MDPI

Article

Optimal Encoding in Stochastic Latent-Variable Models

Michael E. Rule [1,†], Martino Sorbaro [2] and Matthias H. Hennig [3,*]

[1] University of Cambridge
[2] Institute of Neuroinformatics, University of Zürich and ETH, Zürich
[3] University of Edinburgh
* Correspondence: mhennig@inf.ed.ac.uk

Version May 25, 2020 submitted to Entropy

Abstract: In this work we explore emergent encoding strategies learned by statistical models of sensory coding in noisy spiking networks. Early stages of sensory communication in neural systems can be viewed as encoding channels in the information-theoretic sense. However, spiking neural populations face constraints not commonly considered in communications theory. Using Restricted Boltzmann machines, we find that networks with sufficient capacity learn to balance precision and noise-robustness in order to adaptively communicate stimuli with varying information content. Mirroring variability suppression observed in sensory systems, highly informative stimuli are encoded with high precision, at the cost of highly variable responses to frequent, hence less informative stimuli. We find that this coding strategy corresponds to statistical criticality in the neural population code, and emerges at model sizes where the input statistics are well captured. These phenomena have well-defined thermodynamic interpretations, and we discuss their connection to prevailing theories of coding and statistical criticality in neural populations.

Keywords: information theory; encoding; neural networks; sensory systems

1. Introduction

The rate at which information can be conveyed by a finite neural population is limited. Neurons have a maximum firing rate, and spiking communication is affected by noise. The spiking output of a neural population can therefore be viewed as a noisy communication channel in the information-theoretic sense. Neural communications channels must signal reliably under a range of operating conditions. For example, the optic nerve carries visual information from the retina to the brain. The capacity of the optic nerve is fixed, but the amount of information carried along it is not. Certain stimuli can contain more behaviorally-relevant 'bits' than others. To utilize sensory information, the brain must therefore find efficient coding strategies [1]. How might a noisy spiking communications channel structure its available spiking "code words" to reliably communicate stimuli with differing amounts of information?

In conventional communications channels, deriving optimal codes is straightforward: the channel bandwidth is equal to the nominal bandwidth minus the entropy of any noise on the channel. The optimal code-word allocation is given by entropy coding [2], in which the cost (in bits) of a symbol with probability p should be roughly $-\log_2(p)$. Optimal coding strategies are more subtle in spiking channels, since the amount of noise depends on the symbol being transmitted: spiking variability is higher when neurons spend more time close to firing threshold. In addition, limited encoding bandwidth favours models that capture salient latent causes underlying sensory inputs [3–5].

In this work, we used Restricted Boltzmann Machines (RBMs) to study optimal encoding in stochastic spiking channels. Such models balance biological realism and theoretical accessibility. The

stochastic and binary nature of RBMs resembles physiological constraints on spiking communication, while the interpretation of RBMs as Ising spin models also allows access to information-theoretic and thermodynamic quantities [6–9].

We organize this work as follows. We first detail an RBM model of sensory encoding and present evidence of an optimal population size for capturing stimulus statistics. We then show that stimulus-dependent suppression of 'neuronal' variability is an essential feature of the learned encoding strategy in sufficiently large populations. We observe that this corresponds to statistical criticality in the population code, a feature not inherited from the stimulus statistics. By examining a thermodynamics interpretation of the RBM models, we show that statistical criticality connects to the optimization of the underlying network parameters, and that it suggests an optimal model size that balances accuracy verses the number of neurons used for encoding. We conclude with a discussion of the connection between the statistical machine-learning approach used here and other prevailing theories of sensory encoding.

2. Results

2.1. RBMs as a statistical machine-learning analogue of stochastic spiking communication

Restricted Boltzmann Machines (RBMs; Fig. 1a) are stochastic binary neural networks used in statistical machine learning [10]. They consist of two populations of stochastic binary units. One population, the 'visible' layer, is driven by incoming sensory stimuli. The other population, a 'hidden' layer, learns to encode the latent causes of these stimuli. These hidden units can therefore be interpreted as a stochastic spiking communication channel that conveys information about incoming stimuli.

In the RBM, processing of sensory input consists of a linear-nonlinear transformation of a stimulus vector (v) that determines the probability that units in the hidden layer 'spike' (i.e. emit a '1'):

$$\Pr(h_i = 1) = \sigma(W_i v + B_{h_i}), \tag{1}$$

where $\sigma(a) = [1 + \exp(-a)]^{-1}$ is a logistic sigmoid nonlinearity, W is a matrix of 'synaptic' weights between the visible and hidden layers, and B_{h_i} is a per-unit bias that sets the baseline firing rate for hidden units h_i.

In addition to retaining phenomenological aspects of spiking population coding, RBMs can be trained readily using the contrastive divergence algorithm [10,11]. We trained RBMs on binarized regions of natural images in order to study emergent learned encoding strategies (Fig. 1a; Methods §4.1-4.2). We evaluated a range of population sizes for the hidden layer (Fig. 1b-e) to study how encoding strategies change with network size. Small networks did not accurately model the stimulus distribution (Fig. 1b,c), and a minimum population size of ≈30 units was necessary to faithfully model small binary image patches from the CIFAR dataset. Network activity became increasingly sparse (Fig. 1d) and uncorrelated (Fig. 1e) for larger hidden-unit population sizes, mimicking the sparse spiking activity of biological neural networks.

It appears that sufficiently large RBMs can learn stochastic spiking representations of incoming stimuli. We next examine these learned encoding strategies in depth in order to answer two related questions. First, can we understand general principles of sensory encoding in stochastic spiking populations based on the encoding strategies learned by these models? Second, what are the statistical correlates of a model being "sufficiently large" that could be used to identify the optimal population size required for good representations?

2.2. RBMs provide an energy-based interpretation of spiking population codes

For models that capture the stimulus distribution well, we would like to understand how the network allocates its coding space: how do 'visible' stimuli map to spiking patterns in the latent 'hidden' layer, and vice-versa? The limited number of hidden units favors precise neural codes, in

