entropy

MDPI

Article

Minimum and Maximum Entropy Distributions for Binary Systems with Known Means and Pairwise Correlations

Badr F. Albanna [1,2,3], Christopher Hillar [3,4], Jascha Sohl-Dickstein [3,5,6,7] and Michael R. DeWeese [2,3,5,*]

1 Department of Natural Sciences, Fordham University, New York, NY 10023, USA; balbanna@fordham.edu
2 Department of Physics, University of California, Berkeley, CA 94720, USA
3 Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA 94720, USA; chillar@msri.edu (C.H.); jaschasd@google.com (J.S.-D.)
4 Mathematical Sciences Research Institute, Berkeley, CA 94720, USA
5 Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA
6 Biophysics Graduate Group, University of California, Berkeley, CA 94720, USA
7 Google Brain, Google, Mountain View, CA 94043, USA
* Correspondence: deweese@berkeley.edu

Received: 27 June 2017; Accepted: 18 August 2017; Published: 21 August 2017

Abstract: Maximum entropy models are increasingly being used to describe the collective activity of neural populations with measured mean neural activities and pairwise correlations, but the full space of probability distributions consistent with these constraints has not been explored. We provide upper and lower bounds on the entropy for the *minimum* entropy distribution over arbitrarily large collections of binary units with any fixed set of mean values and pairwise correlations. We also construct specific low-entropy distributions for several relevant cases. Surprisingly, the minimum entropy solution has entropy scaling logarithmically with system size for any set of first- and second-order statistics consistent with arbitrarily large systems. We further demonstrate that some sets of these low-order statistics can only be realized by small systems. Our results show how only small amounts of randomness are needed to mimic low-order statistical properties of highly entropic distributions, and we discuss some applications for engineered and biological information transmission systems.

Keywords: information theory; minimum entropy; maximum entropy; statistical mechanics; Ising model; pairwise correlations; compressed sensing; neural networks

1. Introduction

Maximum entropy models are central to the study of physical systems in thermal equilibrium [1], and they have recently been found to model protein folding [2,3], antibody diversity [4], and neural population activity [5–9] quite well (see [10] for a different finding). In part due to this success, these types of models have also been used to infer functional connectivity in complex neural circuits [11–13] and to model collective phenomena of systems of organisms, such as flock behavior [14].

This broad application of maximum entropy models is perhaps surprising since the usual physical arguments involving ergodicity or equality among energetically accessible states are not obviously applicable for such systems, though maximum entropy models have been justified in terms of imposing no structure beyond what is explicitly measured [5,15]. With this approach, the choice of measured constraints fully specifies the corresponding maximum entropy model. More generally, choosing a set of constraints restricts the set of consistent probability distributions—the maximum entropy solution being only one of all possible consistent probability distributions. If the space of distributions were

sufficiently constrained by observations so that only a small number of very similar models were consistent with the data, then agreement between the maximum entropy model and the data would be an unavoidable consequence of the constraints rather than a consequence of the unique suitability of the maximum entropy model for the dataset in question.

In the field of systems neuroscience, understanding the range of allowed entropies for given constraints is an area of active interest. For example, there has been controversy [5,16–19] over the notion that small pairwise correlations can conspire to constrain the behavior of large neural ensembles, which has led to speculation about the possibility that groups of $\sim$200 neurons or more might employ a form of error correction when representing sensory stimuli [5]. Recent work [20] has extended the number of simultaneously recorded neurons modeled using pairwise statistics up to 120 neurons, pressing closer to this predicted error correcting limit. This controversy is in part a reflection of the fact that pairwise models do not always allow accurate extrapolation from small populations to large ensembles [16,17], pointing to the need for exact solutions in the important case of large neural systems. Another recent paper [21] has also examined specific classes of biologically-plausible neural models whose entropy grows linearly with system size. Intriguingly, these authors point out that entropy can be subextensive, at least for one special distribution that is not well matched to most neural data (coincidentally, that special case was originally studied nearly 150 years ago [22] when entropy was a new concept, though not in the present context). Understanding the range of possible scaling properties of the entropy in a more general setting is of particular importance to neuroscience because of its interpretation as a measure of the amount of information communicable by a neural system to groups of downstream neurons.

Previous authors have studied the large scale behavior of these systems with maximum entropy models expanded to second- [5], third- [17], and fourth-order [19]. Here we use non-perturbative methods to derive rigorous upper and lower bounds on the entropy of the *minimum* entropy distribution for *any* fixed sets of means and pairwise correlations possible for arbitrarily large systems (Equations (1)–(9); Figures 1 and 2). We also derive lower bounds on the *maximum* entropy distribution (Equation (8)) and construct explicit low and high entropy models (Equations (28) and (29)) for a broad array of cases including the full range of possible uniform first- and second-order constraints realizable for large systems. Interestingly, we find that entropy differences between models with the same first- and second-order statistics can be nearly as large as is possible between any two arbitrary distributions over the same number of binary variables, provided that the solutions do not run up against the boundary of the space of allowed constraint values (see Section 2.1 and Figure 3 for a simple illustration of this phenomenon). This boundary is structured in such a way that some ranges of values for low order statistics are only satisfiable by systems below some critical size. Thus, for cases away from the boundary, entropy is only weakly constrained by these statistics, and the success of maximum entropy models in biology [2–9,11,14], when it occurs for large enough systems [17], can represent a real triumph of the maximum entropy approach.

Our results also have relevance for engineered information transmission systems. We show that empirically measured first-, second-, and even third-order statistics are essentially inconsequential for testing coding optimality in a broad class of such systems, whereas the existence of other statistical properties, such as finite exchangeability [23], do guarantee information transmission near channel capacity [24,25], the maximum possible information rate given the properties of the information channel. A better understanding of minimum entropy distributions subject to constraints is also important for minimal state space realization [26,27]—a form of optimal model selection based on an interpretation of Occam's Razor complementary to that of Jaynes [15]. Intuitively, maximum entropy models impose no structure beyond that needed to fit the measured properties of a system, whereas minimum entropy models require the fewest "moving parts" in order to fit the data. In addition, our results have implications for computer science as algorithms for generating binary random variables with constrained statistics and low entropy have found many applications (e.g., [28–43]).

Problem Setup

To make these ideas concrete, consider an abstract description of a neural ensemble consisting of N spiking neurons. In any given time bin, each neuron i has binary state s_i denoting whether it is currently firing an action potential ($s_i = 1$) or not ($s_i = 0$). The state of the full network is represented by $\vec{s} = (s_1, \ldots, s_N) \in \{0,1\}^N$. Let $p(\vec{s})$ be the probability of state $\vec{s}$ so that the distribution over all 2^N states of the system is represented by $\mathbf{p} \in [0,1]^{2^N}$, $\sum_{\vec{s}} p(\vec{s}) = 1$. Although we will use this neural framework throughout the paper, note that all of our results will hold for any type of system consisting of binary variables.

In neural studies using maximum entropy models, electrophysiologists typically measure the time-averaged firing rates $\mu_i = \langle s_i \rangle$ and pairwise event rates $\nu_{ij} = \langle s_i s_j \rangle$ and fit the maximum entropy model consistent with these constraints, yielding a Boltzmann distribution for an Ising spin glass [44]. This "inverse" problem of inferring the interaction and magnetic field terms in an Ising spin glass Hamiltonian that produce the measured means and correlations is nontrivial, but there has been progress [19,45–50]. The maximum entropy distribution is not the only one consistent with these observed statistics, however. In fact, there are typically many such distributions, and we will refer to the complete set of these as the *solution space* for a given set of constraints. Little is known about the minimum entropy permitted for a particular solution space.

Our question is: Given a set of observed mean firing rates and pairwise correlations between neurons, what are the possible entropies for the system? We will denote the maximum (minimum) entropy compatible with a given set of imposed correlations up to order n by S_n ($\tilde{S}_n$). The maximum entropy framework [5] provides a hierarchical representation of neural activity: as increasingly higher order correlations are measured, the corresponding model entropy S_n is reduced (or remains the same) until it reaches a lower limit. Here we introduce a complementary, minimum entropy framework: as higher order correlations are specified, the corresponding model entropy $\tilde{S}_n$ is increased (or unchanged) until all correlations are known. The range of possible entropies for any given set of constraints is the gap ($S_n - \tilde{S}_n$) between these two model entropies, and our primary concern is whether this gap is greatly reduced for a given set of observed first- or second-order statistics. We find that, for many cases, the gap grows linearly with the system size N, up to a logarithmic correction.

2. Results

We prove the following bounds on the minimum and maximum entropies for fixed sets of values of first and second order statistics, $\{\mu_i\} = \{\langle s_i \rangle\}$ and $\{\nu_{ij}\} = \{\langle s_i s_j \rangle\}$, respectively. All entropies are given in bits.

For the *minimum entropy*:

$$\log_2 \left(\frac{N}{1 + (N-1)\bar{\alpha}} \right) \leq \tilde{S}_2 \leq \log_2 \left(1 + \frac{N(N+1)}{2} \right), \tag{1}$$

where $\bar{\alpha}$ is the average of $\alpha_{ij} = (4\nu_{ij} - 2\mu_i - 2\mu_j + 1)^2$ over all $i, j \in \{1, \ldots, N\}$, $i \neq j$. Perhaps surprisingly, the scaling behavior of the minimum entropy does not depend on the details of the sets of constraint values—for large systems the entropy floor does not contain tall peaks or deep valleys as one varies $\{\mu_i\}$, $\{\nu_{ij}\}$, or N.

We emphasize that the bounds in Equation (1) are valid for arbitrary sets of mean firing rates and pairwise correlations, but we will often focus on the special class of distributions with *uniform constraints*:

$$\mu_i = \mu, \quad \text{for all } i = 1, \ldots, N \tag{2}$$

$$\nu_{ij} = \nu, \quad \text{for all } i \neq j. \tag{3}$$

The allowed values for μ and ν that can be achieved for arbitrarily large systems (see Appendix A) are

$$\mu^2 \leq \nu \leq \mu. \tag{4}$$

For uniform constraints, Equation (1) reduces to

$$\log_2\left(\frac{N}{1+(N-1)\alpha}\right) \leq \tilde{S}_2 \leq \log_2\left(1+\frac{N(N+1)}{2}\right), \tag{5}$$

where $\alpha(\mu,\nu) = (4(\nu-\mu)+1)^2$. In most cases, the lower bound in Equation (1) asymptotes to a constant, but in the special case for which μ and ν have values consistent with the global maximum entropy solution ($\mu = 1/2$ and $\nu = 1/4$), we can give the higher bound:

$$\log_2(N) \leq \tilde{S}_2 \leq \log_2(N) + 2. \tag{6}$$

For the *maximum entropy*, it is well known that:

$$S_2 \leq N, \tag{7}$$

which is valid for any set of values for $\{\mu_i\}$ and $\{\nu_{ij}\}$. Additionally, for any set of uniform constraints that can be achieved by arbitrarily large systems (Equation (4)) other than $\nu = \mu$, which corresponds to the case in which all N neurons are perfectly correlated, we can derive upper and lower bounds on the maximum entropy that each scale linearly with the system size:

$$w + xN \leq S_2 \leq N, \tag{8}$$

where $w = w(\mu,\nu) \geq 0$ and $x = x(\mu,\nu) > 0$ are independent of N.

An important class of probability distributions are the *exchangeable distributions* [23], which are distributions over multiple variables that are symmetric under any permutation of those variables. For binary systems, exchangeable distributions have the property that the probability of a sequence of ones and zeros is only a function of the number of ones in the binary string. We have constructed a family of exchangeable distributions, with entropy $\tilde{S}_2^{exch}$, that we conjecture to be minimum entropy exchangeable solutions. The entropy of our exchangeable constructions scale linearly with N:

$$C_1 N - \mathcal{O}(\log_2 N) \leq \tilde{S}_2^{exch} \leq C_2 N, \tag{9}$$

where $C_1 = C_1(\mu,\nu)$ and $C_2 = C_2(\mu,\nu)$ do not depend on N. We have computationally confirmed that this is indeed a minimum entropy exchangeable solution for $N \leq 200$.

Figure 1 illustrates the scaling behavior of these various bounds for uniform constraints in two parameter regimes of interest. Figure 2 shows how the entropy depends on the level of correlation (ν) for the maximum entropy solution (S_2), the minimum entropy exchangeable solution($\tilde{S}_2^{exch}$), and a low entropy solution ($\tilde{S}_2^{con}$), for a particular value of mean activity ($\mu = 1/2$) at each of two system sizes, $N = 5$ and $N = 30$.

