temperature, and we find

$$\frac{d\epsilon_*}{dT} = \frac{1}{k_B T^2} \left[-\frac{d^2 s(\epsilon)}{d\epsilon^2}\bigg|_{\epsilon_*} \right]^{-1}. \quad \textbf{[S20]}$$

The change in energy with temperature is called the heat capacity C, and, when we normalize per particle, it is referred to as the specific heat. Combining Eqs. **S19** and **S20**, we see that the specific heat C/N is connected to the variance in energies,

$$\left\langle (\delta\epsilon)^2 \right\rangle = k_B T^2 \frac{C}{N}. \quad \textbf{[21]}$$

This relationship also can be proven without resorting to the approximation in Eq. **S16**.

Our discussion thus far assumes that the second derivative of the entropy with respect to the energy is not zero. If we take all our results at face value, then, when $d^2 s/d\epsilon^2 \to 0$, the specific heat will become infinite (Eq. **S20**), as will the variance of the energy per particle (Eq. **S19**). This is a critical point.

There is much more to be said about the analysis of critical points using the entropy vs. energy. However, our concern here is how these ideas connect to systems that are not in thermal equilibrium, so that temperature and energy are not relevant concepts. What we would like to show is that many of the thermodynamic quantities nonetheless serve to characterize the behavior of any probability distribution for a very large number of variables.

Distributions, More Generally. Rather than trying to compute the partition function, we can ask, for any distribution, how the normalization condition is satisfied. We still imagine that there are states s, built of N different variables, as with the patterns of spiking and silence in a network of neurons. Each state s has a probability P_s, and we must have

$$1 = \sum_s P_s. \quad \textbf{[S22]}$$

We can now follow the same strategy that we used above for the partition function: We do the sum first by summing over all of the states that have the same value of the (log) probability, and then we sum over this value. We start by defining

$$E_s = -\ln P_s, \quad \textbf{[S23]}$$

as in Eq. **1**. Then we have

$$\sum_s P_s = \sum_s \int dE \, \delta(E - E_s) P_s. \quad \textbf{[S24]}$$

However, because $P_s = e^{-E_s}$, we can rewrite this as

$$\sum_s P_s = \int dE \, e^{-E} \sum_s \delta(E - E_s). \quad \textbf{[S25]}$$

Integrating by parts, we obtain

$$\sum_s P_s = \int dE \, e^{-E} \mathcal{N}(E), \quad \textbf{[S26]}$$

where $\mathcal{N}(E)$ is a cumulative density of states, as in Eq. **S7**,

$$\mathcal{N}(E) = \sum_s \Theta(E - E_s). \quad \textbf{[S27]}$$

Again, this is a number of states, so the logarithm of this number is an entropy, exactly as in Eq. **S11**. Thus, the statement that the probability distribution is normalized becomes

$$\sum_s P_s = \int dE \exp[-E + S(E)]. \quad \textbf{[S28]}$$

If we have a system in which the state s is built out of N variables, then we expect that, for large N, both log probabilities (E) and entropies (S) are proportional to N. A standard example is in information theory, where s could label a message built out of N symbols, and the proportionality $E \propto N$ is central to proofs of the classic coding theorems (6). In the case of interest to us here, we can look at the states taken on by groups of N neurons, and we can vary N over some range. The function $\mathcal{N}(E)$, and hence the entropy $S(E)$, is a property of a single system with a particular value of N, and, to remind us of this fact, we can write $S_N(E)$. What happens as N become large is an experimental question. However, in many of the examples that we understand—from statistical physics, from information theory, and indeed from more general examples in probability theory—there is a well-defined limiting behavior at large N, which means that there is a function

$$s(\epsilon) = \lim_{N\to\infty} \frac{1}{N} S_N(E = N\epsilon). \quad \textbf{[S29]}$$

If this limit exists, then the normalization condition on the probability distribution in Eq. **S28** becomes

$$\sum_s P_s \to NP_0 \int d\epsilon\, e^{-Nf(\epsilon)}, \quad \textbf{[S30]}$$

$$f(\epsilon) = \epsilon - s(\epsilon). \quad \textbf{[S31]}$$

Now we can see the correspondence with the description of an equilibrium thermodynamic system, which leads up to the expressions for the partition function in Eqs. **S13** and **S14**:

i) We can assign an energy to every state of the system, which is just the negative log probability. The effective temperature of the system is $k_B T = 1$.
ii) We can count the number of states below a given energy, and the log of this number is an entropy.
iii) If there are N elements (e.g., neurons) in our system, it is natural to ask about the entropy per element as a function of the energy per element. If this function has a smooth limit as N becomes large, $s(\epsilon)$, then we can define a thermodynamics for the system.
iv) When we sum over states, the sum is dominated by states that minimize the free energy, $f(\epsilon) = \epsilon - s(\epsilon)$, just as in ordinary thermodynamics, provided that the curvature of the free energy at this minimum is nonzero.
v) The dominance of states near the minimum of the free energy enforces the notion of "typicality" (6), so that at large N most of the states we actually see have essentially the same value of log probability.
vi) If the curvature at the minimum of the free energy vanishes, then the usual ideas of typicality break down, and we will see large fluctuations in the log probability of states, even if we normalize this log probability by N.
vii) The large variance in log probability is mathematically equivalent to a diverging specific heat in the thermodynamic case. This is a signature of a critical point.

More About Criticality. Before leaving this discussion, we note that there are other signatures of criticality, and even different notions of criticality. In equilibrium systems with interactions that extend only over short distances, correlations typically extend over some longer but finite distance ξ; at the critical point, this correlation length diverges, so that there is no characteristic length scale—all scales between the size of the constituent particles and the size of the system as a whole are relevant (7). Not only does the specific heat diverge at the critical point, but so does the susceptibility to external fields. All of these diverging quantities have a power law dependence on the difference between the actual temperature and the critical temperature, and the exponents of these power laws are quantitatively universal: Many different systems, with different microscopic constituents, exhibit precisely the same exponents, and, in a certain precise sense, these exponents give a complete description of the system in the neighborhood of the critical point (8, 9). In the study of complex, nonequilibrium systems, scale invariance and power law behaviors often are taken as signs of criticality, but seldom is it possible to exhibit these behaviors over the wide range of scales that are the standard in studies of equilibrium critical phenomena, so one must be cautious.

In almost all equilibrium systems, the approach to criticality also is associated with the emergence of long time scales in the dynamics; as with the divergence of the correlation length ξ, the divergence of the correlation time in the dynamics means that there is a form of temporal scale invariance at criticality. Deterministic dynamical systems also exhibit critical phenomena, often called bifurcations, where the system's behavior changes qualitatively in response to an infinitesimal change in parameters (10). These phenomena are easiest to understand when the number of degrees of freedom N is small, but then the sharp bifurcations are rounded if there is noise in the system; the example of equilibrium statistical mechanics shows how noisy dynamical systems can recover sharp transitions in the limit of large N. In general, it is not clear how dynamical and statistical notions of criticality are related to one another in systems with many degrees of freedom.

Experimental Methods

Much of the analysis in this paper is based on the same data set as in ref. 11. For completeness, we review our experimental methods here. Retinae were isolated from the eye in darkness, and the retina was pressed against a custom-fabricated array of 252 electrodes. The retina was superfused with oxygenated Ringer's medium at room temperature. Electrode voltage signals were acquired and digitized at 10 kHz by a 252-channel preamplifier (MultiChannel Systems). The sorting of these signals into action potentials from individual neurons was done offline using the methods of ref. 12.

The repeated natural movie was a movie of a fish tank captured at 30 Hz with a standard camera; it lasted 20 s and was repeated 297 times. As noted in the main text, this experiment allowed us to resolve 160 neurons across the recording array. The random checkerboard consisted of square pixels, 69 μm on a side, each chosen independently black or white 30 times per second, creating a 30-s random movie that was repeated 69 times; this experiment yielded 120 stable, resolved cells. For the spatially uniform flicker, the luminance of the entire screen was chosen randomly from a Gaussian distribution 60 times per second, creating a 10-s-long random sequence that was repeated 98 times; we separated the signals from 111 neurons.