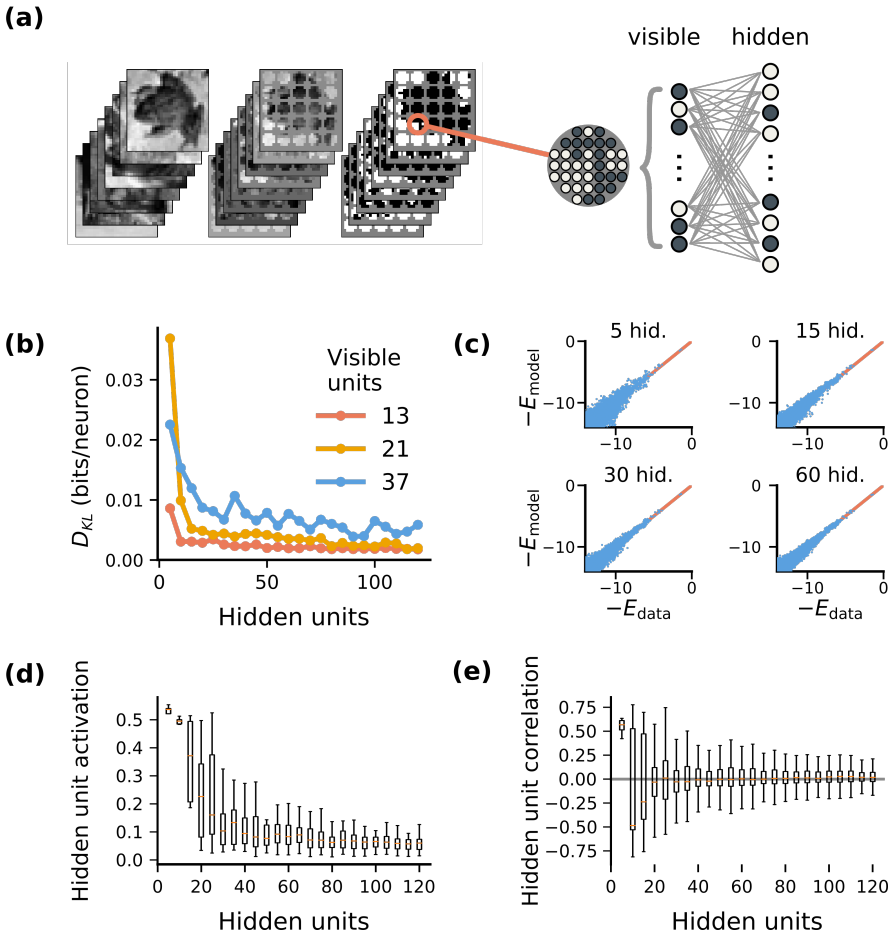

Figure 1. *Effect of the channel size on encoding of stimulus statistics.* (**a**) We trained RBMs to model local regions of (binarized) CIFAR-10 images. We interpret the number of hidden units as the size of a sensory communication channel. (**b**) A minimum number of hidden units is required to faithfully capture stimulus statistics. We quantified model accuracy by the Kullback-Leibler divergence between model samples and held-out training data. Accuracy improves as the hidden-layer size increases, up to a point. Results for three different sizes of stimulus patches (13, 21, 37 pixels) are shown. (**c**) Comparison of actual and predicted pattern probabilities for four hidden-layer sizes. We denote probability in terms of the negative log-probability (in bits), abbreviated as energy $\mathrm{E} = -\log_2 \Pr(\cdot)$. Larger models capture the stimulus distribution better. (**d,e**) Hidden-layer activation becomes sparser (d) as model size increases, and more decorrelated (e). 13 visible units were used for c-e.

which specific stimuli reliably evoke a specific pattern of neuronal spiking. However, noise can limit coding precision, requiring multiple neural states to map to each stimulus to achieve robustness.

Overall, two strategies are available for increasing information content in stochastic spiking codes. Neurons can become reliable, and use precise codes with less noise. Neurons can also increase their firing rates. These strategies have natural analogues in information theory. Increasing codeword precision amounts to decreasing the conditional entropy of evoked neural activity, i.e. reducing the channel noise. Using higher firing rates amounts to increasing the 'energy' of the neural codes, which is equivalent to using longer symbols (or more bandwidth) in a conventional digital code. Hinton et al. (1995) [12] first noted this in the context of spiking latent-variable models, showing that in optimal codes the amount of information in a stimulus should match the average information in the latent spiking pattern minus the entropy (i.e. variability) in that evoked pattern. However, the question remains of how an optimized spiking channel might make use of these two encoding strategies.

Here, we assume that the sensory channel represents all stimuli equally, so that the amount of behaviorally-relevant information in each stimulus is indeed equal to its negative log-probability. This reflects the number of bits required to communicate it in an optimal code in Shannon sense. In reality, early stages of sensory processing filter and discard information, preserving only important details. This issue is minor, however, since one can consider the stimuli in the simulations here as reflecting only the behaviorally-relevant bits.

To explore this, let us first make precise these notions of 'energy' and 'entropy' in the trained RBM networks. For an RBM with weight matrix W and hidden and visible biases B_h and B_v, the probability of any population activity state (h, v) can be written as:

$$\begin{aligned} \Pr(h, v) &= \exp\left(-\mathrm{E}(h, v)\right) \\ \mathrm{E}(h, v) &= -B_v^\top v - B_h^\top h - h^\top W v + \text{constant}, \end{aligned} \tag{2}$$

where $E(h, v)$ is the energy of the state (h, v). Throughout this paper, we will use the term 'energy' synonymously with negative log-probability.

We adopt the compact notation of Dayan et al. (1995) [13], and write energy of a state (h, v) as $\mathrm{E}^{\phi}_{h,v}$, where $\phi = \{W, B_h, B_v\}$ are the model parameters. Probabilities are denoted similarly, and we use Q to denote the distribution of latent factors learned by the RBM network. In this notation, the stimulus-evoked *entropy* of the hidden-unit spiking h given a specific stimulus v is

$$\mathrm{H}^{\phi}_{h|v} = \sum_h \mathrm{Q}^{\phi}_{h|v}\, \mathrm{E}^{\phi}_{h|v} = \left\langle \mathrm{E}^{\phi}_{h|v} \right\rangle_{h|v} \tag{3}$$

Above, $\langle\cdot\rangle_{h|v}$ denotes expectation with respect to the distribution of stimulus-evoked spiking activity in the latent units, $\mathrm{Q}^{\phi}_{h|v}$. In this notation, optimal representations are achieved when the amount of information in a stimulus (E_v) matches the amount of information in the evoked spiking activity minus the entropy of any "noise" in the channel ($\mathrm{H}_{h|v}$) :

$$\mathrm{E}_v = \left\langle \mathrm{E}_{h,v} \right\rangle_{h|v} - \mathrm{H}_{h|v}\,. \tag{4}$$

(c.f. Eq. 5 in Hinton et al. 1995 [12]). In practice, the optimization procedure identifies parameters ϕ that only approximately achieve the above relationship. For models that are too small, not all stimuli are equally well-encoded, as reflected in the increased Kullback-Leibler divergence in the fits for smaller models (Fig. 1b).

2.3. Stimulus-dependent variability suppression is a key feature of optimal encoding

We calculated the stimulus-evoked energy and entropy for a range of network sizes (Methods §4.3). In Fig. 2 we examine how these quantities vary a function of stimulus energy. Here, stimulus energy is equivalent to negative log-probability ($\mathrm{E}_v = -\log \mathrm{P}_v$), and reflects the amount of information

(in bits) needed to specify a particular stimulus. This can be interpreted as the bitrate required to convey a stimulus, and groups of stimuli with similar energy therefore reflect different bitrates required of the sensory communication channel.

We found that RBMs learned to reserve the highest-bandwidth (low noise) parts of coding space for high-information stimuli. This can be seen in Figure 2a, which shows that the stimulus-evoked entropy in the latent spiking activity is reduced when higher bitrates are needed, provided the channel is sufficiently large ($\geq$35 units). This reflects an adaptive code that lowers neuronal variability when more bandwidth is required. Conversely, stimuli that require less bandwidth are represented using noisier parts of encoding space.

Curiously, we found that channels that were too small to properly encode all stimuli (<35 units) exhibited the opposite trend: the reliable parts of coding space are allocated to low-information stimuli. This suggests that variability suppression may emerge above a critical model size, and might be a useful correlate for optimizing the number of latent units in the channel.

To communicate more information, neural codes can either reduce noise ($H_{h|v}$), or they can use more informative code-words. In optimal Shannon coding, more informative codewords are simply rarer (information is negative log-probability), and correspond to specific spiking patterns reserved for rare stimuli. One can summarize how "rare" the spiking patterns for a particular stimulus are in terms of the average energy of the evoked codewords, which we denote as $\langle \mathrm{E}_h^{\phi} \rangle_{h|v}$. Intuitively, $\langle \mathrm{E}_h^{\phi} \rangle_{h|v}$ is the average number of bits needed to specify a particular codeword h evoked by stimulus v (if we do not know v in advance). That is, it is the amount of information needed to encode the stimulus-evoked neural states.