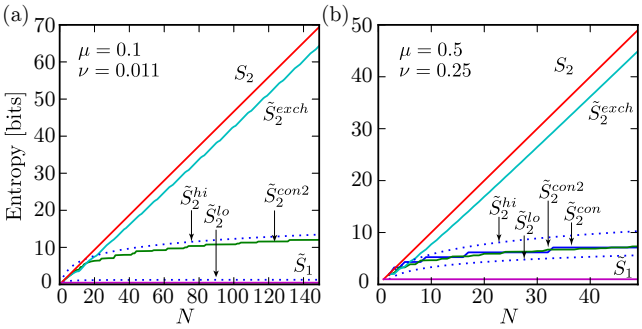

Figure 1. Minimum and maximum entropy for fixed uniform constraints as a function of N. The minimum entropy grows no faster than logarithmically with the system size N for any mean activity level μ and pairwise correlation strength ν. (**a**) In a parameter regime relevant for neural population activity in the retina [5,6] ($\mu = 0.1$, $\nu = 0.011$), we can construct an explicit low entropy solution ($\tilde{S}_2^{con2}$) that grows logarithmically with N, unlike the linear behavior of the maximum entropy solution (S_2). Note that the linear behavior of the maximum entropy solution is only possible because these parameter values remain within the boundary of allowed μ and ν values (See Appendix C); (**b**) Even for mean activities and pairwise correlations matched to the global maximum entropy solution (S_2; $\mu = 1/2$, $\nu = 1/4$), we can construct explicit low entropy solutions ($\tilde{S}_2^{con}$ and $\tilde{S}_2^{con2}$) and a lower bound ($\tilde{S}_2^{lo}$) on the entropy that each grow logarithmically with N, in contrast to the linear behavior of the maximum entropy solution (S_2) and the finitely exchangeable minimum entropy solution ($\tilde{S}_2^{exch}$). $\tilde{S}_1$ is the minimum entropy distribution that is consistent with the mean firing rates. It remains constant as a function of N.

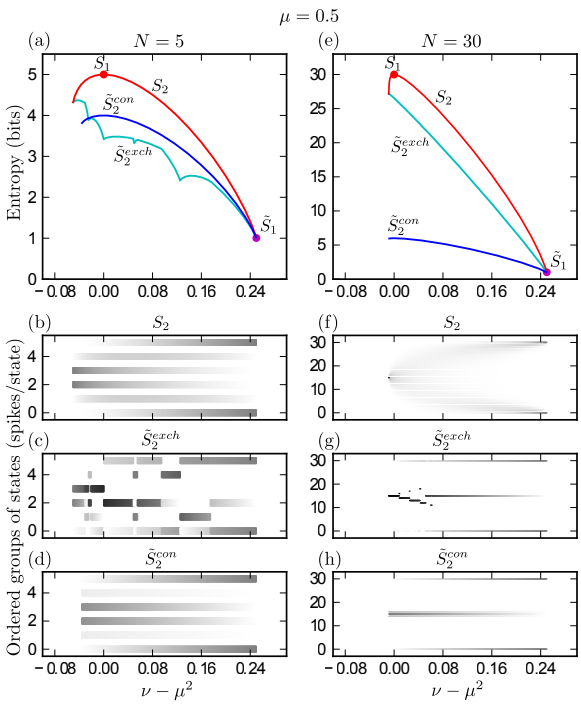

Figure 2. Minimum and maximum entropy models for uniform constraints. (**a**) Entropy as a function of the strength of pairwise correlations for the maximum entropy model (S_2), finitely exchangeable minimum entropy model ($\tilde{S}_2^{exch}$), and a constructed low entropy solution ($\tilde{S}_2^{con}$), all corresponding to $\mu = 1/2$ and $N = 5$. Filled circles indicate the global minimum $\tilde{S}_1$ and maximum S_1 for $\mu = 1/2$; (**b**–**d**) Support for S_2 (b), $\tilde{S}_2^{exch}$ (c), and $\tilde{S}_2^{con}$ (d) corresponding to the three curves in panel (a). States are grouped by the number of active units; darker regions indicate higher total probability for each group of states; (**e**–**h**) Same as for panels (a) through (d), but with $N = 30$. Note that, with rising N, the cusps in the $\tilde{S}_2^{exch}$ curve become much less pronounced.

2.1. Limits on System Growth

Physicists are often faced with the problem of having to determine some set of experimental predictions (e.g., mean values and pairwise correlations of spins in a magnet) for some model system defined by a given Hamiltonian, which specifies the energies associated with any state of the system. Typically, one is interested in the limiting behavior as the system size tends to infinity. In this canonical situation, no matter how the various terms in the Hamiltonian are defined as the system grows, the existence of a well-defined Hamiltonian guarantees that there exists a (maximum entropy) solution for the probability distribution for any system size.

However, here we are studying the *inverse* problem of deducing the underlying model based on measured low-order statistics [19,45–49]. In particular, we are interested in the minimum and maximum entropy models consistent with a given set of specified means and pairwise correlations. Clearly, both of these types of (potentially degenerate) models must exist whenever there exists at least one distribution consistent with the specified statistics, but as we now show, some sets of constraints can only be realized for small systems.

A Simple Example

To illustrate this point, consider the following example, which is arguably the simplest case exhibiting a cap on system size. At least one system consisting of N neurons can be constructed to satisfy the uniform set of constraints: $\mu = 0.1$ and $\nu = 0.0094 < \mu^2 = 0.01$, provided that $2 \leq N \leq 150$, but no solution is possible for $N > 150$. To prove this, we first observe that any set of uniform constraints that admits at least one solution must admit at least one exchangeable solution (see Appendices A and D). Armed with this fact, we can derive closed-form solutions for upper and lower bounds on the minimum value for ν consistent with N and μ (Appendix Equation (A25)), as well as the actual minimum value of ν for any given value of N, as depicted in Figure 3.

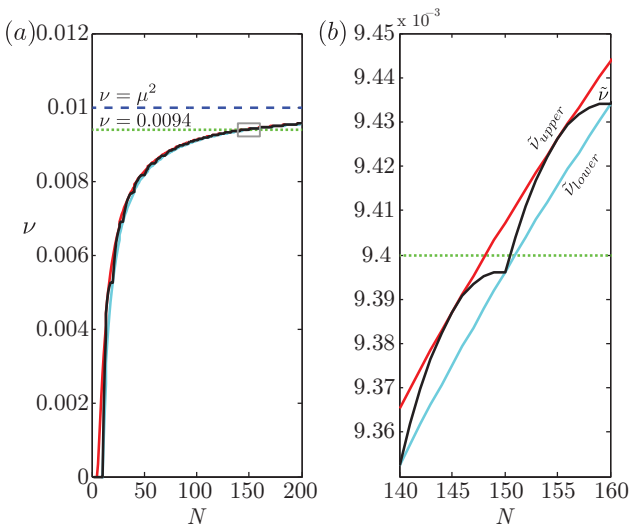

Figure 3. An example of uniform, low-level statistics that can be realized by small groups of neurons but not by any system greater than some critical size. (**a**) Upper (red curve, $\tilde{\nu}_{upper}$) and lower (cyan curve, $\tilde{\nu}_{lower}$) bounds on the minimum value (black curve, $\tilde{\nu}$) for the pairwise correlation ν shared by all pairs of neurons are plotted as a function of system size N assuming that every neuron has mean activity $\mu = 0.1$. Note that all three curves asymptote to $\nu = \mu^2 = 0.01$ as $N \to \infty$ (dashed blue line); (**b**) Enlarged portion of (**a**) outlined in grey reveals that groups of $N \leq 150$ neurons can exhibit uniform constraints $\mu = 0.1$ and $\nu = 0.0094$ (green dotted line), but this is not possible for any larger group.

Thus, this system defined by its low-level statistics cannot be grown indefinitely, not because the entropy decreases beyond some system size, which is impossible (e.g., Theorem 2.2.1 of [25]),

but rather, because no solution of any kind is possible for these particular constraints for ensembles of more than 150 neurons. We believe that this relatively straightforward example captures the essence of the more subtle phenomenon described by Schneidman and colleagues [5], who pointed out that a naive extrapolation of the maximum entropy model fit to their retinal data to larger system sizes results in a conundrum at around $N \approx 200$ neurons.

For our simple example, it was straightforward to decide what rules to follow when growing the system—both the means and pairwise correlations were fixed and perfectly uniform across the network at every stage. In practice, real neural activities and most other types of data exhibit some variation in their mean activities and correlations across the measured population, so in order to extrapolate to larger systems, one must decide how to model the distribution of statistical values involving the added neurons.

Fortunately, despite this complication, we have been able to derive upper and lower bounds on the minimum entropy for arbitrarily large N. In Appendix C, we derive a lower bound on the maximum entropy for the special case of uniform constraints achievable for arbitrarily large systems, but such a bound for the more general case would depend on the details of how the statistics of the system change with N.

Finally, we mention that this toy model can be thought of as an example of a "frustrated" system [51], in that traveling in a closed loop of an odd number of neurons involves an odd number of anticorrelated pairs. By "anticorrelated", we mean that the probability of simultaneous firing of a pair of neurons is less than chance, $\nu < \mu^2$, but note that ν is always non-negative due to our convention of labeling active and inactive units with zeros and ones, respectively. However, frustrated systems are not typically defined by their correlational structure, but rather by their Hamiltonians, and there are often many Hamiltonians that are consistent with a given set of observed sets of constraints, so there does not seem to be a one-to-one relationship between these two notions of frustration. In the more general setting of nonuniform constraints, the relationship between frustrated systems as defined by their correlational structure and those that cannot be grown arbitrarily large is much more complex.

2.2. Bounds on Minimum Entropy

Entropy is a strictly concave function of the probabilities and therefore has a unique maximum that can be identified using standard methods [52], at least for systems that possess the right symmetries or that are sufficiently small. In Section 2.3, we will show that the maximum entropy S_2 for many systems with specified means and pairwise correlations scales linearly with N (Equation (8), Figure 1). We obtain bounds on minimum entropy by exploiting the geometry of the entropy function.

2.2.1. Upper Bound on the Minimum Entropy

We prove below that the minimum entropy distribution exists at a vertex of the allowed space of probabilities, where most states have probability zero [53]. Our challenge then is to determine in which vertex a minimum resides. The entropy function is nonlinear, precluding obvious approaches from linear programming, and the dimensionality of the probability space grows exponentially with N, making exhaustive search and gradient descent techniques intractable for $N \gtrsim 5$. Fortunately, we can compute a lower (upper) bound $\tilde{S}_2^{lo}$ ($\tilde{S}_2^{hi}$) on the entropy of the minimum entropy solution for all N (Figure 1), and we have constructed two families of explicit solutions with low entropies ($\tilde{S}_2^{con}$ and $\tilde{S}_2^{con2}$; Figures 1 and 2) for a broad parameter regime covering all allowed values for μ and ν in the case of uniform constraints that can be achieved by arbitrarily large systems (see Equation (A24), Figure A2 in the Appendix).

Our goal is to minimize the entropy S as a function of the probabilities p_i where

$$S(\mathbf{p}) = -\sum_{i=1}^{n_s} p_i \log_2 p_i, \tag{10}$$

n_s is the number of states, the p_i satisfy a set of n_c independent linear constraints, and $p_i \geq 0$ for all i. For the main problem we consider, $n_s = 2^N$. The number of constraints—normalization, mean firing rates, and pairwise correlations—grows quadratically with N:

$$n_c = 1 + \frac{N(N+1)}{2}. \tag{11}$$

The space of normalized probability distributions $\mathcal{P} = \{p : \sum_{i=1}^{n_s} p_i = 1,\ p_i \geq 0\}$ is the standard simplex in $n_s - 1$ dimensions. Each additional linear constraint on the probabilities introduces a hyperplane in this space. If the constraints are consistent and independent, then the intersection of these hyperplanes defines a $d = n_s - n_c$ affine space, which we call $\mathcal{C}$. All solutions are constrained to the intersection between $\mathcal{P}$ and $\mathcal{C}$, and this solution space is a convex polytope of dimension $\leq d$, which we refer to as $\mathcal{R}$. A point within a convex polytope can always be expressed as a linear combination of its vertices; therefore, if $\{\mathbf{v}_i\}$ are the vertices of $\mathcal{R}$, we may express

$$\mathbf{p} = \sum_i^{n_v} a_i \mathbf{v}_i, \tag{12}$$

where n_v is the total number of vertices and $\sum_i^{n_v} a_i = 1$.

Using the concavity of the entropy function, we will now show that the minimum entropy for a space of probabilities S is attained on one (or possibly more) of the vertices of that space. Moreover, these vertices correspond to probability distributions with small support—specifically a support size no greater than n_c (see Appendix B). This means that the global minimum will occur at the (possibly degenerate) vertex that has the lowest entropy, which we denote as $\mathbf{v}_*$:

$$\begin{aligned} S(\mathbf{p}) &= S\left(\sum_i^{n_v} a_i \mathbf{v}_i\right) \\ &\geq \sum_i^{v_s} a_i S(\mathbf{v}_i) \\ &\geq \left(\sum_i^{v_s} a_i\right) S(\mathbf{v}_*) \\ &= S(\mathbf{v}_*). \end{aligned} \tag{13}$$

It follows that $\tilde{S}_2 = S(\mathbf{v}_*)$.