Effects of Bin Size

We follow earlier work and define the states of the neural network in discrete time bins (13). That is, we slice the time axis in bins of duration $\Delta\tau$, and define $\sigma_i = 1$ at time t if neuron i spikes in the window $[t, t+\Delta\tau)$, and $\sigma_i = 0$ otherwise. We choose $\Delta\tau = 20$ ms because this captures the structure of the correlation functions, but it should be admitted that there is some arbitrariness here. If we make bins too large, surely we are grouping together distinct responses of the network, whereas, if we make the bins too small, then meaningful correlations are spread over multiple bins, and we need to analyze the distribution of state sequences rather than instantaneous states (14).

One might hope, however, that there is a range of bin sizes over which the basic structure of the distribution $P(\{\sigma_i\})$ is constant. We test this in Fig. S1, which should be compared with Fig. 3. Fig. S1 shows the entropy vs. energy, computed directly from the data, with bin widths of $\Delta\tau = 10$ ms and 40 ms, whereas we use $\Delta\tau = 20$ ms in the main text. Although details vary a bit, in all

cases, we see the approach to $s(\epsilon)=\epsilon$ as N becomes large. Although much remains to be understood about the dynamics of the states in this network, Fig. S1 demonstrates that our main results do not depend sensitively on the choice of $\Delta\tau$.

Analysis of Maximum Entropy Models

We can take our maximum entropy model seriously as a statistical mechanics problem and use Monte Carlo simulation to generate samples of the states $\{\sigma_i\}$ drawn from our model distribution. Heat capacity curves were estimated by running a Metropolis Monte Carlo sampler independently at every T. Because the model assigns an energy $E=E(\{\sigma_i\})$ to each state, we can compute the mean and variance of E from a single long Monte Carlo run, and thus estimate the heat capacity through the thermodynamic identity in Eq. **S21**. Samples of the energy were collected at every sweep (roughly N spin flips); 2×10^6 sweeps were performed for every T.

To estimate the function $n(E)$ in the maximum entropy models, including the α-ensembles in Fig. 5, we used Wang–Landau sampling (15). In detail, the complete energy range was divided into 2×10^4 equidistant energy bins (6×10^3 for the α-ensembles), the histogram flatness criterion was 0.9, and the final multiplicative update was $1+10^{-5}$. These measurements, as well as the specific heat curves, can both be used to give an estimate of the entropy of the distribution, and these agree to within better than 1% (11), providing a check on our sampling procedures. For more on these matters, see the methods section, "Computing the entropy . . .," of ref. 11.

In addition to the models described in the main text, we have also considered models that do not include the term $V(K)$ in Eq. **4**; these are maximum entropy models that match exactly the mean spike probabilities of individual neurons, and the pairwise correlations, but not the probability of K neurons spiking simultaneously. As explained in ref. 11, these simpler, purely pairwise models provide noticeably less accurate descriptions of the network activity. [Note that the parameters $\{h_i;J_{ij}\}$ in the two models are not the same, but must be found, independently, to match the relevant expectation values (11).] Importantly, although both models capture all of the pairwise correlations among neurons, the peak of the specific heat is much stronger and more clearly N-dependent in the more accurate model, as shown in Fig. S2.

More About Alternatives

In this section, we expand on alternative interpretations of the data, arguing that the signatures of criticality are unlikely to be explained away as spurious consequences of less interesting models.

Impact of Limited Data. We have tested in detail the reliability with which maximum entropy models can be inferred from the available data. As explained in ref. 11, we can learn these models from 90% of the data and then compare the quality of the model against both the training set and the held-out test set. Even with $N=120$ neurons, the model predicts that the log likelihood of the test data is the same as that of the training data, within error bars, and these errors are less than 1% (figure 4 of ref. 11). Still, one might worry that small errors associated with the finiteness of the data set could have a disproportionate impact on the putative signatures of criticality. To test for this, we have learned models for $N=100$ neurons from fractions of the data ranging down to just 10%; results for the heat capacity vs. temperature (as in Fig. 4) are shown in Fig. S3. We see that the sharp peak in $C(T)$ is essentially independent of the sample size across this wide range, and that the variations in $C(T)$ across different small fractions of the data are only a few percent. Thus, this behavior is not a result of overfitting, nor is it linked in any way to the size of our data set.

It is important that, in Fig. S3, we are always looking at the same 100 neurons; otherwise, variability across subsets could be confused with sampling errors. When we change the size of the data, we are choosing, at random, some fraction of the experiment, and, for each fraction, we examine 10 such random choices. For each choice, we make a completely independent reconstruction of the maximum entropy model, which means that variability includes not just the effects of finite data but also any errors in parameter estimation or in the Monte Carlo estimate of the specific heat. Evidently, all of these errors are quite small.

Are Correlations Inherited from the Visual Stimulus? As discussed in the main text, one possible interpretation of our observations is that correlations among neurons simply reflect correlations in the visual stimulus. In this case, any interesting features in the joint distribution of activity among many neurons would be entirely traceable to the structure of the sensory inputs.

The idea that correlations among neurons should be decomposed into contributions from their inputs and contributions intrinsic to the circuit is very old (16), dating back to a time when it was hoped that measurement of correlations would allow a direct inference of connectivity in the circuit. Before discussing the origin of correlations, it is important to emphasize that the distinction between "stimulus-induced" and "intrinsic" correlations is not a distinction that the brain can make. Experimentally, we make this distinction by providing exact repetitions of the stimulus, but this never happens in the natural world. The only knowledge that the brain has of its visual inputs is the set of signals provided by the population of ganglion cells itself, so there is no way to search for correlations with some other reference signal. We also note that, following decades of experiments on correlations among RGCs (17, 18), there is now direct evidence that triggering spikes in one ganglion cell changes the response of other ganglion cells to sensory signals,* so that these cells certainly are not responding independently to their visual inputs.

Although the dissection of the correlations is irrelevant for brain function, it is interesting to ask, mechanistically, how these correlations arise. If they arise solely from the visual inputs, then changing the statistical structure of these inputs should produce a dramatic effect. We have replaced the natural movies with randomly flickering checkerboards (an approximation to spatiotemporal white noise) and spatially uniform but temporally random flicker. In each case, we have constructed maximum entropy models (Eqs. **3** and **4**) and searched for a peak in the specific heat vs. temperature, as in Fig. 4; results are shown in Fig. S4.

Although there are quantitative differences among the responses to the different stimulus ensembles, we see that there are signatures of criticality in each case. As with the natural movies, there is a peak in the specific heat, the height of the peak grows with the number of neurons, and the location of the peak moves toward $T=1$ at larger N. It thus seems unlikely that these signatures of criticality in the specific heat are merely a reflection of input statistics. Indeed, we should remember that the decomposition of correlations into intrinsic and stimulus-induced is incomplete, because the retina adapts to the distribution of its inputs, on many time scales. It would appear that some combination of anatomical connectivity and adaptation poises the population of RGCs near a peak in the specific heat. This points toward future experiments that should probe more directly the invariance of thermodynamic behavior across adaptation states.†

It seems worth emphasizing that, even if correlations are largely inherited from the visual stimulus, this transformation from input to output correlations is nontrivial. The conventional model for the input–output relations of the neurons is the "linear–nonlinear" model, in which the probability of spiking is determined by an instantaneous nonlinear function of a linearly filtered version of the stimulus. In the retina, with the stimulus given by the light

*Asari H, Meister M, Computational and Systems Neuroscience, February 23–26, 2012, Salta Lake City, UT.