We expected $\langle \mathrm{E}_h^{\phi} \rangle_{h|v}$ to increase for higher stimulus bitrates, but found instead that it closely tracked variability ($H_{h|v}$), decreasing for stimuli that required more bits to communicate. Indeed, above a critical model size ($\geq$35 units in this case), the stimulus-evoked entropy and energy tracked

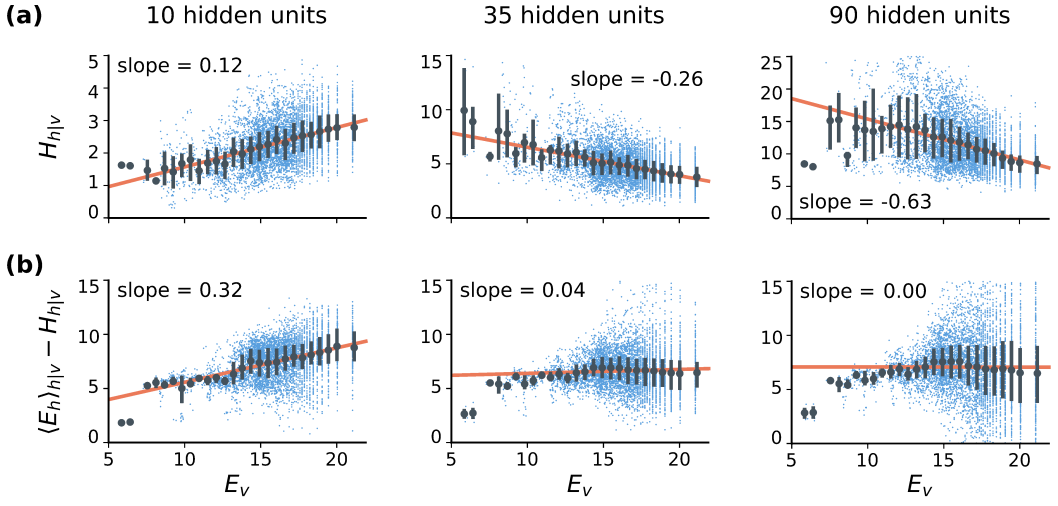

Figure 2. *Informative stimuli suppress variability in stochastic spiking communication channels.* Here, we trained RBMs to encode 13 visible units from circular patches of binarized CIFAR-10 images. Plots show how the statistics of the evoked activity in hidden units (vertical axes) varies as a function of stimulus information content (horizontal axes). Larger 'energies' (E_v) represent stimuli (blue dots) that require more bits to communicate. All units are in bits. (**a**) Sufficiently large models learn to reduce channel entropy (variability) for stimuli that require more information to codify. (**b**) To communicate more information, neural codes can either reduce stimulus-conditioned entropy $H_{h|v}$, or they can use rarer code-words, i.e. increase $\langle E_h \rangle_{h|v}$. In sufficiently large models, we find that energy and entropy both decrease for stimuli that require more information to communicate. (gray bars; dots=mean, bars=inter-quartile range).

each-other with a 1:1 ratio. This is illustrated in Fig. 2b, which plots the difference between these two quantities over a range of stimulus bitrates.

Surprisingly, this 1:1 balance between energy and entropy corresponds to statistical criticality and the emergence of $1/f$ power-law statistics in the latent spiking activity. Criticality in the brain has been the subject of some controversy over the past decades [14,15], and we unpack this observation in more depth in the following sections.

2.4. Optimal codes exhibit statistical criticality

When we say that a collection of observations exhibit "statistical criticality", we mean that they are consistent with being generated by a physical process that lies close to a phase transition in the thermodynamic sense. At first glance, it is unclear how the allocation of codewords in a stochastic spiking code might be related to criticality, or why this relationship might be interesting from the standpoint of neural coding.

Historically, the study of statistical criticality in neural systems was motivated by theories that suggest that dynamical regimes close to a phase transition might be useful for processing information [9]. Indeed, several studies have suggested evidence of statistical criticality in neural data [16–18]. However, other studies call the significance of this into question [19], showing that these statistics can arise under very generic circumstances [20], might be inherited from the environment [21], and could even be a data-processing artefact [22]. We hope to clarify some of this controversy by examining the emergence of statistical criticality in this *in silico* model of spiking population coding.

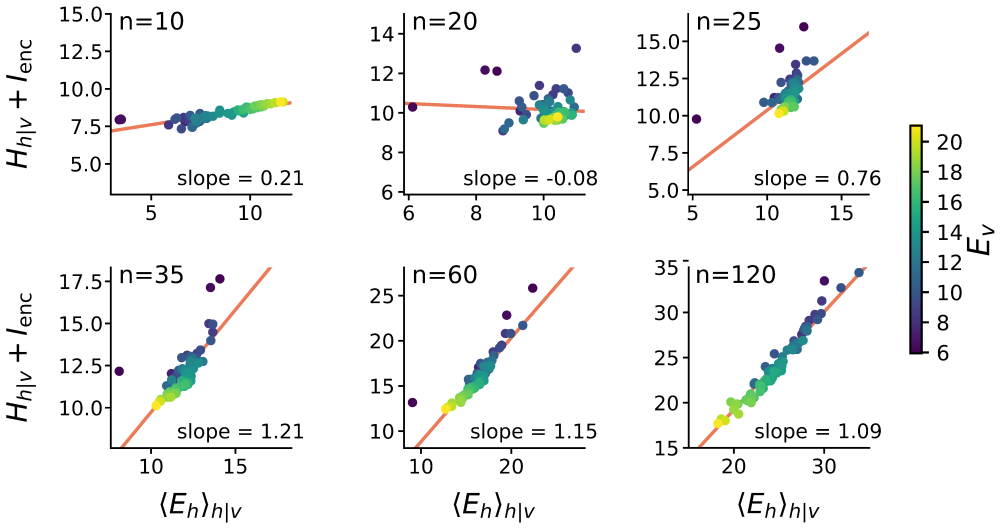

Figure 3. *Stimulus information content predicts energy and entropy of evoked activity in latent units.* Each plot shows the average stimulus-evoked entropy ($H_{h|v}$) plus a constant (I_{enc}) on the vertical axis, against the information content of the code-words evoked by a given stimulus ($\langle E_h \rangle_{h|v}$ horizontal axis). Here, $I_{\text{enc}}=\langle D_{\text{KL}}(Q_{h|v}\|Q_h)\rangle$ is the average energy-entropy relationship for all stimuli, which becomes approximately constant above a critical model size (Fig. 2b). Color indicates the stimulus bitrate E_v. Points reflect the average energy and entropy of hidden patterns evoked for a given E_v. In too-small models (n=10), low-variability states are used to represent common (low-information) stimuli. This relationship shifts as the encoding capacity increases (n=20,25). Above a critical model size (n$\geq$35), an inverse relationship between visible energies and the entropy of latent representations emerges: high-energy visible patterns suppress variability. A 1:1 trade-off between using energy and entropy for modulating bit rate also emerges (red lines). This relationship persists in larger models (n=60,120). This 1:1 trade-off reflects emergence of a 1/f power-law in the statistics of hidden unit activity, which gives rise to statistical criticality. Here, models were trained to encode 13 visible units from circular patches of binarized CIFAR-10 images.