Moreover, if a distribution satisfying the constraints exists, then there is one with at most n_c nonzero p_i (e.g., from arguments as in [41]). Together, these two facts imply that there are minimum entropy distributions with a maximum of n_c nonzero p_i. This means that even though the state space may grow exponentially with N, the support of the minimum entropy solution for fixed means and pairwise correlations will only scale quadratically with N.

The maximum entropy possible for a given support size occurs when the probability distribution is evenly distributed across the support and the entropy is equal to the logarithm of the number of states. This allows us to give an upper bound on the minimum entropy as

$$\begin{aligned} \tilde{S}_2 \leq \tilde{S}_2^{hi} &= \log_2(n_c) \\ &= \log_2\left(1 + \frac{N(N+1)}{2}\right) \\ &\approx 2\log_2(N), \quad N \gg 1. \end{aligned} \tag{14}$$

Note that this bound is quite general: as long as the constraints are independent and consistent this result holds regardless of the specific values of the $\{\mu_i\}$ and $\{\nu_{ij}\}$.

2.2.2. Lower Bound on the Minimum Entropy

We can also use the concavity of the entropy function to derive a lower bound on the entropy $\tilde{S}_2^{lo}$ as in Equation (1):

$$\tilde{S}_2(N, \{\mu_i\}, \{\nu_{ij}\}) \geq \tilde{S}_2^{lo} = \log_2 \left(\frac{N^2}{N + \sum_{i \neq j} \alpha_{ij}} \right), \tag{15}$$

where $\tilde{S}_2(N, \{\mu_i\}, \{\nu_{ij}\})$ is the minimum entropy given a network of size N with constraint values $\{\mu_i\}$ and $\{\nu_{ij}\}$, and the sum is taken over all $i, j \in \{1, \ldots, N\}$, $i \neq j$ and $\alpha_{ij} = (4\nu_{ij} - 2\mu_i - 2\mu_j + 1)^2$. Taken together with the upper bound, this fully characterizes the scaling behavior of the entropy floor as a function of N.

To derive this lower bound, note that the concavity of the entropy function allows us to write

$$S(\mathbf{p}) \geq -\log_2 \|\mathbf{p}\|_2^2. \tag{16}$$

Using this relation to find a lower bound on $S(\mathbf{p})$ requires an upper bound on $\|\mathbf{p}\|_2^2$ provided by the Frobenius norm of the correlation matrix $C \equiv \langle \vec{s}\vec{s}^T \rangle$ (where the states are defined to take vaues in $\{-1, 1\}^N$ rather than $\{0, 1\}^N$)

$$\|\mathbf{p}\|_2^2 \leq \frac{\|C\|_F^2}{N^2}. \tag{17}$$

In this case, $\|C\|_F^2$ is a simple function of the $\{\mu_i\}$, $\{\nu_{ij}\}$:

$$\|C\|_F^2 = N + \sum_{i \neq j} (4\nu_{ij} - 2\mu_i - 2\mu_j + 1)^2. \tag{18}$$

Using Equations (17) and (18) in Equation (16) gives us our result (see Appendix H for further details).

Typically, this bound asymptotes to a constant as the number of terms included in the sum over α_{ij} scales with the number of pairs ($\mathcal{O}(N^2)$), but in certain conditions this bound can grow with the system size. For example, in the case of uniform constraints, this reduces to

$$\tilde{S}_2^{lo} = \log_2 \left(\frac{N}{1 + (N-1)\alpha} \right), \tag{19}$$

where $\alpha(\mu, \nu) = (4(\nu - \mu) + 1)^2$.

In the special case

$$\nu = \mu - 1/4, \tag{20}$$

α vanishes allowing the bound in Equation (15) to scale logarithmically with N.

In the large N limit for uniform constraints, we know $\mu \geq \nu \geq \mu^2$ (see Appendix A), therefore the only values of μ and ν satisfying Equation (20) are

$$\mu = 1/2, \quad \nu = \mu^2 = 1/4. \tag{21}$$

Although here the lower bound grows logarithmically with N, rather than remaining constant, for many large systems this difference is insignificant compared with the linear dependence $S_0 = N$ of the maximum entropy solution (i.e., N fair i.i.d. Bernoulli random variables). In other words, the gap between the minimum and maximum possible entropies consistent with the measured mean activities and pairwise correlations grows linearly in these cases (up to a logarithmic correction), which is as large as the gap for the space of all possible distributions for binary systems of the same size N with the same mean activities but without restriction on the correlations.

2.3. Bounds on Maximum Entropy

An upper bound on the maximum entropy is well-known and easy to state. For a given number of allowed states n_s, the maximum possible entropy occurs when the probability distribution is equally distributed over all allowed states (i.e., the microcanonical ensemble), and the entropy is equal to

$$\begin{aligned} S_2 \leq S_2^{hi} &= \log_2(n_s) \\ &= N. \end{aligned} \tag{22}$$

Lower bounds on the maximum entropy are more difficult to obtain (see Appendix C for further discussion on this point.). However, by restricting ourselves to uniform constraints achievable by arbitrarily large systems (Equations (2) and (3)), we can construct a distribution that provides a useful lower bound on the maximum possible entropy consistent with these constraints:

$$S_2 \geq S_2^{con} = w + xN, \tag{23}$$

where

$$w = -\beta \log_2 \beta - (1-\beta)\log_2(1-\beta), \tag{24}$$

$$x = (1-\beta)\left[-\eta \log_2 \eta - (1-\eta)\log_2(1-\eta)\right] \tag{25}$$

and

$$\beta = \frac{\nu - \mu^2}{1 + \nu - 2\mu}, \tag{26}$$

$$\eta = \frac{\mu - \nu}{1 - \mu}. \tag{27}$$

It is straightforward to verify that w and x are nonnegative constants for all allowed values of μ and ν that can be achieved for arbitrarily large systems. Importantly, x is nonzero provided $\nu < \mu$ (i.e., $\beta \neq 1 \cap \eta \neq \{0,1\}$), so the entropy of the system will grow linearly with N for any set of uniform constraints achievable for arbitrarily large systems except for the case in which all neurons are perfectly correlated.

2.4. Low-Entropy Solutions

In addition to the bounds we derived for the minimum entropy, we can also construct probability distributions between these minimum entropy boundaries for distributions with uniform constraints. These solutions provide concrete examples of distributions that achieve the scaling behavior of the bounds we have derived for the minimum entropy. We include these so that the reader may gain a better intuition for what low entropy models look like in practice. We remark that these models are not intended as an improvement over maximum entropy models for any particular biological system. This would be an interesting direction for future work. Nonetheless, they are of practical importance to other fields as we discuss further in Section 2.6.

Each of our low entropy constructions, $\tilde{S}_2^{con}$ and $\tilde{S}_2^{con2}$, has an entropy that grows logarithmically with N (see Appendices E–G, Equations (1)–(6)):

$$\begin{aligned} \tilde{S}_2^{con} &= \lceil \log_2(N) + 1 \rceil \\ &\leq \log_2(N) + 2, \end{aligned} \tag{28}$$

$$\begin{aligned} \tilde{S}_2^{con2} &\leq \log_2\left(\lceil N \rceil_p \left(\lceil N \rceil_p - 1\right)\right) - 1 + \log_2(3) \\ &\leq \log_2\left(N(2N-1)\right) + \log_2(3), \end{aligned} \tag{29}$$

where $\lceil . \rceil$ is the ceiling function and $\lceil . \rceil_p$ represents the smallest prime at least as large as its argument. Thus, there is always a solution whose entropy grows no faster than logarithmically with the size of the system, for any observed levels of mean activity and pairwise correlation.

As illustrated in Figure 1a, for large binary systems with uniform first- and second-order statistics matched to typical values of many neural populations, which have low firing rates and correlations slightly above chance ([5–9,11,16]; $\mu = 0.1$, $\nu = 0.011$), the range of possible entropies grows almost linearly with N, despite the highly symmetric constraints imposed (Equations (2) and (3)).

Consider the special case of first- and second-order constraints (Equation (21)) that correspond to the unconstrained global maximum entropy distribution. For these highly symmetric constraints, both our upper and lower bounds on the minimum entropy grow logarithmically with N, rather than just the upper bound as we found for the neural regime (Equation (6); Figure 1a). In fact, one can construct [21,22] an explicit solution (Equation (28); Figures 1b and 2a,d,e,h) that matches the mean, pairwise correlations, and triplet-wise correlations of the global maximum entropy solution whose entropy $\tilde{S}_2^{con}$ is never more than two bits above our lower bound (Equation (15)) for all N. Clearly then, these constraints alone do not guarantee a level of independence of the neural activities commensurate with the maximum entropy distribution. By varying the relative probabilities of states in this explicit construction we can make it satisfy a much wider range of μ and ν values than previously considered, covering most of the allowed region (see Appendices F and G) while still remaining a distribution whose entropy grows logarithmically with N.

2.5. Minimum Entropy for Exchangeable Distributions

We consider the exchangeable class of distributions as an example of distributions whose entropy must scale linearly with the size of the system unlike the global entropy minimum which we have shown scales logarithmically. If one has a principled reason to believe some system should be described by an exchangeable distribution, the constraints themselves are sufficient to drastically narrow the allowed range of entropies although the gap between the exchangeable minimum and the maximum will still scale linearly with the size of the system except in special cases. This result is perhaps unsurprising as the restriction to exchangeable distributions is equivalent to imposing a set of additional constraints (e.g., $p(100) = p(010) = p(001)$, for $N = 3$) that is exponential in the size of the system.

While a direct computational solution to the general problem of finding the minimum entropy solution becomes intractable for $N \gtrsim 5$, the situation for the exchangeable case is considerably different. In this case, the high level of symmetry imposed means that there are only $n_s = N + 1$ states (one for each number of active neurons) and $n_c = 3$ constraints (one for normalization, mean, and pairwise firing). This makes the problem of searching for the minimum entropy solution at each vertex of the space computationally tractable up into the hundreds of neurons.

Whereas the global lower bound scales logarithmically, our computation illustrates that the exchangeable case scales with N as seen in Figure 1. The large gap between $\tilde{S}_2^{exch}$ and $\tilde{S}_2$ demonstrates that a distribution can dramatically reduce its entropy if it is allowed to violate sufficiently strong symmetries present in the constraints. This is reminiscent of other examples of symmetry-breaking in physics for which a system finds an equilibrium that breaks symmetries present in the physical laws. However, here the situation can be seen as reversed: Observed first and second order statistics satisfy a symmetry that is not present in the underlying model.

2.6. Implications for Communication and Computer Science

These results are not only important for understanding the validity of maximum entropy models in neuroscience, but they also have consequences in other fields that rely on information entropy. We now examine consequences of our results for engineered communication systems. Specifically, consider a device such as a digital camera that exploits compressed sensing [54,55] to reduce the dimensionality of its image representations. A compressed sensing scheme might involve taking inner

products between the vector of raw pixel values and a set of random vectors, followed by a digitizing step to output N-bit strings. Theorems exist for expected information rates of compressed sensing systems, but we are unaware of any that do not depend on some knowledge about the input signal, such as its sparse structure [56,57]. Without such knowledge, it would be desirable to know which empirically measured output statistics could determine whether such a camera is utilizing as much of the N bits of channel capacity as possible for each photograph.

As we have shown, even if the mean of each bit is $\mu = 1/2$, and the second- and third-order correlations are at chance level ($\nu = 1/4$; $\langle s_i s_j s_k \rangle = 1/8$, for all sets of distinct $\{i, j, k\}$), consistent with the maximum entropy distribution, it is possible that the Shannon mutual information shared by the original pixel values and the compressed signal is only on the order of $\log_2(N)$ bits, well below the channel capacity (N bits) of this (noiseless) output stream. We emphasize that, in such a system, the transmitted information is limited not by corruption due to noise, which can be neglected for many applications involving digital electronic devices, but instead by the nature of the second- and higher-order correlations in the output.

Thus, measuring pairwise or even triplet-wise correlations between all bit pairs and triplets is insufficient to provide a useful floor on the information rate, no matter what values are empirically observed. However, knowing the extent to which other statistical properties are obeyed can yield strong guarantees of system performance. In particular, exchangeability is one such constraint. Figure 1 illustrates the near linear behavior of the lower bound on information ($\tilde{S}_2^{exch}$) for distributions obeying exchangeability, in both the neural regime (cyan curve, panel (a)) and the regime relevant for our engineering example (cyan curve, panel (b)). We find experimentally that any exchangeable distribution has as much entropy as the maximum entropy solution, up to terms of order $\log_2(N)$ (see Appendicies).