†Ioffe M, Tkačik G, Bialek W, Berry MJ, II, Computational and Systems Neuroscience, February 27 to March 2, 2014, Salt Lake City, UT.

intensity as a function of space and time, $I(\vec{x},t)$, the probability of spiking for one cell in one time bin is then

$$p(t)=p_0\,g\left[\int d^2x\int d\tau F(\vec{\mathrm{x}},\tau)I(\vec{\mathrm{x}},t-\tau)\right], \qquad \textbf{[S32]}$$

where p_0 sets the maximum response, $g[\cdot]$ is a nonlinear function that we can normalize to range between 0 and 1, and $F(\vec{\mathrm{x}},\tau)$ is the linear spatiotemporal receptive field of the cell. It is a theorem that, if this model is an accurate description of the neural response, then the receptive field can be determined by correlating the spiking output with the spatiotemporal variations in the input, provided that the inputs are chosen from a Gaussian ensemble. In particular, if the inputs are white noise (down to the spatial and temporal resolution used), as in the random checkerboard experiments, then

$$F(\vec{\mathrm{x}},\tau)\propto\left\langle I(\vec{\mathrm{x}},t-\tau)\delta(t-t_{\mathrm{spike}})\right\rangle, \qquad \textbf{[S33]}$$

where t_{spike} is the time of a spike and the average $\langle\cdots\rangle$ is computed across a long sample of the random checkerboard movie. As described in the supporting information of ref. 19, we have constructed these receptive fields for every cell in the population, and then mapped the nonlinearities $g[\cdot]$ independently for each cell. If we then generate spikes at the output of this population, and compute their pairwise correlation coefficients, we obtain the results shown in Fig. S5.

There is a widely held intuition that correlations among neural responses in the retina should be understood as being shaped largely by the overlap of receptive fields, but Fig. S5 suggests that the situation is more complex. The linear–nonlinear model predicts correlations based on receptive field structure, but these predictions are strongly at variance with what we see in the data. The distribution of correlations in the model is narrower than in the data, failing to access the tail of strong positive correlations and cutting off at only modest negative correlations. Taking each pair of neurons individually, we see that the predicted and observed correlations are almost unrelated to one another.

Zipf's Law, Superposition, and Related Matters. Rather than counting states that have a particular value of the (log) probability, we can simply put the states in order of their probability, highest probability states first. The resulting plot of probability vs. rank is sometimes called the "Zipf plot," with reference to the corresponding analysis of words in written language (20). As we have emphasized elsewhere (4, 5, 21), the Zipf plot is essentially the plot of entropy vs. energy, turned on its side. Concretely, if the state with rank r has probability p_r, then we have r states with probability $p\geq p_r$, or an effective energy $E\leq -\ln p_r$. However, the number of states with energy less that E is what we have called the cumulative density of states, $\mathcal{N}(\mathcal{E})$. Thus, we have

$$\mathcal{N}(E)|_{E=-\ln p_r}=r, \qquad \textbf{[S34]}$$

or, for the entropy,

$$S(E)|_{E=-\ln p_r}=\ln r. \qquad \textbf{[S35]}$$

What Zipf observed about words is that $p_r\approx A/r$, up to some maximum r. If we take this as an exact statement ("Zipf's law"), then $r=A/p_r$, and hence Eq. **S35** becomes

$$S(E)=E+\ln A. \qquad \textbf{[S36]}$$

Thus, Zipf's law is equivalent to a linear relation between entropy and energy, with slope one. Because this means that the second derivative of the entropy with respect to the energy vanishes, Zipf's law seems to imply criticality, in precisely the sense that we are discussing for neurons.

Zipf's law is a power law, $p_r\propto r^{-\gamma}$, in this case with $\gamma=1$, although this is quite different from the usual power law scaling relations among thermodynamic variables near an equilibrium critical point (7–9). The ubiquity of Zipf's law has led many people to wonder if there is some universal underlying mechanism. In several ways, this discussion parallels the discussion of $1/f$ noise: In many systems, fluctuations over time have a spectrum without any obvious scale, and when we plot the spectrum vs. frequency, especially on logarithmic axes, the behavior approximates a power law with exponent close to one.

In the discussion of $1/f$ noise, it was realized, early on, that a system might appear to be scale-invariant if it has a discrete set of scales spread over a sufficiently broad range. Thus, if we look at fluctuations over time, and what we see is the sum or superposition of many processes with correlation times τ_1, τ_2, τ_3, …, then, if these correlation times come from a broad distribution, the net spectrum will be nearly featureless; even a handful of correlation times, with the right spread, can give a good approximation to $1/f$ noise. It seems that this is the correct description of $1/f$ noise in metals (22). Importantly, if the apparent $1/f$ noise really is a superposition of many noise sources with a range of correlation times, then, if we can perturb these time scales, we should see measurable departures from $1/f$ behavior, and this was the experimental strategy used in sorting out the behavior of current noise in metals. What this means, of course, is that this is an example of almost $1/f$ noise, and that the small deviations from truly scale-invariant behavior are crucial.

An extreme version of the mixture model for scale-invariant behavior is discussed by van Opheusden (23), who considered populations of neurons firing independently but with a distribution of mean spike probabilities. With a proper choice of this distribution, completely independent neurons can generate a good approximation to Zipf's law at fixed N. As noted in the main text, however, the actual distribution of spike probabilities that we see in the data does not have this special property. Further, unless the distribution of spike probabilities is singular, the variance of log probability across all of the states of the network will be exactly proportional to the number of neurons that we consider, and hence such models cannot explain the supralinear growth of the heat capacity in Fig. 4, which is one of the key signatures of criticality.

Aitchison et al. (24) have suggested that the original example of Zipf's law for words in English should be explained by adapting the multiple time scale idea in $1/f$ noise. The distribution of words can be thought of as a sum over contributions from several parts of speech (nouns, verbs, adjectives, etc.), and, for each part of speech, we do not see Zipf's law but rather a distribution that has a characteristic scale; the scales for different parts of speech are different, and, when we sum over all parts of speech, we see the emergence of Zipf's law. If this is correct, then, as in the case of $1/f$ noise in metals, we must conclude that Zipf's law is not exact. Further, it should be possible to modulate the characteristic scales, or the weights given to each component of the distribution, and thereby make the deviations from Zipf's law more apparent. In metals, one can do this simply by modulating the temperature (22). In language, the scales and weights for different parts of speech vary across languages, topics, and authors, so one might expect the equivalent of the temperature modulation experiment has been done, implicitly, many times, although this is not discussed in ref. 24. With modern corpora, searching more carefully for departures from Zipf's law should be straightforward. At best, however, explaining Zipf's law as a superposition over multiple parts of speech would be a demonstration that deviations from Zipf's law are important.

In connecting Zipf's law to criticality, one must keep in mind that critical phenomena exist only in the thermodynamic limit. As we have defined it, the entropy vs. energy $S(E)$ is not a smooth function in a system of finite size, because there are discrete states with particular probabilities and hence particular energies. A differentiable function $s(\epsilon)$ emerges, as in Eq. **S29**, only in the limit that we consider a system with many degrees of freedom. In the example of language, to make a connection to criticality thus requires more than counting words. Instead, we should imagine text segments with a length of N letters or words, and ask how the Zipf plot evolves as a function of N. The emergence of a function $s(\epsilon)$ would correspond to the plot of $(\ln p_r)/N$ vs. $(\ln r)/N$ converging to a limit as N grows, and evidence for criticality depends on the properties of this limiting function. Thus, criticality is much more than Zipf's law at fixed N.

More General Hidden Variable Models. The idea that correlations among neurons might be inherited from the visual stimulus is one possibility among many. More generally, we might ask if the pattern of correlations could be understood as the independent response of neurons to some signal that is effectively external to the network, or at least hidden from an observer who sees only the patterns of spikes and silence. To assess this possibility, it is useful to step back and think about simpler models in statistical mechanics. Almost everything that we will say in this section is well-known in the physics literature, but it seems useful to be explicit.