Figure 3 illustrates how the energy and entropy of stimulus-evoked activity varies as a function of stimulus bitrate. We group stimuli into sets "$\mathcal{V}_E$" of similar energy, which correspond to different bitrates required of the spiking channel. For each $\mathcal{V}_E$, we plot the average stimulus-evoked entropy $\mathrm{H}_{h|v}$ (a correlate of the spiking noise), and energy $\langle \mathrm{E}_h^{\phi} \rangle_{h|v}$ (the average number of bits required to specify a particular evoked spiking pattern). To more clearly illustrate the scaling, the entropy is shifted by a constant I_{enc} which reflects the average difference between energy and entropy. For this particular set of stimuli, models with at least 35 hidden unit exhibit a positive correlation between energy and entropy, with a slope that approaches one as the model size increases.

This relationship corresponds to the so-called "Zipf's law" [15]. Zipf's law refers to the frequency (f) of symbols in a dataset. Here, the symbols are the spiking "codewords" (h) in the hidden units. Zipf's law states that frequency of any symbols is inversely proportional to its rank in the frequency table. I.e. the rarer a pattern is the more patterns there are of similar frequency. For example, in a dataset exhibiting Zipf's law we would expect approximately 2^E patterns with frequency above 2^{-E} (up to some multiplicative constant). These statistics are especially curious in the context of the RBM, which can be interpreted as a type of Ising spin model. Ising spin models at a critical point exhibit Zipf's law in their distribution of states [9,15].

We confirm that the 1:1 variation in entropy and energy observed here corresponds to Zipf's law in the codeword frequencies in Figure 5a. The stimulus-evoked entropy $\mathrm{H}_{h|v}$ determines the number of hidden codewords h that correspond to a given stimulus v. Loosely, one can think of a stimulus as eliciting $2^{\mathrm{H}_{h|v}}$ possible patterns. Likewise, the "energy" of a hidden codeword E_h is proportional to its negative log-probability. In the models examined here, the energy and entropy of stimulus-evoked spiking patterns vary similarly as a function of stimulus energy E_v, giving rise to Zipf's law in the frequencies of population spiking patterns. This means that, as the neural code gets noisier, the probability of any specific codeword also decreases. Overall then, we find that stimuli that are encoded in the "noisier" parts of coding space are allocated over a larger pool of increasingly rare, but representationally equivalent, codewords. This strategy is essential for reserving the reliable parts of the coding space for high-information stimuli, while also using a robust code to communicate low-information stimuli in noisier parts of the coding space.

Here we show that statistical criticality emerges naturally in a model of stochastic spiking encoding, but only for models that are large enough to capture the stimulus distribution. Our use of a ground-truth model simulation ensures that these statistics are not an artefact of recording or data-processing [22]. A natural question, however, is whether these statistics arise from the statistics of natural images, which also exhibit Zipf's law [21]. In Figure 4 we confirm that this is not the case, as models trained on synthetic visual stimuli designed to have other statistics still exhibit $1/f$ power-law statistics in latent unit activity.

Many processes can generate similar statistics, and while criticality implies $1/f$ statistics, the converse is not necessarily true [20,24,25]. We next therefore asked whether the observed statistics are associated with true criticality in the thermodynamic sense, and whether this tells us anything significant about the model optimization and the learned encoding strategies.

2.5. Evidence for an optimal population size

So far, we have demonstrated that Zipf's law emerges in optimized RBM models of spiking population codes. Should we attribute any significance to these statistics? Do they imply anything meaningful about the underling spiking population code, or could they arise from more mundane explanations [14,20]? To address these questions in depth, we leverage the fact that the RBM can be interpreted as a thermodynamic system. This means that one can define signatures of a true phase transition, and therefore examine whether these critical statistics imply anything meaningful with respect to model parameters and their optimization.