This result has potential applications in the field of symbolic dynamics and computational mechanics, which study the consequences of viewing a complex system through a finite state measuring device [27,58]. If we view each of the various models presented here as a time series of binary measurements from a system, our results indicate that bitstreams with identical mean and pairwise statistics can have profoundly different scaling as a function of the number of measurements (N), indicating radically different complexity. It would be interesting to explore whether the models presented here appear differently when viewed through the ϵ-machine framework [27].

In computer science, it is sometimes possible to construct efficient deterministic algorithms from randomized ones by utilizing low entropy distributions. One common technique is to replace the independent binary random variables used in a randomized algorithm with those satisfying only pairwise independence [59]. In many cases, such a randomized algorithm can be shown to succeed even if the original independent random bits are replaced by pairwise independent ones having significantly less entropy. In particular, efficient derandomization can be accomplished in these instances by finding pairwise independent distributions with small sample spaces. Several such designs are known and use tools from finite fields and linear codes [33,34,60–62], combinatorial block designs [32,63], Hadamard matrix theory [42,64], and linear programming [41], among others [22]. Our construction here of two families of low entropy distributions fit to specified mean activities and pairwise statistics adds to this literature.

3. Discussion

Ideas and approaches from statistical mechanics are finding many applications in systems neuroscience [65,66]. In particular, maximum entropy models are powerful tools for understanding physical systems, and they are proving to be useful for describing biology as well, but a deeper understanding of the full solution space is needed as we explore systems less amenable to arguments involving ergodicity or equally accessible states. Here we have shown that second order statistics do not significantly constrain the range of allowed entropies, though other constraints, such as exchangeability, do guarantee extensive entropy (i.e., entropy proportional to system size N).

We have shown that in order for the constraints themselves to impose a linear scaling on the entropy, the number of experimentally measured quantities that provide those constraints must scale exponentially with the size of the system. In neuroscience, this is an unlikely scenario, suggesting that whatever means we use to infer probability distributions from the data (whether maximum entropy or otherwise) will most likely have to agree with other, more direct, estimates of the entropy [67–73]. The fact that maximum entropy models chosen to fit a somewhat arbitrary selection of measured statistics are able to match the entropy of the system they model lends credence to the merits of this approach.

Neural systems typically exhibit a range of values for the correlations between pairs of neurons, with some firing coincidently more often than chance and others firing together less often than chance. Such systems can exhibit a form of frustration, such that they apparently cannot be scaled up to arbitrarily large sizes in such a way that the distribution of correlations and mean firing rates for the added neurons resembles that of the original system. We have presented a particularly simple example of a small system with uniform pairwise correlations and mean firing rates that cannot be grown beyond a specific finite size while maintaining these statistics throughout the network.

We have also indicated how, in some settings, minimum entropy models can provide a floor on information transmission, complementary to channel capacity, which provides a ceiling on system performance. Moreover, we show how highly entropic processes can be mimicked by low entropy processes with matching low-order statistics, which has applications in computer science.

Acknowledgments: The authors would like to thank Jonathan Landy, Tony Bell, Michael Berry, Bill Bialek, Amir Khosrowshahi, Peter Latham, Lionel Levine, Fritz Sommer, and all members of the Redwood Center for many useful discussions. M.R.D. is grateful for support from the Hellman Foundation, the McDonnell Foundation, the McKnight Foundation, the Mary Elizabeth Rennie Endowment for Epilepsy Research, and the National Science Foundation through Grant No. IIS-1219199. C.H. was supported under an NSF All-Institutes Postdoctoral Fellowship administered by the Mathematical Sciences Research Institute through its core grant DMS-0441170. This material is based upon work supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract number W911NF-13-1-0390.

Author Contributions: Badr F. Albanna and Michael R. DeWeese contributed equally to this project and are resposible for the bulk of the results and writing of the paper. Jascha Sohl-Dickstein contributed substantially to clarifications of the main results and improvements to the text including the connection of these results to computer science. Christopher Hillar also contributed substantially to clarifications of the main results, the method used to identify an analytical lower bound on the minimum entropy solution, and improvements to the clarity of the text. All authors approved of this manuscript before submission.

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

Appendix A. Allowed Range of ν Given μ Across All Distributions for Large N

In this section we will only consider distributions satisfying uniform constraints

$$\mu_i = \mu, \quad \forall i = 1, \ldots, N \tag{A1}$$

$$\nu_{ij} = \nu, \quad \forall i \neq j, \tag{A2}$$

and we will show that

$$\mu^2 \leq \nu \leq \mu \tag{A3}$$

in the large N limit. One could concievably extend the linear programming methods below to find bounds in the case of general non-uniform constraints, but as of this time we have not been able to do so without resorting to numerical algorithms on a case-by-case basis.

We begin by determining the upper bound on ν, the probability of any pair of neurons being simultaneously active, given μ, the probability of any one neuron being active, in the large N regime, where N is the total number of neurons. Time is discretized and we assume any neuron can spike no more than once in a time bin. We have $\nu \leq \mu$ because ν is the probability of a pair of neurons firing together and thus each neuron in that pair must have at least a firing probability of ν. Furthermore, it is easy to see that the case $\mu = \nu$ is feasible when there are only two states with non-zero probabilities:

all neurons silent (p_0) or all neurons active (p_1). In this case, $p_1 = \mu = \nu$. We use the term "active" to refer to neurons that are spiking, and thus equal to one, in a given time bin, and we also refer to "active" states in a distribution, which are those with non-zero probabilities.

We now proceed to show that the lower bound on ν in the large N limit is μ^2, the value of ν consistent with statistical independence among all N neurons. We can find the lower bound by viewing this as a linear programming problem [52,74], where the goal is to maximize $-\nu$ given the normalization constraint and the constraints on μ.

It will be useful to introduce the notion of an *exchangeable distribution* [23], for which any permutation of the neurons in the binary words labeling the states leaves the probability of each state unaffected. For example if $N = 3$, an exchangeable solution satisfies

$$p(100) = p(010) = p(001),$$
$$p(110) = p(101) = p(011).$$

In other words, the probability of any given word depends only on the number of ones it contains, not their particular locations, for an exchangeable distribution.

In order to find the allowed values of μ and ν, we need only consider exchangeable distributions. If there exists a probability distribution that satisfies our constraints, we can always construct an exchangeable one that also does given that the constraints themselves are symmetric (Equations (1) and (2)). Let us do this explicitly: Suppose we have a probability distribution $p(\vec{s})$ over binary words $\vec{s} = (s_1, \ldots, s_N) \in \{0,1\}^N$ that satisfies our constraints but is not exchangeable. We construct an exchangeable distribution $p_e(w)$ with the same constraints as follows:

$$p_e(\vec{s}) \equiv \sum_{\sigma} \frac{p(\sigma(\vec{s}))}{N!}, \tag{A4}$$

where σ is an element of the permutation group $\mathcal{P}_N$ on N elements. This distribution is exchangeable by construction, and it is easy to verify that it satisfies the same uniform constraints as does the original distribution, $p(\vec{s})$.

Therefore, if we wish to find the maximum $-\nu$ for a given value of μ, it is sufficient to consider exchangeable distributions. From now on in this section we will drop the e subscript on our earlier notation, define p to be exchangeable, and let $p(i)$ be the probability of a state with i spikes.

The normalization constraint is

$$1 = \sum_{i=0}^{N} \binom{N}{i} p(i). \tag{A5}$$

Here the binomial coefficient $\binom{N}{i}$ counts the number of states with i active neurons.

The firing rate constraint is similar, only now we must consider summing only those probabilities that have a particular neuron active. How many states are there with only a pair of active neurons given that a particular neuron must be active in all of the states? We have the freedom to place the remaining active neuron in any of the $N-1$ remaining sites, which gives us $\binom{N-1}{1}$ states with probability $p(2)$. In general if we consider states with i active neurons, we will have the freedom to place $i-1$ of them in $N-1$ sites, yielding:

$$\mu = \sum_{i=1}^{N} \binom{N-1}{i-1} p(i). \tag{A6}$$

Finally, for the pairwise firing rate, we must add up states containing a specific pair of active neurons, but the remaining $i-2$ active neurons can be anywhere else:

$$\nu = \sum_{i=2}^{N} \binom{N-2}{i-2} p(i). \tag{A7}$$

Now our task can be formalized as finding the maximum value of

$$-\nu = -\sum_{i=2}^{N} \binom{N-2}{i-2} p(i) \tag{A8}$$

subject to

$$1 = \sum_{i=0}^{N} \binom{N}{i} p(i), \tag{A9}$$

$$\mu = \sum_{i=1}^{N} \binom{N-1}{i-1} p(i), \tag{A10}$$

$$p(i) \geq 0, \quad \text{for all } i.$$

This gives us the following dual problem: Minimize

$$\mathcal{E} \equiv \lambda_0 + \mu\lambda_1, \tag{A11}$$

given the following $N+1$ constraints (each labeled by i)

$$\binom{N}{i}\lambda_0 + \binom{N-1}{i-1}\lambda_1 \geq -\binom{N-2}{i-2}, \quad N \geq i \geq 0, \tag{A12}$$

where $\binom{a}{b}$ is taken to be zero for $b < 0$. The principle of strong duality [52] ensures that the value of the objective function at the solution is equal to the extremal value of the original objective function $-\nu$.

The set of constraints defines a convex region in the λ_1, λ_0 plane as seen in Figure A1. The minimum of our dual objective generically occurs at a vertex of the boundary of the allowed region, or possibly degenerate minima can occur anywhere along an edge of the region. From Figure A1 is is clear that this occurs where Equation (A12) is an equality for two (or three in the degenerate case) consecutive values of i. Calling the first of these two values i_0, we then have the following two equations that allow us to determine the optimal values of λ_0 and λ_1 (λ_0^* and λ_1^*, respectively) as a function of i_0

$$\binom{N}{i}\lambda_0^* + \binom{N-1}{i-1}\lambda_1^* = -\binom{N-2}{i_0-2} \tag{A13}$$

$$\binom{N}{i_0+1}\lambda_0^* + \binom{N-1}{i_0}\lambda_1^* = -\binom{N-2}{i_0-1}. \tag{A14}$$

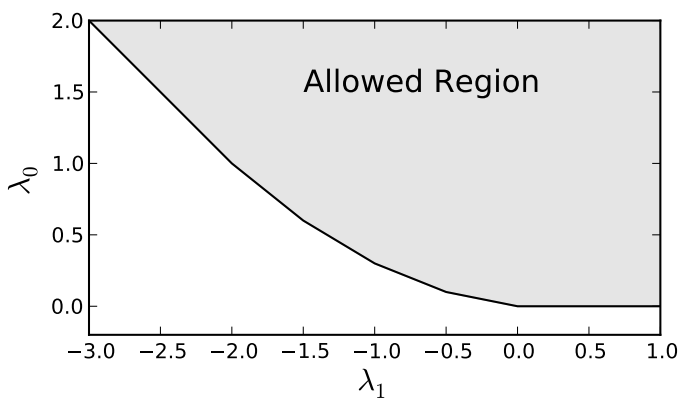

Figure A1. An example of the allowed values of λ_0 and λ_1 for the dual problem ($N = 5$).

Solving for λ_0^* and λ_1^*, we find

$$\lambda_0^* = \frac{i_0(i_0+1)}{N(N-1)} \tag{A15}$$

$$\lambda_1^* = \frac{-2i_0}{(N-1)}. \tag{A16}$$

Plugging this into Equation (A11) we find the optimal value $\mathcal{E}^*$ is

$$\begin{aligned}\mathcal{E}^* &= \lambda_0^* + \mu\lambda_1^* \\ &= \frac{i_0(i_0+1)}{N(N-1)} - \mu\frac{2i_0}{(N-1)} \\ &= \frac{i_0(i_0+1-2\mu N)}{N(N-1)}.\end{aligned} \tag{A17}$$

Now all that is left is to express i_0 as a function of μ and take the limit as N becomes large. This expression can be found by noting from Equation (A11) and Figure A1 that at the solution, i_0 satisfies

$$-m(i_0) \leq \mu \leq -m(i_0+1), \tag{A18}$$

where $m(i)$ is the slope, $d\lambda_0/d\lambda_1$, of constraint i. The expression for $m(i)$ is determined from Equation (A12),

$$\begin{aligned}m(i) &= -\frac{\binom{N-1}{i-1}}{\binom{N}{i}} \\ &= -\frac{i}{N}.\end{aligned} \tag{A19}$$

Substituting Equation (A19) into Equation (A18), we find

$$\frac{i_0}{N} \leq \mu \leq \frac{i_0+1}{N}. \tag{A20}$$

This allows us to write

$$\mu = \frac{i_0 + b(N)}{N}. \tag{A21}$$

where $b(N)$ is between 0 and 1 for all N. Solving this for i_0, we obtain

$$i_0 = N\mu - b(N) \tag{A22}$$

Substituting Equation (A22) into Equation (A17), we find

$$\begin{aligned}\mathcal{E}^* &= \frac{(N\mu - b(N))(N\mu - b(N) + 1 - 2N\mu)}{N(N-1)} \\ &= \frac{(N\mu - b(N))(-N\mu - b(N) + 1)}{N(N-1)} \\ &= -\mu^2 + \mathcal{O}\left(\frac{1}{N}\right)\end{aligned} \tag{A23}$$

Taking the large N limit we find that $\mathcal{E}^* = -\mu^2$ and by the principle of strong duality [52] the maximum value of $-\nu$ is $-\mu^2$. Therefore we have shown that for large N, the region of satisfiable constraints is simply

$$\mu^2 \leq \nu \leq \mu, \tag{A24}$$

as illustrated in Figure A2.