Consider the mean field Ising ferromagnet, in which spins $\sigma_i = \pm 1$ experience an effective magnetic field that is proportional to the average over all of the other spins in the system, so that

$$E(\{\sigma_i\}) = -\frac{J}{2N}\sum_{i\neq j}\sigma_i\sigma_j. \quad \textbf{[S37]}$$

Note that the sum is over all pairs, and the factor of N ensures that the energy of the system is proportional to N. The sum over all distinct pairs is missing the term $i=j$, but, because $\sigma_i^2 = 1$, we have

$$E(\{\sigma_i\}) = -\frac{J}{2N}\sum_{i,j}\sigma_i\sigma_j + \frac{J}{2} \quad \textbf{[S38]}$$

$$= -\frac{J}{2N}\left(\sum_i \sigma_i\right)^2 + \frac{J}{2}. \quad \textbf{[S39]}$$

The probability of finding the system in any particular state $\{\sigma_i\}$ is given by (choosing units where $k_B T = 1$)

$$P(\{\sigma_i\}) \equiv \frac{1}{Z}e^{-E(\{\sigma_i\})} \quad \textbf{[S40]}$$

$$= \frac{1}{Z}\exp\left[\frac{J}{2N}\left(\sum_i \sigma_i\right)^2 - \frac{J}{2}\right]. \quad \textbf{[S41]}$$

However, we can always write

$$\exp\left[\frac{A}{2}x^2\right] = \int \frac{dh}{\sqrt{2\pi A}}\exp\left[-\frac{1}{2A}h^2 + hx\right]. \quad \textbf{[S42]}$$

Applying this identity to Eq. **S41**, we have

$$P(\{\sigma_i\}) = e^{-J/2}\frac{1}{Z}\int \frac{dh}{\sqrt{2\pi N/J}}\sum_{\{\sigma_i\}}\exp\left[-\frac{N}{2J}h^2 + h\sum_i \sigma_i\right]. \quad \textbf{[S43]}$$

We can think of this, more suggestively, as

$$P(\{\sigma_i\}) \propto \int dh\, P(h)\prod_{i=1}^{N} P(\sigma_i|h), \quad \textbf{[S44]}$$

where $P(\sigma_i|h)$ describes the response of a single spin to an external field,

$$P(\sigma_i|h) = \frac{1}{2\cosh(h)}e^{h\sigma_i}, \quad \textbf{[S45]}$$

and $P(h)$ is a distribution of fields,

$$P(h) = \frac{1}{Z_h}\exp\left[-\frac{N}{2J}h^2 + N\,\ln\cosh(h)\right]. \quad \textbf{[S46]}$$

Thus, a model in which all spins interact with one another, equally, is mathematically identical to a model in which each spin responds independently to a magnetic field chosen at random.

Once we transform from interacting spins to a distribution of fields, all of the thermodynamic behavior of the system is determined by $P(h)$ (Eq. **S46**). We notice that if J is small (equivalently, if T is large), the distribution of fields is close to being Gaussian with standard deviation $\delta h = \sqrt{J/N}$. If J is large, the distribution $P(h)$ becomes bimodal, with peaks at $\pm h_0(J)$ whose locations do not depend on N; this corresponds to the spontaneous magnetization of the system. Finally, at the critical value of $J = 1$, the distribution of fields is unimodal, centered at $h = 0$, but broad,

$$P_{\rm crit}(h) \approx \exp\left[-\frac{N}{12}h^4 + \cdots\right], \quad \textbf{[S47]}$$

so that typical fields are $\delta h \approx 1/N^{1/4}$, much larger than $\delta h \approx 1/N^{1/2}$ in the high-temperature phase. In this sense, criticality is the statement that the equivalent fields have anomalously large fluctuations (25).

We can find essentially the same equivalence in a much broader class of models. Consider a collection of spins that interact through some matrix J_{ij}, so that the energy

$$E(\{\sigma_i\}) = -\frac{1}{2}\sum_{i,j}J_{ij}\sigma_i\sigma_j. \quad \textbf{[S48]}$$

The Hopfield model corresponds to the choice

$$J_{ij} = \frac{J}{N}\sum_{\mu=1}^{K}\xi_i^\mu\xi_j^\mu, \quad \textbf{[S49]}$$

where there are K stored memories,

$$\boldsymbol{\xi}^\mu \equiv \{\xi_1^\mu, \xi_2^\mu, \cdots, \xi_N^\mu\}. \quad \textbf{[S50]}$$

In this case, the same arguments that lead to Eqs. **S44** and **S46** now give

$$P(\{\sigma_i\}) \propto \int d^k\phi\, P(\phi)\prod_{i=1}^{N}P(\sigma_i|h_i), \quad \textbf{[S51]}$$

where the local fields

$$h_i = \sum_{\mu=1}^{K} \xi_i^\mu \phi_\mu, \qquad \textbf{[S52]}$$

and

$$P(\phi) = \frac{1}{Z_\phi} \exp\left[-\frac{N}{2J} \sum_{\mu=1}^{K} \phi_\mu^2 + \sum_{i=1}^{N} \ln\cosh\left(\sum_{\mu=1}^{K} \xi_i^\mu \phi_\mu \right) \right]. \qquad \textbf{[S53]}$$

As in the mean field model, if J is small, then the effective fields have a standard deviation $\delta h \approx 1/\sqrt{N}$, and, as the system approaches the critical point, this scale becomes larger by a (fractional) power of N. This shouldn't be surprising because, with K fixed as N becomes large, the Hopfield model is a mean field model (26, 27).

If we can have K independent fields, could we have as many as there are neurons in the network? This is more subtle. If we imagine that the field acting on each neuron, as defined in Eq. **S52**, is built out of N independent components, so that

$$h_i = \sum_{\mu=1}^{N} \xi_i^\mu \phi_\mu, \qquad \textbf{[S54]}$$

then we have to be careful to be sure that the typical field is bounded. Specifically, if all of the ϕ_μ have the same variance, $\langle \phi^2 \rangle$, then the variance of the field is

$$\left\langle (\delta h_i)^2 \right\rangle = \langle \phi^2 \rangle \sum_{\mu=1}^{N} \left(\xi_i^\mu \right)^2. \qquad \textbf{[S55]}$$

Clearly, we need to have $\xi_\mu \approx 1/\sqrt{N}$ to be sure that the variance of the fields is not proportional to the size of the system. Therefore, we should write $\xi_\mu = \alpha_i^\mu/\sqrt{N}$, where α_i^μ is a number of order 1. Then the correlation between the fields acting on different neurons becomes

$$\langle \delta h_i \delta h_j \rangle = \frac{1}{N} \langle \phi^2 \rangle \sum_{\mu=1}^{N} \alpha_i^\mu \alpha_j^\mu. \qquad \textbf{[S56]}$$

Now, if the influences of the different field components on the different neurons (the coefficients α_i^μ) are essentially random—e.g., some neurons are "off cells" with respect to one field and "on cells" with respect to another, with no pattern in this assignment—then the sum in Eq. **S56** is of N random numbers with zero mean, and hence the typical scale for the sum is $\langle \delta h_i \delta h_j \rangle \approx 1/\sqrt{N}$. This is not at all what one expects in a critical system. The only way to escape from this conclusion is for the terms α_i^μ to have some structure, which is equivalent to fixing some correlations among the fields acting on different spins. Put another way, if the system we are studying is equivalent to one in which N spins (or neurons) are reacting independently to N distinct fields, then criticality requires some form of correlation among these fields.

The role of correlations in critical behavior is even clearer in the general case where we have an arbitrary matrix of interactions J_{ij}. Then we can write

$$P(\{\sigma_i\}) \equiv \frac{1}{Z} \exp\left[\frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j \right] \qquad \textbf{[S57]}$$

$$= \int d^N h \; P(\{h_i\}) \prod_{i=1}^{N} P(\sigma_i | h_i), \qquad \textbf{[S58]}$$

where the distribution of fields is given by

$$P(\{h_i\}) = \frac{1}{Z'} \exp\left[-\frac{1}{2} \sum_{i,j} K_{ij} h_i h_j + \sum_i \ln\cosh h_i \right], \qquad \textbf{[S59]}$$

where

$$Z' = Z\sqrt{(2\pi)^N \det J}, \qquad \textbf{[S60]}$$

and K_{ij} is the matrix inverse of J_{ij}, $K_{ij} = (J^{-1})_{ij}$. This implies that the partition function can be written as an integral over the fluctuating fields,

$$Z \propto \int d^N h \; \exp\left[-\frac{1}{2} \sum_{i,j} K_{ij} h_i h_j + \sum_i \ln\cosh h_i \right]. \qquad \textbf{[S61]}$$

If we think about a family of models in which the interactions J_{ij} are scaled up and down in strength, e.g., with an effective temperature $J_{ij} \to J_{ij}/T$, then there is (often) a critical point at some value of the temperature T. What happens to the probability distribution of the equivalent fields, $P(\{h_i\})$, at this critical point? It is hard to answer this question in general, but, in the well-studied examples from statistical mechanics—where the elements of the network live on a regular lattice, and the matrix K_{ij} has a structure that depends only the distance between lattice points i and j, decaying rapidly so that the dominant terms connect near neighbors—the structure of $P(\{h_i\})$ near criticality approaches the structure of ϕ^4 field theory (9). Crucially, the variance of the field at a single point, $\langle (\delta h_i)^2 \rangle$, does not acquire any anomalous N dependence at the critical point. Instead, criticality is marked by the appearance of long-range correlations among the fields at different points, so that the sum of the fields over the entire sample (the $\mathbf{k} = 0$ Fourier component) does have a diverging variance.