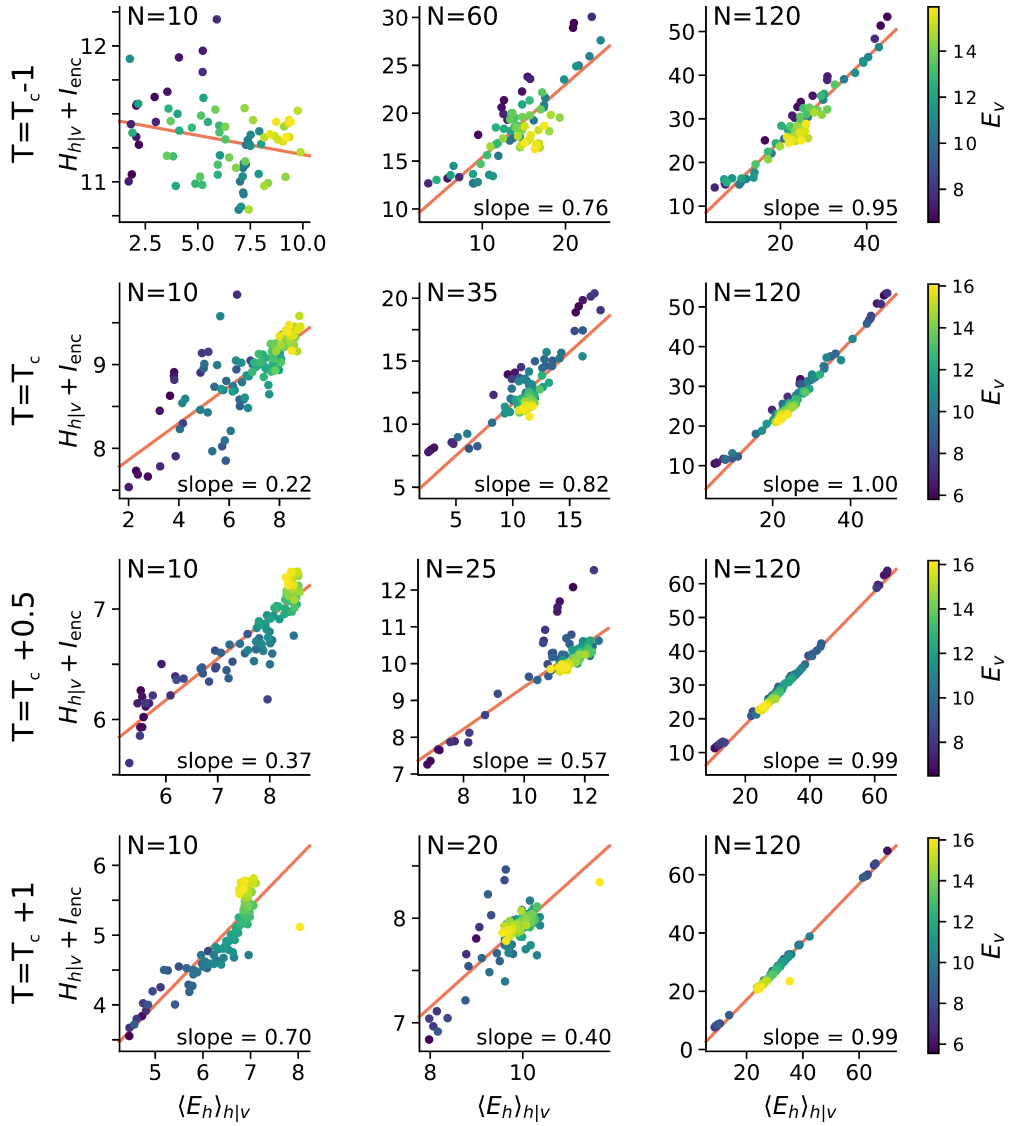

Figure 4. *Learned encoding strategies do not depend on the statistics of the stimulus distribution.* In natural visual stimuli, the visible samples themselves display 1/f power-law statistics. This might encourage similar statistics in the activations of hidden units, explaining the 1:1 trade-off between modulating entropy and energy that we observed. Here we show the energy-entropy balance as a function of stimulus information content (i.e. bit-rate, E_v) for RBMs fit to two-dimensional lattice Ising models, sampled at a range of temperature above and below the critical temperature of $T_c = 2/\ln(1+\sqrt{2}) \approx 2.269$. The energy-entropy balance converges to identity regardless of the data temperature (right column). However, the critical hidden-layer size (N) does decrease with temperature, illustrated here (middle column) by the increasing hidden-layer size displaying intermediate energy-entropy statistics. Small models (left column) exhibit a correlation between visible energy and entropy for training-data temperatures above T_c. Ising models were simulated on a 10×10 grid, and sampled via the Swendsen-Wang algorithm [23] with 10k steps burn-in and 100k training patterns drawn every 100 samples. 13-unit patches were presented to the RBM for training. All units are in bits.

To explore the thermodynamic interpretation of the RBM, one can extend the energy-based definition of the RBM (Eq. 2) to include an inverse temperature parameter β:

$$\mathrm{P}_{h,v} \propto \exp\left(-\beta\, \mathrm{E}_{h,v}\right). \tag{5}$$

This corresponds to scaling the biases and weights by β, and controls a single direction in parameter space that determines how ordered or disordered the spiking activity is. High temperatures ($\beta \to 0$) corresponding to a noisy phase where the probability of all states are equal. Low temperatures ($\beta \to \infty$) exhibit only a few fixed patterns. Critical models exists at a specific temperature $\beta = 1/T_c$ that defines a transition between the these two phases.

To generalize this idea, we can study the Fisher Information Matrix (FIM), which defines a local measure of "importance" to various directions in the space of RBM parameters $\phi = \{W, B_h, B_v\}$. The FIM provides an infinitesimal equivalent of the Kullback-Leibler divergence between the model and a neighbouring model, which differs by an infinitesimal deviation in the parameter space, and is defined as follows:

$$F_{ij}(\phi) = \sum_{v,h} P_{v,h} \frac{\partial^2 E_{v,h}}{\partial \phi_i \partial \phi_j}. \tag{6}$$

For RBMs, one can calculate the FIM from the activity statistics (Methods §4.4). The FIM is a generalized measure of susceptibility or specific heat [26], and it diverges at the point of phase transition $\beta = 1/T_c$. Intuitively, this is because the model's statistics change abruptly at the critical temperature, where the model's behavior as a function of parameters approaches a non-differentiable point with infinite curvature (i.e. diverging FIM) for increasing system sizes. For small, finite models, there is no true phase transition per-se. Instead, the FIM exhibits a local peak around $\beta = 1/T_c$ which indicates the finite-size analogue of a critical temperature [26].

One can assess whether a given model is close to a phase transition by examining the structure of the FIM for a range of temperatures. Analyzing the behavior of the largest FIM eigenvalue is analogous to studying the divergence of specific heat [9], but its interpretation is more general. In Figure 5a we find that a local peak in the maximum FIM eigenvalue (the direction in parameter space with the largest curvature) emerges for models with ≥ 30 units. This is also the model size at which statistical criticality emerges (Fig. 3, 5a rightmost column). We conclude that the emergence of statistical criticality corresponds to a true critical point in the thermodynamics sense. Empirically we find that models that are sufficiently large to fit the data exhibit a localized peak in the FIM curvature for β=1. We conjecture that these statistics might be useful in identifying the optimal model size that balances accuracy vs. size cost.

Above the model size at which criticality emerges ("critical model size"), we find diminishing returns in terms of model accuracy (Fig. 1b,c). We examined the structure of the FIM to determine whether the model exhibited "sloppy" [27–29] parameters that might be removed without degrading accuracy. Indeed, we found that many single units or weights become relatively unimportant in larger models (Fig. 5a) . This suggests that the activity statistics may reveal superfluous neurons or synapses that could be removed or "pruned" with relatively little damage to the network's function.

However, the parameter importance as assessed by FIM should be interpreted with caution. We find that the least "important" units, in terms of FIM curvature, have receptive fields corresponding to complex or high spatial-frequency features (Fig. 5d). These units therefore encode fine details of images. While removing a single unit might have a minor effect of the model accuracy, collectively many unimportant units may be necessary for maximising the encoded information.

In these simulations, the emergence of Zipf's law can be connected to an energy-based description of spiking correlations that lies close to a phase transition, and is not inherited from the statistics of the stimuli, nor is it a data-processing artefact. While neural networks *in vivo* do not use the optimization procedure that we used here, any learning procedure that adapts its internal states to optimally encode