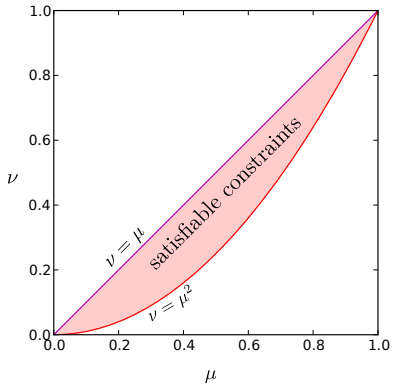

Figure A2. The red shaded region is the set of values for μ and ν that can be satisfied for at least one probability distribution in the $N \to \infty$ limit. The purple line along the diagonal where $\nu = \mu$ is the distribution for which only the all active and all inactive states have non-zero probability. It represents the global entropy minimum for a given value of μ. The red parabola, $\nu = \mu^2$, at the bottom border of the allowed region corresponds to a wide range of probability distributions, including the global maximum entropy solution for given μ in which each neuron fires independently. We find that low entropy solutions reside at this low ν boundary as well.

We can also compute upper and lower bounds on the minimum possible value for ν for finite N by taking the derivative of $\mathcal{E}^*$ (Equation (A23)) with respect to $b(N)$ and setting that to zero, to obtain $b(N) = 0.5$. Recalling that $0 \leq b(N) \leq 1$, it is clear that the only candidates for extremizing $\mathcal{E}^*$ are $b(N) \in \{0, 0.5, 1\}$, and we have:

$$\frac{N\mu(N\mu - 1)}{N(N-1)} \leq \nu_{min}(N) \leq \frac{(N\mu - 0.5)^2}{N(N-1)}. \tag{A25}$$

To obtain the exact value of the minimum of ν for finite N, we substitute the greatest integer less than or equal to μN for i_0 in Equation (A17) to obtain

$$\nu_{min} = -\frac{\lfloor \mu N \rfloor \left(\lfloor \mu N \rfloor + 1 - 2\mu N \right)}{N(N-1)}, \tag{A26}$$

where $\lfloor . \rfloor$ is the floor function. Both of the bounds in (A25) and the true ν_{min} are plotted as functions of N in Figure 3 of the main text for $\mu = 0.1$.

Appendix B. Minimum Entropy Occurs at Small Support

Our goal is to minimize the entropy function

$$S(\mathbf{p}) = \sum_{i=0}^{n_s} -p_i \log_2 p_i, \tag{A27}$$

where n_s is the number of states, the p_i satisfy a set of n_c independent linear constraints, and $p_i \geq 0$ for all i. For the main problem we consider, $n_s = 2^N$. The constraints for normalization, mean firing rates, and pairwise firing rates give

$$\begin{aligned} n_c &= 1 + N + \frac{N(N-1)}{2} \\ &= 1 + \frac{N(N+1)}{2}. \end{aligned} \tag{A28}$$

In this section we will show that the minimum occurs at the vertices of the space of allowed probabilities. Moreover, these vertices correspond to probabilities of small support—specifically a support size equal to n_c in most cases. These two facts allow us to put an upper bound on the minimum entropy of

$$\tilde{S}_2 \leq \tilde{S}_2^{hi} \approx 2\log_2(N), \tag{A29}$$

for large N.

We begin by noting that the space of normalized probability distributions $\mathcal{P} = \{p : \sum_{i\in 1}^{n_s} p_i = 1,\ p_i \geq 0\}$ is the standard simplex in $n_s - 1$ dimensions. Each linear constraint on the probabilities introduces a hyperplane in this space. If the constraints are consistent and independent, the intersection of these hyperplanes defines a $d = n_s - n_c$ affine space, which we call $\mathcal{C}$. All solutions are constrained to the intersection between $\mathcal{P}$ and $\mathcal{C}$ and this solution space is a convex polytope of dimension $\leq d$, which we refer to as $\mathcal{R}$. A point within a convex polytope can always be expressed as a linear combination of its vertices, therefore if $\{\mathbf{v}_i\}$ are the vertices of $\mathcal{R}$ we may write

$$\mathbf{p} = \sum_i^{n_v} a_i \mathbf{v}_i, \tag{A30}$$

where n_v is the total number of vertices and $\sum_i^{n_v} a_i = 1$.

Using the concavity of the entropy function, we will now show that the minimum entropy for a space of probabilities S is attained on the vertices of that space. Of course, this means that the global minimum will occur at the vertex that has the lowest entropy, $\mathbf{v}_*$

$$\begin{aligned} S(\mathbf{p}) &= S\left(\sum_i^{n_v} a_i \mathbf{v}_i\right) \\ &\geq \sum_i^{v_s} a_i S(\mathbf{v}_i) \\ &\geq \left(\sum_i^{v_s} a_i\right) S(\mathbf{v}_*) \\ &= S(\mathbf{v}_*). \end{aligned} \tag{A31}$$

Therefore,

$$\tilde{S}_2 = S(\mathbf{v}_*). \tag{A32}$$

Moreover, if a distribution satisfying the constraints exists, then there is one with at most n_c non-zero p_i (e.g., from arguments as in [41]). Together, these two facts imply that there are minimum entropy distributions with a maximum of n_c non-zero p_i. This means that even though the state space may grow exponentially with N, the support of the minimum entropy solution for fixed means and pairwise correlations will only scale quadratically with N.

This allows us to give an upper bound on the minimum entropy as,

$$\begin{aligned} \tilde{S}_2 \leq \tilde{S}_2^{hi} &= \log_2(n_c) \\ &= \log_2\left(1 + \frac{N(N+1)}{2}\right) \\ &\approx 2\log_2(N), \end{aligned} \tag{A33}$$

for large N. It is important to note how general this bound is: as long as the constraints are independant and consistent this result holds *regardless* of the specific values of the $\{\mu_i\}$ and $\{\nu_{ij}\}$.

Appendix C. The Maximum Entropy Solution

In the previous Appendix, we derived a useful upper bound on the minimum entropy solution valid for any values of $\{\mu_i\}$ and $\{\nu_{ij}\}$ that can be achieved by at least one probability distribution. In Appendix H below, we obtain a useful lower bound on the minimum entropy solution. It is straightforward to obtain an upper bound on the maximum entropy distribution valid for arbitrary achievable $\{\mu_i\}$ and $\{\nu_{ij}\}$: the greatest possible entropy for N neurons is achieved if they all fire independently with probability $1/2$, resulting in the bound $S \leq N$.

Deriving a useful lower bound for the maximum entropy for arbitrary allowed constraints $\{\mu_i\}$ and $\{\nu_{ij}\}$ is a subtle problem. In fact, merely specifying how an ensemble of binary units should be grown from some finite initial size to arbitrarily large N in such a way as to "maintain" the low-level statistics of the original system raises many questions.

For example, typical neural populations consist of units with varying mean activities, so how should the mean activities of newly added neurons be chosen? For that matter, what correlational structure should be imposed among these added neurons and between each new neuron and the existing units? For each added neuron, any choice will inevitably change the histograms of mean activities and pairwise correlations for any initial ensemble consisting of more than one neuron, except in the special case of uniform constraints. Even for the relatively simple case of uniform constraints, we have seen that there are small systems that cannot be grown beyond some critical size while maintaining uniform values for $\{\mu\}$ and $\{\nu\}$ (see Figure 2 in the main text). The problem of determining whether it is even mathematically possible to add a neuron with some predetermined set of pairwise correlations with existing neurons can be much more challenging for the general case of nonuniform constraints.

For these reasons, we will leave the general problem of finding a lower bound on the maximum entropy for future work and focus here on the special case of uniform constraints:

$$\mu_i = \mu, \quad \forall i = 1, \ldots, N \tag{A34}$$

$$\nu_{ij} = \nu, \quad \forall i \neq j. \tag{A35}$$

We will obtain a lower bound on the maximum entropy of the system with the use of an explicit construction, which will necessarily have an entropy, S_2^{con}, that does not exceed that of the maximum entropy solution. We construct our model system as follows: with probability β, all N neurons are active and set to 1, otherwise each neuron is independently set to 1 with probability η. The conditions required for this model to match the desired mean activities and pairwise correlations across the population are given by

$$\mu = \beta + (1-\beta)\eta, \tag{A36}$$

$$\nu = \beta + (1-\beta)\eta^2. \tag{A37}$$

Inverting these equations to isolate β and η yields

$$\beta = \frac{\nu - \mu^2}{1 + \nu - 2\mu}, \tag{A38}$$

$$\eta = \frac{\mu - \nu}{1 - \mu}. \tag{A39}$$

The entropy of the system is then

$$\begin{aligned} S_2^{con} &= -\beta \log_2 \beta - (1-\beta)\log_2(1-\beta) \\ &\quad + (1-\beta)N\left[-\eta \log_2 \eta - (1-\eta)\log_2(1-\eta)\right] \\ &= w + xN, \end{aligned} \tag{A40}$$

where w and x are nonnegative constants for all allowed values of μ and ν that can be achieved for arbitrarily large systems. x is nonzero provided $\nu < \mu$ (i.e., $\beta \neq 1 \cap \eta \neq \{0,1\}$), so the entropy of the system will grow linearly with N for any uniform constraints achievable for arbitrarily large systems except for the case in which all neurons are perfectly correlated.

Using numerical methods, we can empirically confirm the linear scaling of the entropy of the true maximum entropy solution for the case of uniform constraints that can be achieved by arbitrarily large systems. In general, the constraints can be written

$$\mu_i = \langle s_i \rangle = \sum_{\vec{s}} p(\vec{s}) s_i, \quad i = 1, \ldots, N, \tag{A41}$$

$$\nu_{ij} = \langle s_i s_j \rangle = \sum_{\vec{s}} p(\vec{s}) s_i s_j, \quad i \neq j, \tag{A42}$$

where the sums run over all 2^N states of the system. In order to enforce the constraints, we can add terms involving Lagrange multipliers λ_i and γ_{ij} to the entropy in the usual fashion to arrive at a function to be maximized

$$\begin{aligned} \mathcal{S}(p(\vec{s})) = &- \sum_{\vec{s}} p(\vec{s}) \log_2 p(\vec{s}) \\ &- \sum_i \lambda_i \left(\sum_{\vec{s}} p(\vec{s}) s_i - \mu_i \right) \\ &- \sum_{i<j} \gamma_{ij} \left(\sum_{\vec{s}} p(\vec{s}) s_i s_j - \nu_{ij} \right). \end{aligned} \tag{A43}$$

Maximizing this function gives us the Boltzmann distribution for an Ising spin glass

$$p(\vec{s}) = \frac{1}{\mathcal{Z}} \exp\left(- \sum_i \lambda_i s_i - \sum_{i<j} \gamma_{ij} s_i s_j \right), \tag{A44}$$

where $\mathcal{Z}$ is the normalization factor or partition function. The values of λ_i and γ_{ij} are left to be determined by ensuring this distribution is consistent with our constraints $\{\mu_i\}$ and $\{\nu_{ij}\}$.

For the case of uniform constraints, the Lagrange multipliers are themselves uniform:

$$\lambda_i = \lambda, \quad \forall i, \tag{A45}$$

$$\gamma_{ij} = \gamma, \quad \forall i < j. \tag{A46}$$

This allows us to write the following expression for the maximum entropy distribution:

$$p(\vec{s}) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda \sum_i s_i - \gamma \sum_{i<j} s_i s_j \right). \tag{A47}$$

If there are k neurons active, this becomes

$$p(k) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda k - \gamma \frac{k(k-1)}{2} \right). \tag{A48}$$

Note that there are $\binom{N}{k}$ states with probability $p(k)$. Using expression (A48), we find the maximum entropy by using the `fsolve` function from the `SciPy` package of Python subject to constraints (A34) and (A35).