To summarize, almost any model of interacting spins (or neurons) can be rewritten as a model of spins that respond independently to external signals; the thermodynamic behavior is then controlled by the distribution of these signals. If the number of signals is small compared with the number of elements in the network, which corresponds to a mean field model, then, away from criticality, the typical scale of these signals is small (e.g., $\sim 1/\sqrt{N}$), and the approach to the critical point involves this scale becoming anomalously large. In the more general case where the number of signals is comparable to the number of neurons, criticality is associated not with an anomalous scale for the fluctuations of any single signal but rather with large-scale correlations among these signals.

The correlations among neurons are described by a matrix $\chi_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$, and it is useful to think about the eigenvalues of this matrix. In a mean field model at criticality, or in the scenario described in ref. 25, there will be one eigenvalue separated from all of the others, which carries most of the variance of the entire system. In a system with homogeneous local interactions, the eigenmodes of χ_{ij} are Fourier modes, and, at criticality, the spectrum $\tilde{\chi}(\mathbf{k})$ diverges as $\mathbf{k} \to 0$, but continuously, so that no single mode separates cleanly from all of the others. Similarly, a mean field model is equivalent to an interacting model in which the matrix J_{ij} is of low rank (in the simplest case, rank one). Analyzing the raw data from our population of $N = 160$ neurons, we find that the largest eigenvalue of χ_{ij} captures less than 10% of the total variance, and is separated from the second largest eigenvalue by a factor of less than 2. Analyzing the models we have constructed, the spectrum of J_{ij} is nearly continuous, with no sign of a single dominant mode. These observations indicate that the network we are studying is not in the mean field regime and, more generally, that its collective behavior cannot be captured by linear dimensionality reduction strategies.

The idea that we can explain what we observe in the population of RGCs as being the result of neurons responding to other signals

evidently has clear limits. Except in special cases, assigning apparent critical behavior to such a model effectively transfers the problem to explaining the strong correlations among the signals that are driving the neurons. In this regard, it is interesting that, although details vary, we see signs of near-critical behavior in response to naturalistic movies, random checkerboards, and full-field flicker (Fig. S4). Across these different stimulus ensembles, the correlation structure of signals at the input to the retina is changing dramatically, and so, even if we think that the behavior of the ganglion cell population should be ascribed to the statistics of input signals, one has to explain how the correlations needed to mimic criticality are maintained by the retinal circuitry.

Latent Variables Redux. Aitchison et al. claim that critical phenomenology is a generic consequence of large fluctuations in latent variables (24), arguing that the behavior of mean field systems discussed by Schwab et al. (25) is typical of what we should see in complex, biological contexts. They also propose explicit candidates for the latent variables in the case of RGCs. We have argued in the preceding paragraphs that the mean field case is not the typical one in statistical physics, and is unlikely to describe the data we are discussing here. Nonetheless, one might worry that the "large variance" scenario does succeed in producing something like the critical behavior we have identified, without any of the fine tuning that one might have expected by analogy with known equilibrium critical phenomena. Here we test the suggestions of ref. 24 in more detail.

Concretely, Aitchison et al. (24) propose that all of the phenomenology of criticality should be understood in terms of models where different neurons spike or remain silent independently given the value of a latent variable that is broadcast to the entire network. Their first suggestion is that this variable is the visual stimulus itself, parameterized by time during the stimulus movie (figure 3 B and C of ref. 24). If this model is correct, then, in experiments with repeated presentations of the same stimulus, we should find zero correlations among neurons when we average over repetitions at the same moment in time. This test is not as simple as it sounds, however, because natural movies have many epochs in which the probability of one cell spiking is essentially zero. In our data, we record from $N = 160$ neurons and we have $T/\Delta\tau = 953$ distinct time bins in the natural movie, so there are $\sim 10^5$ entries in the matrix of spike probability vs. time; of these, a fraction 0.68 are consistent with zero, in that we see no spikes across 297 repetitions of the movie. Evidently, in such silent bins, one cannot estimate correlations. Conversely, a significant component of the overall correlations between two neurons may be contributed by the temporal coincidence of these silent epochs.

As an aside about silent epochs, we note that the maximum entropy model (Eqs. **3** and **4**) predicts that individual neurons should have near-zero probability of spiking when the effective field contributed by the other neurons in the network is sufficiently negative. This prediction is quantitatively correct, down to probabilities of ~ 0.001, as shown in figure 9 of ref. 11. By tracking the effective field vs. time during the stimulus movie, we can correctly predict continuous epochs of silence, characteristic of the neural response to natural stimuli, as shown in figure 15A of ref. 11.

To test the hypothesis of independence given the latent time variable, we compute the correlation coefficient between the binary variables σ_i and σ_j for every pair of cells (i, j) at a fixed time t, but only at times in the movie where each cell in the pair generates at least five spikes across the 297 repetitions of the stimulus; results are shown in Fig. S6. Although we can find moments in time where neurons that are strongly correlated across the whole experiment have near-zero correlation, we can also find the opposite. In fact, the range of correlation coefficients that we observe while conditioning on a particular time in the stimulus movie is broader than the distribution that we see in the overall correlations. It is perhaps most striking that neurons with near-zero pairwise correlation across the whole experiment can have large positive or negative correlations when conditioned on the stimulus movie, exactly the opposite of what Aitchison et al. (24) predict.

A further difficulty in testing the hypothesis of conditional independence arises from the limited size of our data set, or any reasonable data set. As we have discussed in ref. 11, the long duration of the experiments we are analyzing means that overall correlations can be estimated with high precision, and the threshold for reliable detection of a correlation is correspondingly small. However, if we are trying to estimate the correlations at a single moment in time, even uncorrelated neurons will exhibit spurious correlations with typical scale $1/\sqrt{N_{\rm reps}}$, where $N_{\rm reps}$ is the number of repetitions of the stimulus movie; even with the relatively large $N_{\rm reps} = 297$ in this experiment, we expect spurious correlations of ~ 0.05, as indicated in Fig. S6. The fact that many of the correlations we observe are smaller than this, of course, does not mean that neurons are conditionally independent but rather that we can't tell. This is a serious problem in drawing conclusions about the collective behavior of the network, because we know that widespread correlations on the order of $1/N$ can be signatures of nontrivial collective behavior (13, 28). To reliably exclude correlations on this scale at one moment in time, with no further assumptions, would require $N_{\rm reps} \approx N^2$, which rapidly becomes impossible in larger networks; even if we are more optimistic and assume that the relevant scale of correlations is $\sim 1/\sqrt{N}$, we still need $N_{\rm reps} \approx N$. This means that, with reasonable data sets on large networks, one could easily conclude that the data are statistically consistent with the hypothesis of conditional independence when, in fact, the correlations are sufficiently strong to provide the signature of dramatic collective behavior. For a different approach to this problem, which reaches similar conclusions to our Fig. S6, see ref. 29.