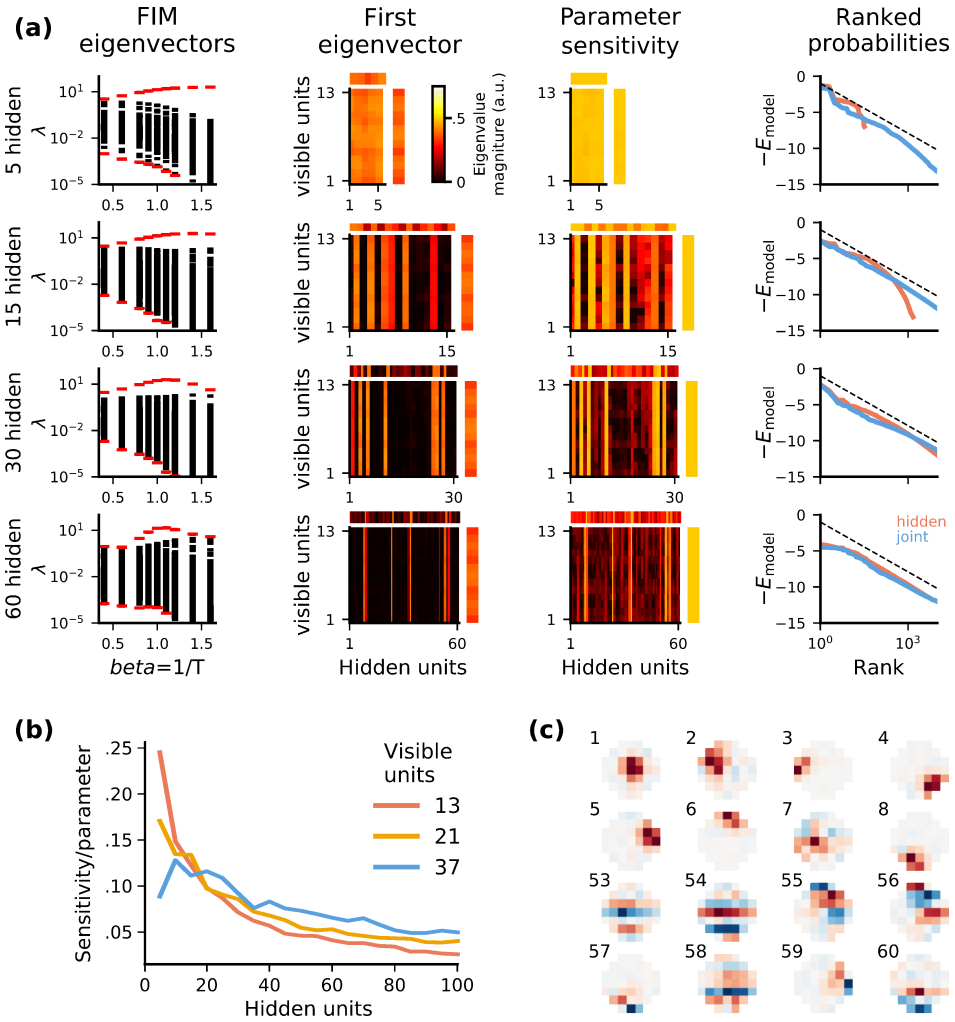

Figure 5. *Analyses of parameter sensitivity suggests an optimal model size for encoding sensory statistics.* (**a**) Analysis of the Fisher Information Matrix (FIM) over a range of hidden-layer sizes (top to bottom; 13 visible units). From left to right, (1) FIM eigenvalue spectra λ_i (y-axis) over a range of inverse temperatures β indicate that model fits (β=1) past a certain size lie at a peak in their generalized susceptibility. This is a correlate of criticality in Ising spin models. Eigenvalues below 10^{-5} are truncated, and the largest and smallest eigenvalues are in red; (2) Important parameters in the leading FIM eigenvector align with individual hidden units, and become sparse for larger hidden layers. The eigenvector is displayed separately for the weights (matrix), and the visible (vertical) and hidden (horizontal) biases; (3) The average sensitivity of each parameter over all FIM eigenvectors, shown here as the square root of the FIM diagonal, also shows sparsity, indicating that beyond a certain size additional hidden units contribute little to model accuracy. Data is shown as in column 2; (4) Variance of the hidden unit activation as a function of stimulus energy. In larger models, units with sensitive parameters contribute to encoding low energy, less informative patterns. (**b**) The average sensitivity of each parameter, measured by the trace of the FIM, normalized by hidden-layer size, decreases as hidden-layer size grows. (**c**) Hidden unit projective fields from a model with 37 visible and 60 hidden units, ordered by relative sensitivity (rank indicated above each image). More important units (ranks 1-8) encode spatially simple features such as localized patches, while the least important ones (ranks 53-60) have complex features.

the external world should (approximately) optimize representational cost (Eq. 4; so-called "free-energy" minimization [12,13,30,31]). These modelling results raise the question of whether statistical criticality is a natural outcome of this optimization, and whether it can be interpreted as an adaptive strategy to represent stimuli with variable bitrates in a stochastic neural code.

3. Discussion

Understanding neural population codes in the context of communications theory is challenging, since stochastic spiking channels differ in many aspects from the communications channels studied in engineering. In this work, we used restricted Boltzmann machines to study optimal encoding in stochastic spiking channels. Analogously to sensory systems, such models learn to encode the latent causes of natural images in terms of a stochastic binary representation. Although different stimuli require different number of bits to encode, the number of hidden units available for this representation is fixed, and different parts of the encoding space exhibit more channel noise than others.

Under these constraints, RBMs learned to represent higher bitrate stimuli by suppressing variability, which mirrors the behavior of *in vivo* neural populations [32]. Surprisingly, we found that high-information stimuli were often associated with lower energy code-words, a result which may connect to the synergy-by-silence observed in the retina [33]. This coding strategy can be explained by a competitive allocation of encoding space in a stochastic channel. Noise is largest when neurons are close to firing threshold, and so the noisiest parts of activity space exhibit intermediate firing rates. To handle higher bitrates it is necessary to signal reliably, and also to avoid overlapping with these noisy parts of the coding space. Suppressing firing in a selective population of cells is one way to achieve this.

A central prediction of this coding strategy is that common (low information) stimuli are associated with less precise (more noisy) encoding. It would be interesting to revisit data recorded from sensory systems such as the retina, to see if the effective stimulus bitrate predicts the observed neuronal variability. This result also highlights that that the fundamental unit of "neural coding" is not a specific pattern of spiking activity *per se*. We found that many stimuli can be encoded by a large volume of equivalent spiking population codewords. The equivalence between different evoked spiking patterns ensures robust representations despite noise.

We found that variability suppression corresponded to the emergence of Zipf's law in the spiking population statistics. We explored this further, since criticality in neural codes has been the subject of intense debate. We found that statistical criticality was not inherited from the stimulus distribution. Criticality was also not an epiphenomenon arising from any of the more common-place theories that have been put forth. Instead, the emergence of Zipf's law was a signature of the underlying system lying close to a phase transition, and this regime correlated with the emergence of optimal encoding strategies.

Spiking systems can also exhibit statistical criticality in the sparse, large-network limit [9]. In contrast, the statistical criticality observed here emerges abruptly at a finite optimal model size, which depends on the data being encoded (Fig. 4), and correlates with the channel learning to modulate variability based on stimulus bitrate. This association of critical statistics with the modulation of bitrate is connected to Aitchison et al. [20], which notes that that $1/f$ power-law statistics arise whenever data are generated from hidden underlying causes. If the observed code-words arise from a mixture of different explanations, it can cause the overall distribution to exhibit scale-free statistics corresponding to Zipf's law. Our modelling work reveals a specific example of this phenomenon in systems that encode the external world. Here, we found that the bitrate of the underlying stimulus is the underlying, unobserved variable, and that statistical criticality signifies an adaptive strategy for handling variable bitrates in a stochastic spiking channel.

Theoretical work predicts that critical $1/f$ statistics might be common in large latent-variable models like the RBM [34,35]. These existing results, however, were derived in the limit of an infinite (or at least very large) number of hidden units. In contrast, we found that criticality emerges in small

models—but only if there are enough latent units to accurately encode the stimulus distribution. The result of Mastromatteo et al. (2011) [34] in particular shows that most large random spin models lie close to a phase transition, simply by chance. If statistical criticality is, in a sense, the default, this implies that there is something interesting about models that *do not* exhibit statistical criticality. We found that the absence of criticality was a symptom of a model being too small to properly explain the stimulus distribution, and corresponded to an inability to vary the channel bandwidth as needed. More generally, departure from $1/f$ statistics may reveal important clues about physiological constraints on, or the operating regime of, stochastic spiking channels.

In conclusion, we found that statistical machine learning models of spiking communication employ variability-suppression as an optimal encoding strategy. This is a very general phenomenon that must occur if a noisy channel with a fixed number of unit is to communicate stimuli with variable bitrates. We also found that this strategy correlates with statistical criticality. These statistical signatures may be useful in identifying optimal model sizes in machine learning, and may provide clues about the operating regime of biological neural networks. Our findings might also relate to work that finds emergent critical statistics at an optimal layer depth in deep neural networks [36,37]. It remains to be seen whether the variability suppression observed here corresponds to the structure of the neural code *in vivo*.

4. Materials and Methods

4.1. Datasets

Images from the CIFAR-10 [38] data set were converted to gray scale, and binarized around the median pixel intensity. 90,000 randomly-selected circular patches of different radii were used as training data (Fig. 1a).

4.2. Restricted Boltzmann Machines

RBMs were fit using one-step contrastive divergence (CD1) [10,39] implemented in Theano (github.com/martinosorb/rbm_utils) on NVIDIA GeForce GTX 980 GPUs. The learning rate was reduced in stages: 0.2, 0.1, 0.05, 0.01, $5 \cdot 10^{-3}$, 10^{-3}. 8 epochs were trained at each rate with mini-batch size 4. To estimate model energies, 350,000 states were sampled via 500 chains of Gibbs sampling, keeping one sample every 150 steps.