As Figure A3 shows, for fixed uniform constraints the entropy scales linearly as a function of N, even in cases where the correlations between all pairs of neurons (ν) are quite large, provided that $\mu^2 \leq \nu < \mu$. Moreover, for uniform constraints with anticorrelated units (i.e., pairwise correlations

below the level one would observe for independent units), we find empirically that the maximum entropy still scales approximately linearly up to the maximum possible system size consistent with these constraints, as illustrated by Figure 3 in the main text.

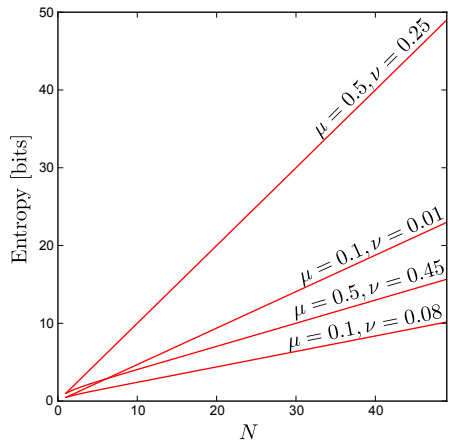

Figure A3. For the case of uniform constraints achievable by arbitrarily large systems, the maximum possible entropy scales linearly with system size, N, as shown here for various values of μ and ν. Note that this linear scaling holds even for large correlations, provided that $\nu < \mu$. For the extreme case $\nu = \mu$, all the neurons are perfectly correlated so the entropy of the ensemble does not change with increasing N.

Appendix D. Minimum Entropy for Exchangeable Probability Distributions

Although the values of the firing rate (μ) and pairwise correlations (ν) may be identical for each neuron and pair of neurons, the probability distribution that gives rise to these statistics need not be exchangeable as we have already shown. Indeed, as we explain below, it is possible to construct non-exchangeable probability distributions that have dramatically lower entropy then both the maximum and the minimum entropy for exchangeable distributions. That said, exchangeable solutions are interesting in their own right because they have large N scaling behavior that is distinct from the global entropy minimum, and they provide a symmetry that can be used to lower bound the information transmission rate close to the maximum possible across all distributions.

Restricting ourselves to exchangeable solutions represents a significant simplification. In the general case, there are 2^N probabilities to consider for a system of N neurons. There are N constraints on the firing rates (one for each neuron) and $\binom{N}{2}$ pairwise constraints (one for each pair of neurons). This gives us a total number of constrains (n_c) that grows quadratically with N:

$$n_c = 1 + \frac{N(N+1)}{2}. \tag{A49}$$

However in the exchangeable case, all states with the same number of spikes have the same probability so there are only $N+1$ free parameters. Moreover, the number of constraints becomes 3 as there is only one constraint each for normalization, firing rate, and pairwise firing rate (as expressed in Equations (A5)–(A7), respectively). In general, the minimum entropy solution for exchangeable distributions should have the minimum support consistent with these three constraints. Therefore, the minimum entropy solution should have at most three non-zero probabilities (see Appendix B).

For the highly symmetrical case with $\mu = 1/2$ and $\nu = 1/4$, we can construct the exchangeable distribution with minimum entropy for all even N. This distribution consists of the all ones state, the all zeros state, and all states with $N/2$ ones. The constraint $\mu = 1/2$ implies that $p(0) = p(N)$, and the condition $\nu = 1/4$ implies

$$p(N/2) = \frac{N-1}{N}\frac{(N/2)!^2}{N!}, \quad N \text{ even}, \tag{A50}$$

which corresponds to an entropy of

$$\tilde{S}_2^{exch} = \frac{\log_2(2N)}{N} + \frac{N-1}{N}\log_2\left(\frac{NN!}{(N/2)!^2(N-1)}\right). \tag{A51}$$

Using Sterling's approximation and taking the large N limit, this simplifies to

$$\tilde{S}_2^{exch} \approx N - {}^1\!/_2\log_2(N) - {}^1\!/_2\log_2(2\pi) + O\left[\frac{\log_2(N)}{N}\right]. \tag{A52}$$

For arbitrary values of μ, ν and N, it is difficult to determine from first principles which three probabilities are non-zero for the minimum entropy solution, but fortunately the number of possibilities $\binom{N+1}{3}$ is now small enough that we can exhaustively search by computer to find the set of non-zero probabilities corresponding to the lowest entropy.

Using this technique, we find that the scaling behavior of the exchangeable minimum entropy is linear with N as shown in Figure A4. We find that the asymptotic slope is positive, but less than that of the maximum entropy curve, for all $\nu \neq \mu^2$. For the symmetrical case, $\nu = \mu^2$, our exact expression Equation (A51) for the exchangeable distribution consisting of the all ones state, the all zeros state, and all states with $N/2$ ones agrees with the minimum entropy exchangeable solution found by exhaustive search, and in this special case the asymptotic slope is identical to that of the maximum entropy curve.

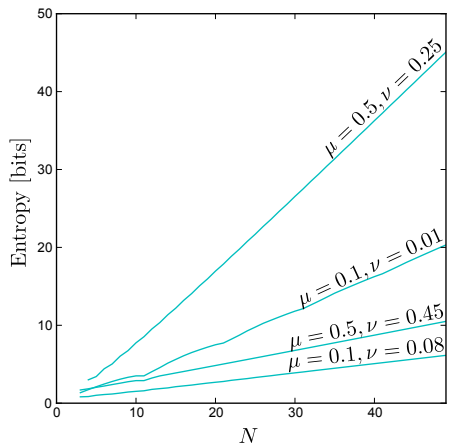

Figure A4. The minimum entropy for exchangeable distributions versus N for various values of μ and ν. Note that, like the maximum entropy, the exchangeable minimum entropy scales linearly with N as $N \to \infty$, albeit with a smaller slope for $\nu \neq \mu^2$. We can calculate the entropy exactly for $\mu = 0.5$ and $\nu = 0.25$ as $N \to \infty$, and we find that the leading term is indeed linear: $\tilde{S}_2^{exch} \approx N - 1/2\log_2(N) - 1/2\log_2(2\pi) + O[\log_2(N)/N]$.

Appendix E. Construction of a Low Entropy Distribution for All Values of μ and ν

We can construct a probability distribution with roughly N^2 states with nonzero probability out of the full 2^N possible states of the system such that

$$\mu = \frac{n}{N}, \quad \nu = \frac{n(n-1)}{N(N-1)}, \tag{A53}$$

where N is the number of neurons in the network and n is the number of neurons that are active in every state. Using this solution as a basis, we can include the states with all neurons active and all

neurons inactive to create a low entropy solution for all allowed values for μ and ν (See Appendix G). We refer to the entropy of this low entropy construction $\tilde{S}_2^{con2}$ to distinguish it from the entropy ($\tilde{S}_2^{con}$) of another low entropy solution described in the next section. Our construction essentially goes back to Joffe [60] as explained by Luby in [33].

We derive our construction by first assuming that N is a prime number, but this is not actually a limitation as we will be able to extend the result to all values of N. Specifically, non-prime system sizes are handled by taking a solution for a larger prime number and removing the appropriate number of neurons. It should be noted that occasionally the solution derived using the next largest prime number does not necessarily have the lowest entropy and occasionally we must use even larger primes to find the minimum entropy possible using this technique; all plots in the main text were obtained by searching for the lowest entropy solution using the 10 smallest primes that are each at least as great as the system size N.

We begin by illustrating our algorithm with a concrete example; following this illustrative case we will prove that each step does what we expect in general. Consider $N = 5$, and $n = 3$. The algorithm is as follows:

1. Begin with the state with $n = 3$ active neurons in a row:

   11100

2. Generate new states by inserting progressively larger gaps of 0s before each 1 and wrapping active states that go beyond the last neuron back to the beginning. This yields $N - 1 = 4$ unique states including the original state:

   11100
   10101
   11010
   10011

3. Finally, "rotate" each state by shifting each pattern of ones and zeros to the right (again wrapping states that go beyond the last neuron). This yields a total of $N(N-1)$ states:

11100	01110	00111	10011	11001
10101	11010	01101	10110	01011
11010	01101	10110	01011	10101
10011	11001	11100	01110	00111

4. Note that each state is represented twice in this collection, removing duplicates we are left with $N(N-1)/2$ total states. By inspection we can verify that each neuron is active in $n(N-1)/2$ states and each pair of neurons is represented in $n(n-1)/2$ states. Weighting each state with equal probability gives us the values for μ and ν stated in Equation (A53).

Now we will prove that this construction works in general for N prime and any value of n by establishing (1) that step 2 of the above algorithm produces a set of states with n spikes; (2) that this method produces a set of states that when weighted with equal probability yield neurons that all have the same firing rates and pairwise statistics; and (3) that this method produces at least double redundancy in the states generated as stated in step 4 (although in general there may be a greater redundancy). In discussing (1) and (2) we will neglect the issue of redundancy and consider the states produced through step 3 as distinct.

First we prove that step 2 always produces states with n neurons, which is to say that no two spikes are mapped to the same location as we shift them around. We will refer to the identity of the spikes by their location in the original starting state; this is important as the operations in step 2 and 3 will change the relative ordering of the original spikes in their new states. With this in mind, the location of the ith spike with a spacing of s between them will result in the new location l (here the original state with all spikes in a row is $s = 1$):

$$l = (s \cdot i) \bmod N, \tag{A54}$$

where $i \in \{0, 1, 2, ..., n-1\}$. In this form, our statement of the problem reduces to demonstrating that for given values of s and N, no two values of i will result in the same l. This is easy to show by contradiction. If this were the case,

$$\begin{aligned} (s \cdot i_1) \bmod N &= (s \cdot i_2) \bmod N \\ \Rightarrow (s \cdot (i_1 - i_2)) \bmod N &= 0. \end{aligned} \tag{A55}$$

For this to be true, either s or $(i_1 - i_2)$ must contain a factor of N, but each are smaller than N so we have a contradiction. This also demonstrates why N must be prime—if it were not, it would be possible to satisfy this equation in cases where s and $(i_1 - i_2)$ contain between them all the factors of N.

It is worth noting that this also shows that there is a one-to-one mapping between s and l given i. In other words, each spike is taken to every possible neuron in step 2. For example, if $N = 5$, and we fix $i = 2$:

$$\begin{aligned} 0 \cdot 2 \bmod 5 &= 0 \\ 1 \cdot 2 \bmod 5 &= 2 \\ 2 \cdot 2 \bmod 5 &= 4 \\ 3 \cdot 2 \bmod 5 &= 1 \\ 4 \cdot 2 \bmod 5 &= 3 \end{aligned}$$

If we now perform the operation in step 3, then the location l of spike i becomes

$$l = (s \cdot i + d) \bmod N, \tag{A56}$$

where d is the amount by which the state has been rotated (the first column in step 3 is $d = 0$, the second is $d = 1$, etc.). It should be noted that step 3 trivially preserves the number of spikes in our states so we have established that steps 2 and 3 produce only states with n spikes.

We now show that each neuron is active, and each pair of neurons is simultaneously active, in the same number of states. This way when each of these states is weighted with equal probability, we find symmetric statistics for these two quantities.

Beginning with the firing rate, we ask how many states contain a spike at location l. In other words, how many combinations of s, i, and d can we take such that Equation (A56) is satisfied for a given l. For each choice of s and i there is a unique value of d that satisfies the equation. s can take values between 1 and $N-1$, and i takes values from 0 to $n-1$, which gives us $n(N-1)$ states that include a spike at location l. Dividing by the total number of states $N(N-1)$ we obtain an average firing rate of

$$\mu = \frac{n}{N}. \tag{A57}$$

Consider neurons at l_1 and l_2; we wish to know how many values of s, d, i_1 and i_2 we can pick so that

$$l_1 = (s \cdot i_1 + d) \bmod N, \tag{A58}$$

$$l_2 = (s \cdot i_2 + d) \bmod N. \tag{A59}$$

Taking the difference between these two equations, we find

$$\Delta l = (s \cdot (i_2 - i_1)) \bmod N. \tag{A60}$$

From our discussion above, we know that this equation uniquely specifies s for any choice of i_1 and i_2. Furthermore, we must pick d such that Equations (A58) and (A59) are satisfied. This means that for each choice of i_1 and i_2 there is a unique choice of s and d, which results in a state that includes active neurons at locations l_1 and l_2. Swapping i_1 and i_2 will result in a different s and d. Therefore, we have $n(n-1)$ states that include any given pair—one for each choice of i_1 and i_2. Dividing this number by the total number of states, we find a correlation ν equal to

$$\nu = \frac{n(n-1)}{N(N-1)}, \tag{A61}$$

where N is prime.

Finally we return to the question of redundancy among states generated by steps 1 through 3 of the algorithm. Although in general there may be a high level of redundancy for choices of n that are small or close to N, we can show that in general there is at least a twofold degeneracy. Although this does not impact our calculation of μ and ν above, it does alter the number of states, which will affect the entropy of system.