The second suggestion of Aitchison et al. (24) is that the relevant latent variable is the total number of spikes generated by the network, $K = \sum_{i=1}^{N} \sigma_i$ (figure 3 D and E of ref. 24). This is difficult to understand because K is a collective variable, not a latent variable. For a network with a finite number of neurons, for example, it is not possible for the activity of each cell to be independent given K; at a minimum, there must be anticorrelations that hold the number of spikes fixed. In trying to make sense out of these ideas, we have examined the correlations between pairs of cells at fixed K, and find almost all possible behaviors, including strong positive correlations (the opposite of what is required to hold K fixed) with K-dependent strengths.

Our earlier work emphasized that the distribution of K itself is anomalous, and that maximum entropy models that capture this distribution already exhibit signatures of criticality (30). Focusing on the summed activity of a network of neurons is analogous to focusing on the total magnetization of a magnet. Indeed, criticality in ferromagnets is associated with an anomalously broad distribution of magnetizations, just as the signatures that we see of critical behavior in a neural network are associated with an anomalously broad distribution of K. However, in no sense does this explain the critical phenomena. In particular, the qualitative observation of large fluctuations in magnetization is consistent with many different quantitative critical behaviors, including the mean field case where there is no divergence of the specific heat.

To summarize, the suggestion by Aitchison et al. (24) that time in the stimulus movie provides a latent variable whose variation explains the behavior that we see fails because the correlations conditioned on this latent variable are as large and structured as observed without conditioning. Their suggestion that the total number of spikes is the relevant variable confuses latent with collective variables, and we find that conditioning on this collective variable also does not simplify the correlation structure of the network. We also have examples in equilibrium statistical mechanics where (qualitatively) large fluctuations in a collective variable are associated with critical phenomena of different universality classes, so that such fluctuations alone cannot single out behaviors such as those we observe in Figs. 3 and 4.

We can also assess the claim that large fluctuations in a latent variable lead generically to critical behavior by exploring a biologically plausible model. Imagine that the sensory stimulus can be parameterized by a variable $\vec{x}$. This could represent, in the retina, the position of a single object. More abstractly, we can think about the parameters in a space of possible stimuli, so that $\vec{x}$ represents position in a "feature space." We could also imagine that we are recording from a part of the brain that represents the organism's own position in space, as with place cells in the hippocampus, in which case $\vec{x}$ is again a literal position variable. As a model, we will consider neurons such that each cell n generates spikes when $\vec{x}$ is in the neighborhood of that cell's preferred stimulus $\vec{x}_n$. More quantitatively, we give each cell a receptive field such that the probability of spiking in a small window of time is

$$p_n(\vec{x}) = P_0 \exp\left[-\frac{|\vec{x}-\vec{x}_n|^2}{2\sigma^2}\right], \qquad \textbf{[S62]}$$

and each cell is independent of the rest given the value of the stimulus $\vec{x}$. Notice that because we will be analyzing only the distribution of responses in a single small window of time $\Delta\tau$, as with our analysis of the real data, we don't need to make any assumptions about the temporal statistics of the spikes.

We focus on the simplest case, where $\vec{x}$ is one-dimensional, and take the distribution of this variable to be uniform across some interval; without loss of generality, we can take $0 < x < 1$. We assume that the N neurons have preferred stimuli x_n that are random but uniformly distributed throughout this interval. Then the only parameters to be adjusted are the width σ of the receptive fields and the peak spike probability P_0. Fig. S7 shows an example with $N = 100$ neurons, $P_0 = 0.3$, and $\sigma = 0.1$; reasonable variations in these parameters do not change the qualitative picture. We can generate long samples of data from this model, and then perform exactly the same analysis that we have done for the real neurons. We see that, although spike probabilities are being modulated in a correlated fashion across the entire population, there is no hint of Zipf's law (Fig. S7*B*), and the plot of entropy vs. energy is far from linear (Fig. S7*D*). This thermodynamic signature of criticality thus is not a generic consequence of strong driving by some latent variable.

1. Ruelle D (1978) *Thermodynamic Formalism: The Mathematical Structures of Classical Equilibrium Statistical Mechanics* (Addison-Wesley, Reading, MA).
2. Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: The characterization of strange sets. *Phys Rev A* 33(2):1141–1151.
3. Feigenbaum MJ, Jensen MH, Procaccia I (1986) Time ordering and the thermodynamics of strange sets: Theory and experimental tests. *Phys Rev Lett* 57(13):1503–1506.
4. Stephens GJ, Mora T, Tkačik G, Bialek W (2013) Statistical thermodynamics of natural images. *Phys Rev Lett* 110(1):018701.
5. Mora T, Bialek W (2011) Are biological systems poised at criticality? *J Stat Phys* 144(2): 268–302.
6. TM Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York).
7. Wilson KG (1979) Problems in physics with many scales of length. *Sci Am* 241(2): 158–179.
8. Sethna JP (2006) *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford Univ Press, Oxford, UK).
9. Parisi G (1988) *Statistical Field Theory* (Addison-Wesley, Redwood City, CA).
10. Guckenheimer J, Holmes P (1983) *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer-Verlag, New York).
11. Tkačik G, et al. (2014) Searching for collective behavior in a large network of sensory neurons. *PLoS Comput Biol* 10(1):e1003408.
12. Marre O, et al. (2012) Mapping a complete neural population in the retina. *J Neurosci* 32(43):14859–14873.
13. Schneidman E, Berry MJ, 2nd, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087): 1007–1012.
14. Mora T, Deny S, Marre O (2014) Dynamical criticality in the collective activity of a population of retinal neurons. arXiv:1410.6769 [q-bio.NC].
15. Wang F, Landau DP (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys Rev Lett* 86(10):2050–2053.
16. Perkel DH, Bullock TH (1968) Neural coding. *Neurosci Res Prog Sum* 3:221–348.
17. Mastronarde DN (1983) Interactions between ganglion cells in cat retina. *J Neurophysiol* 49(2):350–365.
18. Brivanlou IH, Warland DK, Meister M (1998) Mechanisms of concerted firing among retinal ganglion cells. *Neuron* 20(3):527–539.
19. Palmer SE, Marre O, Berry MJ, 2nd, Bialek W (2015) Predictive information in a sensory population. *Proc Natl Acad Sci USA* 112(22):6908–6913.
20. Zipf GK (1932) *Selected Studies of the Principles of Relative Frequency in Language* (Harvard Univ Press, Cambridge, MA).
21. Mora T, Walczak AM, Bialek W, Callan CG, Jr (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107(12):5405–5410.
22. Dutta P, Horn PM (1981) Low-frequency fluctuations in solids: 1/*f* noise. *Rev Mod Phys* 53:497–516.
23. van Opheusden SCF (2013) Critical states in retinal population codes, Masters thesis (Universiteit Leiden, Leiden, The Netherlands).
24. Aitchison L, Corradi N, Latham PE (2014) Zipf's law arises naturally in structured, high–dimensional data. arXiv:1407.7135 [q-bio.NC].
25. Schwab DJ, Nemenman I, Mehta P (2014) Zipf's law and criticality in multivariate data without fine-tuning. *Phys Rev Lett* 113(6):068102.
26. Amit DJ, Gutfreund H, Sompolinsky H (1987) Statistical mechanics of neural networks near saturation. *Ann Phys* 173:30–67.
27. Amit DJ (1989) *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge Univ Press, Cambridge, UK).
28. Castellana M, Bialek W (2014) Inverse spin glass and related maximum entropy problems. *Phys Rev Lett* 113(11):117204.
29. Granot-Atedgi E, Tkačik G, Segev R, Schneidman E (2013) Stimulus-dependent maximum entropy models of neural population codes. *PLoS Comput Biol* 9(3):e1002922.
30. Tkačik G, et al. (2013) The simplest maximum entropy model for collective behavior in a neural network. *J Stat Mech* 2013:P03011.