4.3. Energy and entropy

In the RBM, hidden-layer entropy conditioned on stimulus v can be calculated in closed form as:

$$\mathrm{H}_{h|v} = \sum_{i=1..N_h} g(a^i_{h|v}) - a^i_{h|v} f(a^i_{h|v}),$$

where N_h is the number of hidden units, $a_{h|v}{=}v^\top W{+}B_h$ is the stimulus-conditioned hidden-layer activation vector, $f(x){=}1/(1{+}e^{-x})$ is the sigmoid function, and $g(x){=}\log(1{+}e^x)$. The expected conditional energy $\langle E_h \rangle_{h|v}$ is computed via sampling, where each individual E_h is computed, up to a constant, as:

$$\mathrm{E}_h = -B_h h - \sum_{i=1..N_v} g(W^i h + B^i_v) + \text{const.},$$

where N_v is the number of visible units, B_v is the vector of visible biases and W^i is the row of the weight matrix associated with the i^{th} visible unit. Energies are normalized using the energy of the lowest-energy (most frequent) pattern, estimated by sampling.

4.4. Fisher Information

The Fisher information matrix (FIM, Eq. 6) is a positive semidefinite matrix that defines the curvature of a metric on the manifold of parameters, and indicates the sensitivity of the model to parameter changes. Divergence of an eigenvalue of the FIM indicates an abrupt change in the model distribution, i.e. a phase transition. The FIM generalizes susceptibility and specific heat, physical quantities that diverge at critical points. For a vector $\vec{w}$ in parameter space, we define sensitivity as

$$S(\vec{w}) = \sqrt{\vec{w}^T F \vec{w}}.$$

The distribution of parameter sensitivity has in itself attracted interest [27,28]. For directions corresponding to eigenvectors of the Fisher information, the sensitivity is the square root of the corresponding eigenvalue. For changes in the k^{th} parameter, $S_k = \sqrt{F_{kk}}$. In the case of RBMs (Eq. 2), we can consider the definition of the FIM (Eq. 6) with the biases and weights being possible values of ϕ. Expanding the derivatives, one gets to FIM entries of the form

$$\begin{aligned}
F_{w_{ij},w_{kl}} &= \langle v_i h_j v_k h_l \rangle - \langle v_i h_j \rangle \langle v_k h_l \rangle & F_{b_i^v,b_k^h} &= \langle v_i h_k \rangle - \langle v_i \rangle \langle h_k \rangle \\
F_{w_{ij},b_k^v} &= \langle v_i h_j v_k \rangle - \langle v_i h_j \rangle \langle v_k \rangle & F_{b_i^v,b_k^v} &= \langle v_i v_k \rangle - \langle v_i \rangle \langle v_k \rangle \\
F_{w_{ij},b_k^h} &= \langle v_i h_j h_k \rangle - \langle v_i h_j \rangle \langle h_k \rangle & F_{b_i^h,b_k^h} &= \langle h_i h_k \rangle - \langle h_i \rangle \langle h_k \rangle,
\end{aligned}$$

where the brackets indicate averaging over the distribution $\Pr(v,h)$; these can be computed by sampling. The FIM diagonal summarizes the importance of individual units, and can be computed from locally-available variances and covariances:

$$F_{b_i^v,b_i^v} = \sigma_{v_i}^2, \quad F_{b_i^h,b_i^h} = \sigma_{h_i}^2, \quad F_{w_{ij},w_{ij}} = \langle v_i^2 h_j^2 \rangle - \langle v_i h_j \rangle^2.$$

Free energy in RBMs

We review the derivation of free energy in the context of RBMs [12]. Consider the problem of approximating a data distribution P_v with a model distribution Q_v^ϕ parameterized by ϕ. In a latent variable model, one identifies a distribution on latent factors Q_h^ϕ, as well as a mapping from latent factors to data patterns $Q_{v|h}^\phi$. The latent variables approximate the distribution over the data, i.e.

$$Q_v^\phi = \sum_h Q_{h,v}^\phi = \sum_h Q_{v|h}^\phi Q_h^\phi.$$

Such a model model can be optimized by minimizing the negative log-likelihood of data given model parameters:

$$\operatorname*{argmin}_\theta \left[-\sum_v P_v \log Q_v^\phi \right] = \operatorname*{argmin}_\theta \left[-\sum_v P_v \log \sum_h Q_{h,v}^\phi \right].$$

Jensen's inequality provides an upper bound on the negative log-likelihood that can be easier to minimize. This minimization is equivalent to minimizing the KL divergence from the model to the data distribution:

$$\begin{aligned}
-\sum_v P_v \log \sum_h Q_{h,v}^\phi &= -\sum_v P_v \log \sum_h Q_{h|v}^\phi \frac{Q_{h,v}^\phi}{Q_{h|v}^\phi} \\
&\leq \sum_v P_v \underbrace{\left[-\sum_h Q_{h|v}^\phi \log \frac{Q_{h,v}^\phi}{Q_{h|v}^\phi} \right]}_{E_v^\phi}.
\end{aligned}$$

This connects to the free-energy equation derived by Hinton *et al.* [12], which highlights the relationship between conditional distributions $Q^{\phi}_{h|v}$ and the visible pattern energies $E_v = -\log P_v$. When free energy is minimized over the data distribution, the model energies E^{ϕ}_v approximate the data energies and:

$$\begin{aligned} E^{\phi}_v &= -\sum_h Q^{\phi}_{h|v} \log \frac{Q^{\phi}_{h,v}}{Q^{\phi}_{h|v}} \\ &= \underbrace{-\sum_h Q^{\phi}_{h|v} \log Q^{\phi}_{h,v}}_{\left\langle E^{\phi}_{h,v} \right\rangle_{h|v}} + \underbrace{\sum_h Q^{\phi}_{h|v} \log Q^{\phi}_{h|v}}_{-H^{\phi}_{h|v}} \end{aligned}$$

This relation is derived by Hinton *et al.* [12], equation 5, from the perspective of minimizing communication cost, and in analogy to the Helmholtz free-energy from thermodynamics. This brief derivation illustrates the free-energy relationship in the context of minimizing an upper-bound on the negative log-likelihood of a latent-variable model.

Author Contributions: conceptualization, methodology, validation, and formal analysis M.E.R, M.S, M.H.H.; software, M.S, M.H.H.; writing–original draft preparation, review, and editing, M.E.R, M.S, M.H.H.; supervision, project administration, and funding acquisition M.H.H.;

Funding: Funding provided by the Engineering and Physical Sciences Research Council grant EP/L027208/1. M.S. was supported by the EuroSPIN Erasmus Mundus Program, the EPSRC Doctoral Training Centre in Neuroinformatics (EP/F500385/1 and BB/F529254/1), and a Google Doctoral Fellowship.