The source of the twofold symmetry can be seen immediately by noting that the third and fourth rows of our example contain the same set of states as the second and first respectively. The reason for this is that each state in the $s = 4$ case involves spikes that are one leftward step away from each other just as $s = 1$ involves spikes that are one rightward shift away from each other. The labels we have been using to refer to the spikes have reversed order but the set of states are identical. Similarly the $s = 3$ case contains all states with spikes separated by two leftward shifts just as the $s = 2$ case. Therefore, the set of states with $s = a$ is equivalent to the set of states with $s = N - a$. Taking this degeneracy into account, there are at most $N(N-1)/2$ unique states; each neuron spikes in $n(N-1)/2$ of these states and any given pair spikes together in $n(n-1)/2$ states.

Because these states each have equal probability the entropy of this system is bounded from above by

$$\tilde{S}_2^{con2} \leq \log_2\left(\frac{N(N-1)}{2}\right), \tag{A62}$$

where N is prime. As mentioned above, we write this as an inequality because further degeneracies among states beyond the factor of two that always occurs are possible for some prime numbers. In fact, in order to avoid non-monotonic behavior, the curves for S_2^{con2} shown in Figures 1 and 3 of the main text were generated using the lowest entropy found for the 10 smallest primes greater than N for each value of N.

We can extend this result to arbitrary values for N including non-primes by invoking the Bertrand-Chebyshev theorem, which states that there always exists at least one prime number p with $n < p < 2n - 2$ for any integer $n > 1$:

$$\tilde{S}_2^{con2} \leq \log_2\left(N(2N-1)\right), \tag{A63}$$

where N is any integer. Unlike the maximum entropy and the entropy of the exchangeable solution, which we have shown to both be extensive quantities, this scales only logarithmically with the system size N.

Appendix F. Another Low Entropy Construction for the Communications Regime, $\mu = 1/2$ & $\nu = 1/4$

In addition to the probability distribution described in the previous section, we also rediscovered another low entropy construction in the regime most relevant for engineered communications systems ($\mu = 1/2$, $\nu = 1/4$) that allows us to satisfy our constraints for a system of N neurons with only $2N$ active states. Below we describe a recursive algorithm for determining the states for arbitrarily large systems—states needed for $N = 2^q$ are built from the states needed for $N = 2^{q-1}$, where q is any

integer greater than 2. This is sometimes referred to as a Hadamard matrix. Interestingly, this specific example goes back to Sylvester in 1867 [22], and it was recently discussed in the context of neural modeling by Macke and colleagues [21].

We begin with $N = 2^1 = 2$. Here we can easily write down a set of states that when weighted equally lead to the desired statistics. Listing these states as rows of zeros and ones, we see that they include all possible two-neuron states:

$$\begin{matrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{matrix} \tag{A64}$$

In order to find the states needed for $N = 2^2 = 4$ we replace each 1 in the above by

$$\begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix} \tag{A65}$$

and each 0 by

$$\begin{matrix} 0 & 0 \\ 1 & 0 \end{matrix} \tag{A66}$$

to arrive at a new array for twice as many neurons and twice as many states with nonzero probability:

$$\begin{matrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{matrix} \tag{A67}$$

By inspection, we can verify that each new neuron is spiking in half of the above states and each pair is spiking in a quarter of the above states. This procedure preserves $\mu = 1/2$, $\nu = 1/4$, and $\langle s_i s_j s_k \rangle = 1/8$ for all neurons; thus providing a distribution that mimics the statistics of independent binary variables up to third order (although it does not for higher orders). Let us consider the the proof that $\mu = 1/2$ is preserved by this transformation. In the process of doubling the number of states from N^q to N^{q+1}, each neuron with firing rate $\mu^{(q)}$ "produces" two new neurons with firing rates $\mu_1^{(q+1)}$ and $\mu_2^{(q+1)}$. It is clear from Equations (A65) and (A66) that we obtain the following two relations,

$$\mu_1^{(q+1)} = \mu^{(q)}, \tag{A68}$$

$$\mu_2^{(q+1)} = 1/2. \tag{A69}$$

It is clear from these equations that if we begin with $\mu^{(1)} = 1/2$ that this will be preserved by this transformation. By similar, but more tedious, methods one can show that $\nu = 1/4$, and $\langle s_i s_j s_k \rangle = 1/8$.

Therefore, we are able to build up arbitrarily large groups of neurons that satisfy our statistics using only $2N$ states by repeating the procedure that took us from $N = 2$ to $N = 4$. Since these states are weighted with equal probability we have an entropy that grows only logarithmically with N

$$\tilde{S}_2^{con} = \log_2(2N), \quad N = 2^q,\ q = 2, 3, 4, \ldots. \tag{A70}$$

We mention briefly a geometrical interpretation of this probability distribution. The active states in this distribution can be thought of as a subset of $2N$ corners on an N dimensional hypercube with the property that the separation of almost every pair is the same. Specifically, for each active state, all but one of the other active states has a Hamming distance of exactly $N/2$ from the original state; the remaining state is on the opposite side of the cube, and thus has a Hamming distance of N. In other words, for any pair of polar opposite active states, there are $2N-2$ active states around the "equator".

We can extend Equation (A70) to arbitrary numbers of neurons that are not multiples of 2 by taking the least multiple of 2 at least as great as N, so that in general:

$$\tilde{S}_2^{con} = \lceil \log_2(2N) \rceil \leq \log_2(N) + 2, \quad N \geq 2. \tag{A71}$$

By adding two other states we can extend this probability distribution so that it covers most of the allowed region for μ and ν while remaining a low entropy solution, as we now describe.

We remark that the authors of [31,34] provide a lower bound of Ω_N for the sample size possible for a pairwise independent binary distribution, making the sample size of our novel construction essentially optimal.

Appendix G. Extending the Range of Validity for the Constructions

We now show that each of these low entropy probability distributions can be generalized to cover much of the allowed region depicted in Figure A2; in fact, the distribution derived in Appendix E can be extended to include all possible combinations of the constraints μ and ν. This can be accomplished by including two additional states: the state where all neurons are silent and the state where all neurons are active. If we weight these states by probabilities p_0 and p_1 respectively and allow the $N(N-1)/2$ original states to carry probability p_n in total, normalization requires

$$p_0 + p_n + p_1 = 1. \tag{A72}$$

We can express the value of the new constraints (μ' and ν') in terms of the original constraint values (μ and ν) as follows:

$$\begin{aligned} \mu' &= (1 - p_0 - p_1)\mu + p_1 \\ &= (1 - p_0)\mu + p_1(1 - \mu), && \text{(A73)} \\ \nu' &= (1 - p_0)\nu + p_1(1 - \nu). && \text{(A74)} \end{aligned}$$

These values span a triangular region in the μ-ν plane that covers the majority of satisfiable constraints. Figure A5 illustrates the situation for $\mu = 1/2$. Note that by starting with other values of μ, we can construct a low entropy solution for any possible constraints μ' and ν'.

With the addition of these two states, the entropy of the expanded system $\tilde{S}_2^{con2'}$ is bounded from above by

$$\tilde{S}_2^{con2'} = p_n \tilde{S}_2^{con2} - \sum_{i \in \{0,1,n\}} p_i \log_2(p_i) \tag{A75}$$

For given values of μ' and ν', the p_i are fixed and only the first term depends on N. Using Equations (A73) and (A74),

$$p_n = \frac{\mu' - \nu'}{\mu - \nu}. \tag{A76}$$

This allows us to rewrite Equation (A75) as

$$\tilde{S}_2^{con2'} \leq \left(\frac{\mu' - \nu'}{\mu - \nu} \right) \tilde{S}_2^{con2} + \log_2(3). \tag{A77}$$

We are free to select μ and ν to minimize the first coefficient for a desired μ' and ν', but in general we know this coefficient is less than 1 giving us a simple bound,

$$\tilde{S}_2^{con2'} \leq \tilde{S}_2^{con2} + \log_2(3). \tag{A78}$$

Like the original distribution, the entropy of this distribution scales logarithmically with N. Therefore, by picking our original distribution properly, we can find low entropy distributions for any μ and ν for which the number of active states grows as a polynomial in N (see Figure A5).

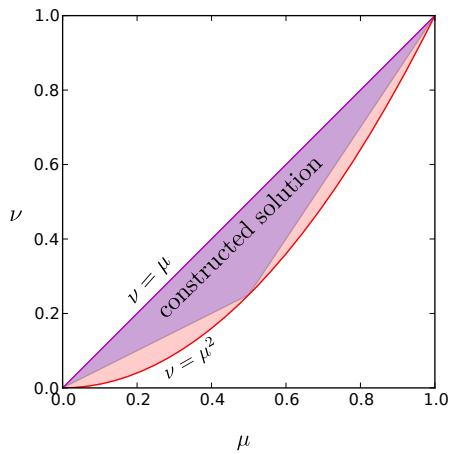

Figure A5. The full shaded region includes all allowed values for the constraints μ and ν for all possible probability distributions, replotted from Figure A2. As described in Appendices E and G, one of our low-entropy constructed solutions can be matched to any of the allowed constraint values in the full shaded region, whereas the constructed solution described in Appendix F can achieve any of the values within the triangular purple shaded region. Note that even with this second solution, we can cover most of the allowed region. Each of our constructed solutions have entropies that scale as $S \sim \ln(N)$.

Similarly, we can extend the range of validity for the construction described in Appendix F to the triangular region shown in Figure A2 by assigning probabilities p_0, p_1, and $p_{N/2}$ to the all silent state, all active state, and the total probability assigned to the remaining $2N-2$ states of the original model, respectively. The entropy of this extended distribution must be no greater than the entropy of the original distribution (Equation (A71)), since the same number of states are active, but now they are not weighted equally, so this remains a low entropy distribution.

Appendix H. Proof of the Lower Bound on Entropy for Any Distribution Consistent with Given $\{\mu_i\}$ & $\{\nu_{ij}\}$

Using the concavity of the logarithm function, we can derive a *lower* bound on the minimum entropy. Our lower bound asymptotes to a constant except for the special case $\mu_i = 1/2$, $\forall\ i$, and $\nu_{ij} = 1/4$, $\forall\ i, j$, which is especially relevant for communication systems since it matches the low order statistics of the global maximum entropy distribution for an unconstrained set of binary variables.

We begin by bounding the entropy from below as follows:

$$\begin{aligned} S(\mathbf{p}) &= -\sum_w p(w) \log_2 p(w) \\ &= \langle -\log_2 p(w) \rangle \\ &\geq -\log_2 \langle p(w) \rangle \\ &\geq -\log_2 \|\mathbf{p}\|_2^2, \end{aligned} \tag{A79}$$

where $\mathbf{p}$ represents the full vector of all 2^N state probabilities, and we have used $\langle\cdot\rangle$ to denote an average over the distribution $p(w)$. The third step follows from Jensen's inequality applied to the convex function $-\log(x)$.

Now we seek an upper bound on $\mathbf{p}^2$. This can be obtained by starting with the matrix representation C of the constraints (for now, we consider each state of the system, $\vec{s}_i$, as a binary column vector, where i denotes the state and each of the N components is either 1 or 0):

$$\begin{aligned} C &= \left\langle \vec{s}\vec{s}^T \right\rangle \\ &= \sum_i p(s_i)\vec{s}_i\vec{s}_i^T, \end{aligned} \tag{A80}$$

where C is an $N \times N$ matrix. In this form, the diagonal entries of C, c_{mm}, are equal to μ_m and the off diagonal entries, c_{mn}, are equal to ν_{mn}.