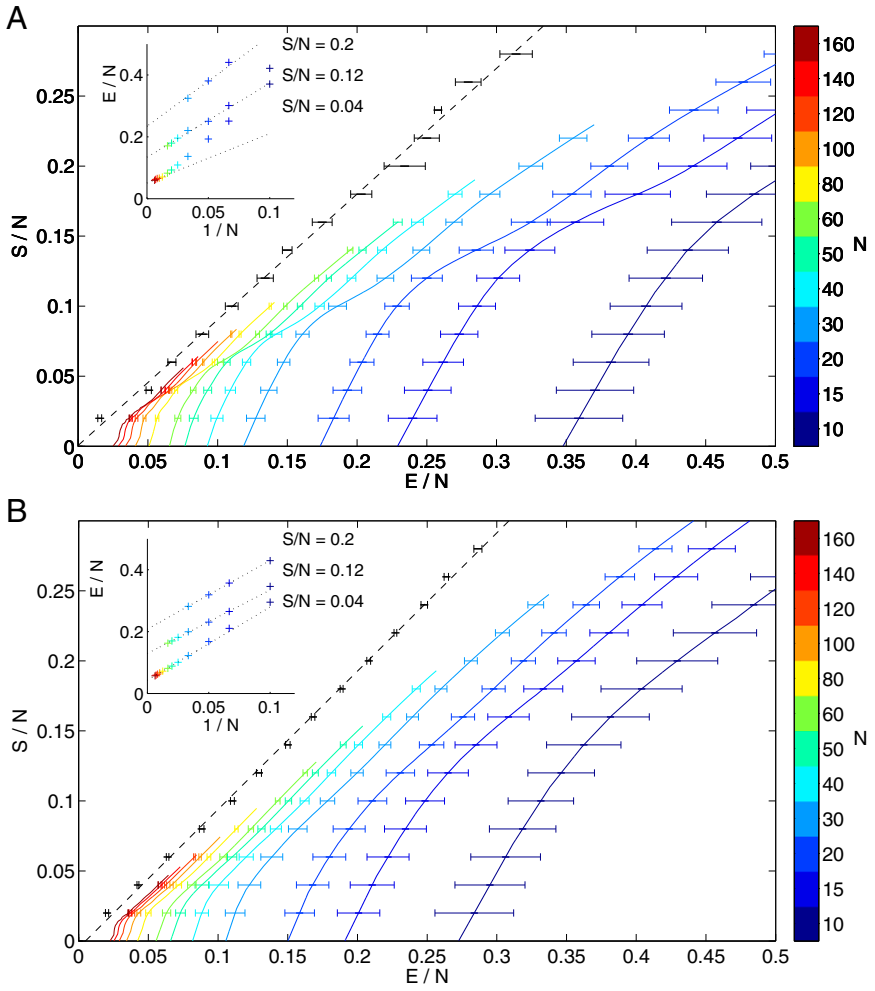

Fig. S1. Entropy vs. energy, directly from data, at different bin widths. Analysis as in Fig. 3. (*A*) $\Delta\tau = 10$ ms. (*B*) $\Delta\tau = 40$ ms.

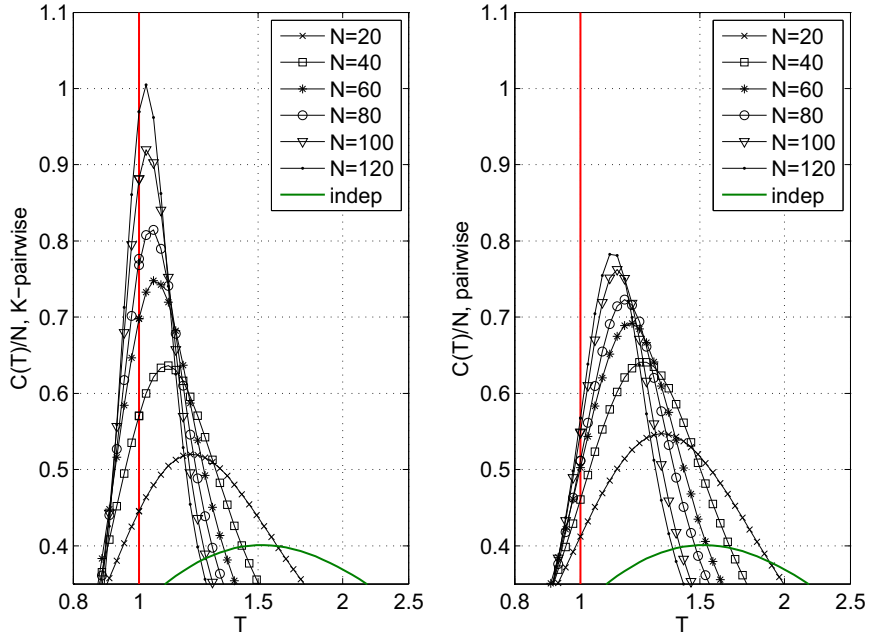

Fig. S2. Signatures of criticality are stronger in more precise models. (*Left*) Specific heat vs. *T* and *N* for the full model, redrawn from Fig. 4*B*. (*Right*) Specific heat in a maximum entropy model that matches mean spike probabilities and pairwise correlations but not the global activity distribution.

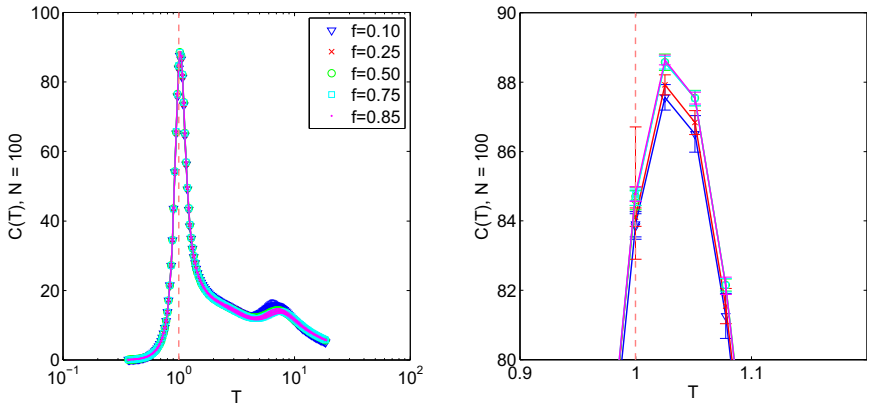

Fig. S3. Heat capacity in maximum entropy models learned from limited data. (*Left*) $C(T)$ for models learned from different fractions f of the data. (*Right*) Zoom-in on the peak near $T = 1$; error bars show SDs across different randomly chosen fractions of the data.

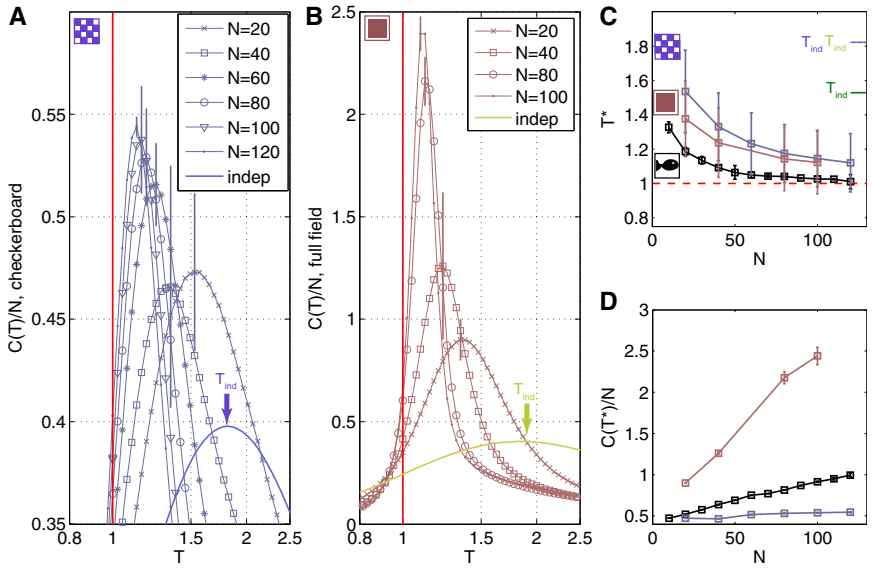

Fig. S4. Specific heat in different stimulus ensembles. (*A*) Random checkerboard stimuli, averaged over five subnetworks at each N. (*B*) Full field flicker stimuli, averaged over 30 subnetworks at each N. In *A* and *B*, arrows indicate the peak of heat capacity for a matched network of independent neurons; for clarity, error bars are shown only at the peak. (*C*) The approach of the peak of specific heat, T^*, to $T = 1$ (model reconstructions), for fish movie (black, from Fig. 4), full field flicker (brown), and checkerboard (violet). (*D*) The growth of specific heat with N for the same stimulus ensembles. Error bars in *C* and *D* are jackknife error estimates; note that these errors are much smaller in the natural stimulus ensemble.