Acknowledgments: We are grateful to Dr. Timothy O'Leary for helpful comments on an early draft of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RBM	Restricted Boltzmann Machine
FIM	Fisher Information Matrix
v	"Visible stimulus" pattern, the input to a neural sensory "encoder"
h	"Hidden activation" pattern of stimulus-driven binary neural activity (interpreted as spiking)
W	The weight matrix for an RBM mapping visible activations to hidden-unit drive
B_h	The biases on the hidden units for an RBM
B_v	The biases on the visible units for an RBM
ϕ	Parameters $\{W, B_h, B_v\}$ associated with an RBM model
E	"Energy", defined here as negative log-probability
H	"Entropy", in the Shannon sense
$\mathrm{E}_{h,v}$	The log-probability of simultaneously observing stimulus v and neural pattern h
$H_{h\|v}$	The entropy of the distribution of neural patterns h evoked by stimulus v
$\mathcal{V}_E$	Set of input stimuli with similar energy (log-probability i.e. bitrate)
T_c	The critical temperature of an RBM interpreted as an Ising spin model
β	Inverse temperature

1. Barlow, H.B. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1972**, *1*, 371–394.
2. Shannon, C.E. A mathematical theory of communication, Part I, Part II. *Bell Syst. Tech. J.* **1948**, *27*, 623–656.
3. Field, D.J. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a* **1987**, *4*, 2379–2394.
4. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation* **1995**, *7*, 1129–1159.

5. Vinje, W.E.; Gallant, J.L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **2000**, *287*, 1273–1276.
6. Schneidman, E.; Berry, M.J.; Segev, R.; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012.
7. Shlens, J.; Field, G.D.; Gauthier, J.L.; Grivich, M.I.; Petrusca, D.; Sher, A.; Litke, A.M.; Chichilnisky, E. The structure of multi-neuron firing patterns in primate retina. *Journal of Neuroscience* **2006**, *26*, 8254–8266.
8. Köster, U.; Sohl-Dickstein, J.; Gray, C.M.; Olshausen, B.A. Modeling higher-order correlations within cortical microcolumns. *PLoS computational biology* **2014**, *10*.
9. Tkačik, G.; Mora, T.; Marre, O.; Amodei, D.; Palmer, S.E.; Berry, M.J.; Bialek, W. Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences* **2015**, *112*, 11508–11513.
10. Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural computation* **2002**, *14*, 1771–1800.
11. Hinton, G.E. A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade*; Springer, 2012; pp. 599–619.
12. Hinton, G.E.; Dayan, P.; Frey, B.J.; Neal, R.M. The "wake-sleep" algorithm for unsupervised neural networks. *Science* **1995**, *268*, 1158–61. doi:10.1126/science.7761831.
13. Dayan, P.; Hinton, G.E.; Neal, R.M.; Zemel, R.S. The Helmholtz Machine. *Neural Comput.* **1995**, *7*, 889–904. doi:10.1162/neco.1995.7.5.889.
14. Mora, T.; Bialek, W. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics* **2011**, *144*, 268–302. doi:10.1007/s10955-011-0229-4.
15. Sorbaro, M.; Herrmann, J.M.; Hennig, M. Statistical models of neural activity, criticality, and Zipf's law. In *The Functional Role of Critical Dynamics in Neural Systems*; Springer, 2019; pp. 265–287.
16. Bradde, S.; Bialek, W. Pca meets rg. *Journal of statistical physics* **2017**, *167*, 462–475.
17. Meshulam, L.; Gauthier, J.L.; Brody, C.D.; Tank, D.W.; Bialek, W. Coarse graining, fixed points, and scaling in a large population of neurons. *Physical review letters* **2019**, *123*, 178103.
18. Stringer, C.; Pachitariu, M.; Steinmetz, N.; Carandini, M.; Harris, K.D. High-dimensional geometry of population responses in visual cortex. *Nature* **2019**, *571*, 361–365.
19. Ioffe, M.L.; Berry II, M.J. The structured 'low temperature'phase of the retinal population code. *PLoS computational biology* **2017**, *13*, e1005792.
20. Aitchison, L.; Corradi, N.; Latham, P.E. Zipf's Law Arises Naturally When There Are Underlying, Unobserved Variables. *PLOS Comput. Biol.* **2016**, *12*, e1005110. doi:10.1371/journal.pcbi.1005110.
21. Stephens, G.J.; Mora, T.; Tkačik, G.; Bialek, W. Statistical Thermodynamics of Natural Images. *Phys. Rev. Lett.* **2013**, *110*, 018701. doi:10.1103/PhysRevLett.110.018701.
22. Nonnenmacher, M.; Behrens, C.; Berens, P.; Bethge, M.; Macke, J.H. Signatures of criticality arise from random subsampling in simple population models. *PLoS computational biology* **2017**, *13*, e1005718.
23. Swendsen, R.H.; Wang, J.S. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical review letters* **1987**, *58*, 86.
24. Bedard, C.; Kroeger, H.; Destexhe, A. Does the 1/f frequency scaling of brain signals reflect self-organized critical states? *Physical review letters* **2006**, *97*, 118102.
25. Beggs, J.M.; Timme, N. Being critical of criticality in the brain. *Frontiers in physiology* **2012**, *3*, 163.
26. Prokopenko, M.; Lizier, J.T.; Obst, O.; Wang, X.R. Relating Fisher information to order parameters. *Physical Review E* **2011**, *84*, 041116.
27. Daniels, B.C.; Chen, Y.J.; Sethna, J.P.; Gutenkunst, R.N.; Myers, C.R. Sloppiness, robustness, and evolvability in systems biology. *Current opinion in biotechnology* **2008**, *19*, 389–395.
28. Gutenkunst, R.N.; Waterfall, J.J.; Casey, F.P.; Brown, K.S.; Myers, C.R.; Sethna, J.P. Universally sloppy parameter sensitivities in systems biology models. *PLoS computational biology* **2007**, *3*, e189.
29. Panas, D.; Amin, H.; Maccione, A.; Muthmann, O.; van Rossum, M.; Berdondini, L.; Hennig, M.H. Sloppiness in spontaneously active neuronal networks. *Journal of Neuroscience* **2015**, *35*, 8480–8492.
30. Friston, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **2010**, *11*, 127–138.
31. LaMont, C.H.; Wiggins, P.A. Correspondence between thermodynamics and inference. *Physical Review E* **2019**, *99*, 052140.

32. Churchland, M.M.; Byron, M.Y.; Cunningham, J.P.; Sugrue, L.P.; Cohen, M.R.; Corrado, G.S.; Newsome, W.T.; Clark, A.M.; Hosseini, P.; Scott, B.B.; others. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience* **2010**, *13*, 369–378.
33. Schneidman, E.; Puchalla, J.L.; Segev, R.; Harris, R.A.; Bialek, W.; Berry, M.J. Synergy from silence in a combinatorial neural code. *Journal of Neuroscience* **2011**, *31*, 15732–15741.
34. Mastromatteo, I.; Marsili, M. On the criticality of inferred models. *Journal of Statistical Mechanics: Theory and Experiment* **2011**, p. 6.
35. Schwab, D.J.; Nemenman, I.; Mehta, P. Zipf's law and criticality in multivariate data without fine-tuning. *Phys. Rev. Lett.* **2014**, *113*, 1–5. doi:10.1103/PhysRevLett.113.068102.
36. Song, J.; Marsili, M.; Jo, J. Resolution and relevance trade-offs in deep learning. *Journal of Statistical Mechanics: Theory and Experiment* **2018**, *2018*, 123406.
37. Cubero, R.J.; Jo, J.; Marsili, M.; Roudi, Y.; Song, J. Statistical criticality arises in most informative representations. *Journal of Statistical Mechanics: Theory and Experiment* **2019**, *2019*, 063402.
38. Krizhevsky, A.; Hinton, G. Learning multiple layers of features from tiny images **2009**.
39. Bengio, Y.; others. Learning deep architectures for AI. *Foundations and trends in Machine Learning* **2009**, *2*, 1–127.