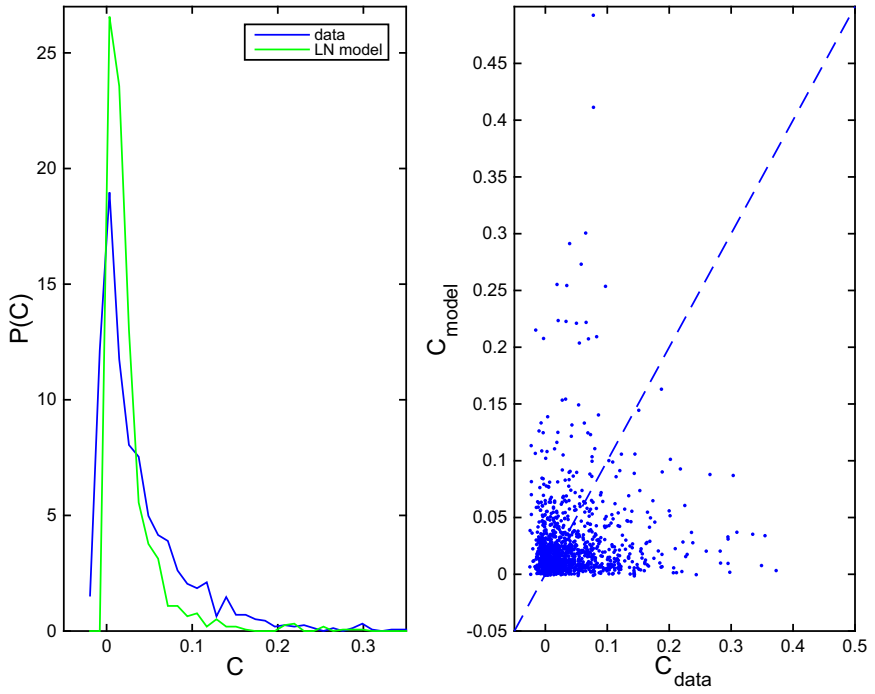

Fig. S5. LN models provide a poor account of pairwise correlations. The model in Eq. **S32** predicts correlations among neural outputs even for white noise stimuli, because of overlaps in the receptive fields. (*Left*) The distribution of these predicted correlations (green) is compared with the distribution of observed correlations (blue). (*Right*) A scatter plot of predicted vs. observed correlations, with each pair of cells in the population corresponding to one point on the plot.

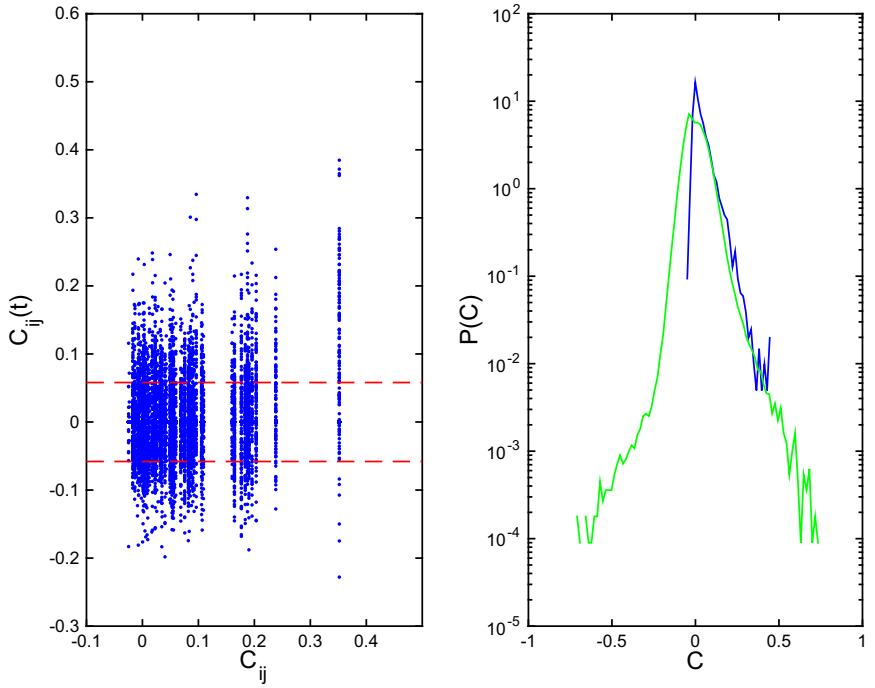

Fig. S6. Correlations at fixed time in the stimulus movie. (*Left*) A scatter plot of correlation coefficients for pairs of neurons across repetitions at fixed time [$C_{ij}(t)$] vs. the correlations across the entire experiment (C_{ij}); for clarity, we show only 100 randomly chosen pairs, and the dashed lines indicate ±1 SD of the correlations expected from uncorrelated cells in data sets of this size. (*Right*) The distribution of correlation coefficients C_{ij} (blue) and $C_{ij}(t)$ (green), now including all 12,720 pairs in the full data set.

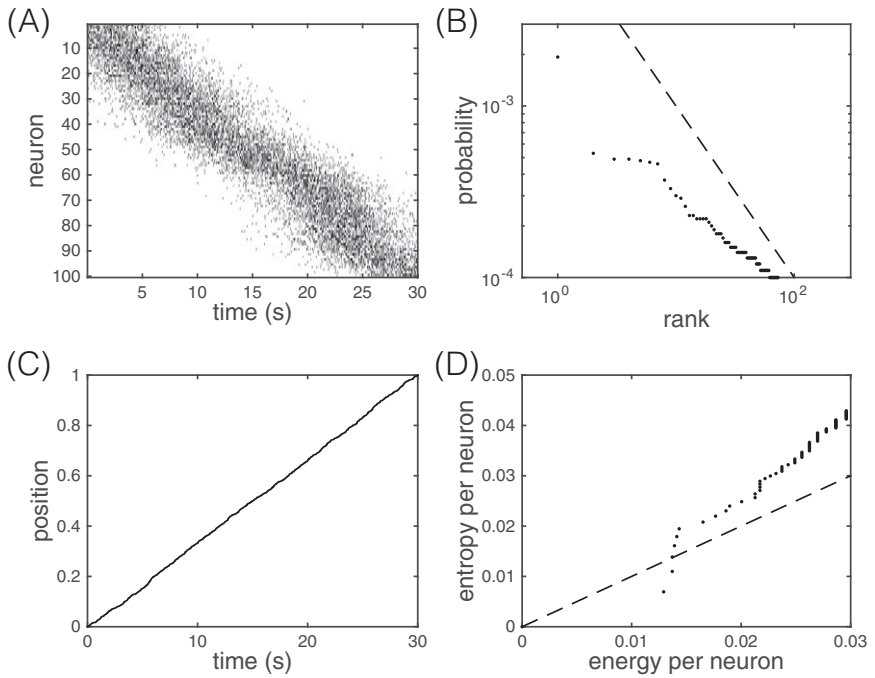

Fig. S7. Responses and thermodynamics for a population of model neurons. (*A*) Spike raster from a population of neurons with responses determined by Eq. **S62**, as the stimulus variable *x* moves along the trajectory shown in *C*. (*B*) Zipf plot—log(probability) vs. log(rank)—for the "words" describing the patterns of response in the model population of 100 cells; dashed line is Zipf's law, for comparison. (*D*) Entropy vs. energy per neuron in the model population of 100 cells, computed as for the real data in Fig. 3*A*; dashed line is of unit slope, for comparison.