For the calculation that follows, it is expedient to represent words of the system as $\vec{\bar{s}} \in \{-1,1\}^N$ rather than $\vec{s} \in \{0,1\}^N$ (i.e., -1 represents a silent neuron instead of 0). The relationship between the two can be written

$$\vec{\bar{s}} = 2\vec{s} - \vec{1}, \tag{A81}$$

where $\vec{1}$ is the vector of all ones. Using this expression, we can relate $\bar{C}$ to C:

$$\begin{aligned} \bar{C} &= \left\langle \vec{\bar{s}}\vec{\bar{s}}^T \right\rangle \\ &= \left\langle (2\vec{s} - \vec{1})(2\vec{s}^T - \vec{1}^T) \right\rangle \\ &= 4\left\langle \vec{s}\vec{s}^T \right\rangle - 2\left\langle \vec{s}\vec{1}^T \right\rangle - 2\left\langle \vec{1}\vec{s}^T \right\rangle + \vec{1}\vec{1}^T, \end{aligned} \tag{A82}$$

$$\bar{c}_{mn} = 4c_{mn} - 2c_{mm} - 2c_{nn} + 1. \tag{A83}$$

This reduces to

$$\bar{c}_{mm} = 1, \tag{A84}$$

$$\bar{c}_{mn} = 4\nu_{mn} - 2(\mu_m + \mu_n) + 1, \quad m \neq n. \tag{A85}$$

Returning to Equation (A80) to find an upper bound on $\mathbf{p}^2$, we take the square of the Frobenius norm of $\bar{C}$:

$$\begin{aligned} \|\bar{C}\|_F^2 &= \mathrm{Tr}(\bar{C}^T\bar{C}) \\ &= \mathrm{Tr}\left(\left(\sum_i p(\vec{\bar{s}}_i)\vec{\bar{s}}_i\vec{\bar{s}}_i^T \right) \times \left(\sum_j p(\vec{\bar{s}}_j)\vec{\bar{s}}_j\vec{\bar{s}}_j^T \right) \right) \\ &= \mathrm{Tr}\left(\sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\vec{\bar{s}}_i\vec{\bar{s}}_i^T\vec{\bar{s}}_j\vec{\bar{s}}_j^T \right) \\ &= \sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\,\mathrm{Tr}\left(\vec{\bar{s}}_i\vec{\bar{s}}_i^T\vec{\bar{s}}_j\vec{\bar{s}}_j^T \right) \\ &= \sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\,\mathrm{Tr}\left(\vec{\bar{s}}_j^T\vec{\bar{s}}_i\vec{\bar{s}}_i^T\vec{\bar{s}}_j \right) \\ &= \sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\left(\vec{\bar{s}}_i \cdot \vec{\bar{s}}_j \right)^2 \\ &\geq \sum_i p(\vec{\bar{s}}_i)^2\left(\vec{\bar{s}}_i \cdot \vec{\bar{s}}_i \right)^2 \\ &= N^2\|\mathbf{p}\|_2^2. \end{aligned} \tag{A86}$$

The final line is where our new representation pays off: in this representation, $\vec{\bar{s}}_i \cdot \vec{\bar{s}}_i = N$. This gives us the desired upper bound for $\mathbf{p}^2$:

$$\frac{\|\bar{C}\|_F^2}{N^2} \geq \|\mathbf{p}\|_2^2. \tag{A87}$$

Using Equations (A84)–(A86), we can express $\|\bar{C}\|_F^2$ in terms of μ and ν:

$$\begin{aligned} \|\bar{C}\|_F^2 &= \sum_m \bar{c}_{mm}^2 + \sum_{m \neq n} \bar{c}_{mn}^2 \\ &= N + (4\nu_{mn} - 2(\mu_m + \mu_n) + 1)^2. \end{aligned} \tag{A88}$$

Combining this result with Equations (A87) and (A79), we obtain a lower bound for the entropy for any distribution consistent with any given sets of values $\{\mu_i\}$ and $\{\nu_{ij}\}$:

$$\begin{aligned} S(\mathbf{p}) \geq \tilde{S}_2^{lo} &= \log_2 \left(\frac{N^2}{N + \sum_{i \neq j} \alpha_{ij}} \right) \\ &= \log_2 \left(\frac{N}{1 + (N-1)\bar{\alpha}} \right) \end{aligned} \tag{A89}$$

where $\alpha_{ij} = (4\nu_{ij} - 2(\mu_i + \mu_j) + 1)^2$ and $\bar{\alpha}$ is the average value of α_{ij} over all i, j with $i \neq j$.

In the case of uniform constraints, this becomes

$$S(\mathbf{p}) \geq \tilde{S}_2^{lo} = \log_2 \left(\frac{N}{1 + (N-1)\alpha} \right), \tag{A90}$$

where $\alpha = (4(\nu - \mu) + 1)^2$.

For large values of N this lower bound asymptotes to a constant

$$\lim_{N \to \infty} \tilde{S}_2^{lo} = \log_2 (1/\bar{\alpha}). \tag{A91}$$

The one exception is when $\bar{\alpha} = 0$. In the large N limit, this case is limited to when $\mu_i = 1/2$ and $\nu_{ij} = 1/4$ for all i, j. Each α_{ij} is positive semi-definite; therefore, $\bar{\alpha} = 0$ only when each $\alpha_{ij} = 0$. In other words,

$$4\nu_{ij} - 2(\mu_i + \mu_j) + 1 = 0 \quad \forall i \neq j \tag{A92}$$

But in the large N limit,

$$\mu_i \mu_j \leq \nu_{ij} \leq \min(\mu_i, \mu_j). \tag{A93}$$

Without loss of generality, we assume that $\mu_i \leq \mu_j$. In this case,

$$0 \leq \mu_i \leq \frac{1}{2} \tag{A94}$$

$$\frac{1}{2} \leq \mu_j \leq \mu_i + \frac{1}{2}, \tag{A95}$$

and

$$\nu_{ij} = \frac{\mu_i + \mu_j}{2} - \frac{1}{4}. \tag{A96}$$

Of course, that means that in order to satisfy $\bar{\alpha} = 0$ each pair must have one μ less than of equal to $1/2$ and the other greater than or equal to $1/2$. The only way this may be true for all possible pairs is if all μ_i are equal to $1/2$. According to Equation (A96), all ν_{ij} must then be equal to $1/4$. This is precisely the communication regime, and in this case our lower-bound scales logarithmically with N,

$$\tilde{S}_2^{lo} = \log_2(N). \tag{A97}$$

References

1. Pathria, R. *Statistical Mechanics*; Butterworth Hein: Oxford, UK, 1972.
2. Russ, W.; Lowery, D.; Mishra, P.; Yaffe, M.; Ranganathan, R. Natural-like function in artificial WW domains. *Nature* **2005**, *437*, 579–583.
3. Socolich, M.; Lockless, S.W.; Russ, W.P.; Lee, H.; Gardner, K.H.; Ranganathan, R. Evolutionary information for specifying a protein fold. *Nature* **2005**, *437*, 512–518.
4. Mora, T.; Walczak, A.M.; Bialek, W.; Callan, C.G. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 5405–5410.
5. 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.
6. Shlens, J.; Field, G.D.; Gauthier, J.L.; Grivich, M.I.; Petrusca, D.; Sher, A.; Litke, A.M.; Chichilnisky, E.J. The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* **2006**, *26*, 8254–8266.
7. Tkacik, G.; Schneidman, E.; Berry, I.; Michael, J.; Bialek, W. Ising models for networks of real neurons. *arXiv* **2006**, arXiv:q-bio/0611072.
8. Tang, A.; Jackson, D.; Hobbs, J.; Chen, W.; Smith, J.; Patel, H.; Prieto, A.; Petrusca, D.; Grivich, M.I.; Sher, A.; et al. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.* **2008**, *28*, 505–518.
9. Shlens, J.; Field, G.D.; Gauthier, J.L.; Greschner, M.; Sher, A.; Litke, A.M.; Chichilnisky, E.J. The Structure of Large-Scale Synchronized Firing in Primate Retina. *J. Neurosci.* **2009**, *29*, 5022–5031.
10. Ganmor, E.; Segev, R.; Schneidman, E. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9679–9684.
11. Yu, S.; Huang, D.; Singer, W.; Nikolic, D. A small world of neuronal synchrony. *Cereb. Cortex* **2008**, *18*, 2891–2901.
12. Köster, U.; Sohl-Dickstein, J. Higher order correlations within cortical layers dominate functional connectivity in microcolumns. *arXiv* **2013**, arXiv:1301.0050v1.
13. Hamilton, L.S.; SohlDickstein, J.; Huth, A.G.; Carels, V.M.; Deisseroth, K.; Bao, S. Optogenetic activation of an inhibitory network enhances feedforward functional connectivity in auditory cortex. *Neuron* **2013**, *80*, 10661076.
14. Bialek, W.; Cavagna, A.; Giardina, I.; Mora, T.; Silvestri, E.; Viale, M.; Walczak, A. Statistical mechanics for natural flocks of birds. *arXiv* **2011**, arXiv:org:1107.0604.
15. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
16. Bethge, M.; Berens, P. Near-maximum entropy models for binary neural representations of natural images. In *Advances in Neural Information Processing Systems*; Platt, J., Koller, D., Singer, Y., Roweis, S., Eds.; MIT Press: Cambridge, MA, USA, 2008; Volume 20, pp. 97–104.
17. Roudi, Y.; Nirenberg, S.H.; Latham, P.E. Pairwise maximum entropy models for studying large biological systems: When they can and when they can't work. *PLoS Comput. Biol.* **2009**, *5*, e1000380.
18. Nirenberg, S.H.; Victor, J.D. Analyzing the activity of large populations of neurons: How tractable is the problem? *Curr. Opin. Neurobiol.* **2007**, *17*, 397–400.
19. Azhar, F.; Bialek, W. When are correlations strong? *arXiv* **2010**, arXiv:org:1012.5987.
20. Tkačik, G.; Marre, O.; Amodei, D.; Schneidman, E.; Bialek, W.; Berry, M.J. Searching for collective behavior in a large network of sensory neurons. *PLoS Comput. Biol.* **2014**, *10*, e1003408.
21. Macke, J.H.; Opper, M.; Bethge, M. Common Input Explains Higher-Order Correlations and Entropy in a Simple Model of Neural Population Activity. *Phys. Rev. Lett.* **2011**, *106*, 208102.
22. Sylvester, J. Thoughts on inverse orthogonal matrices, simultaneous sign successions, and tessellated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers. *Philos. Mag.* **1867**, *34*, 461–475.
23. Diaconis, P. Finite forms of de Finetti's theorem on exchangeability. *Synthese* **1977**, *36*, 271–281.
24. Shannon, C. A mathematical theory of communications, I and II. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
25. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 1991.

26. De Schutter, B. Minimal state-space realization in linear system theory: An overview. *J. Comput. Appl. Math.* **2000**, *121*, 331–354.
27. Shalizi, C.; Crutchfield, J. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.* **2001**, *104*, 817–879.
28. Carter, J.; Wegman, M. Universal classes of hash functions. *J. Comput. Syst. Sci.* **1979**, *18*, 143–154.
29. Sipser, M. A complexity theoretic approach to randomness. In Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing, Boston, MA, USA, 25–27 April 1983; pp. 330–335.
30. Stockmeyer, L. The complexity of approximate counting. In Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing, Boston, MA, USA, 25–27 April 1983; pp. 118–126.
31. Chor, B.; Goldreich, O.; Hasted, J.; Freidmann, J.; Rudich, S.; Smolensky, R. The bit extraction problem or t-resilient functions. In Proceedings of the 26th Annual Symposium on Foundations of Computer Science, Portland, OR, USA, 21–23 October 1985; pp. 396–407.
32. Karp, R.; Wigderson, A. A fast parallel algorithm for the maximal independent set problem. *JACM* **1985**, *32*, 762–773.
33. Luby, M. A simple parallel algorithm for the maximal independent set problem. *SIAM J. Comput.* **1986**, *15*, 1036–1053.
34. Alon, N.; Babai, L.; Itai, A. A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms* **1986**, *7*, 567–583.
35. Alexi, W.; Chor, B.; Goldreich, O.; Schnorr, C. RSA and Rabin functions: Certain parts are as hard as the whole. *SIAM J. Comput.* **1988**, *17*, 194–209.
36. Chor, B.; Goldreich, O. On the power of two-point based sampling. *J. Complex.* **1989**, *5*, 96–106.
37. Berger, B.; Rompel, J. Simulating (log cn)-wise independence in NC. *JACM* **1991**, *38*, 1026–1046.
38. Schulman, L. Sample spaces uniform on neighborhoods. In Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing, Victoria, BC, Canada, 4–6 May 1992; pp. 17–25.
39. Luby, M. Removing randomness in parallel computation without a processor penalty. *J. Comput. Syst. Sci.* **1993**, *47*, 250–286.
40. Motwani, R.; Naor, J.; Naor, M. The probabilistic method yields deterministic parallel algorithms. *J. Comput. Syst. Sci.* **1994**, *49*, 478–516.
41. Koller, D.; Megiddo, N. Constructing small sample spaces satisfying given constraints. In Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, San Diego, CA, USA, 16–18 May 1993; pp. 268–277.
42. Karloff, H.; Mansour, Y. On construction of k-wise independent random variables. In Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, Montreal, QC, Canada, 23–25 May 1994; pp. 564–573.
43. Castellana, M.; Bialek, W. Inverse spin glass and related maximum entropy problems. *Phys. Rev. Lett.* **2014**, *113*, 117204.
44. Fischer, K.H.; Hertz, J.A. *Spin Glasses*; Cambridge University Press: Cambridge, UK, 1991.
45. Tanaka, T. Mean-field theory of Boltzmann machine learning. *Phys. Rev. Lett. E* **1998**, *58*, 2302.
46. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554.
47. Hyvärinen, A. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Trans. Neural Netw.* **2007**, *18*, 1529–1531.
48. Broderick, T.; Dudík, M.; Tkačik, G.; Schapire, R.; Bialek, W. Faster solutions of the inverse pairwise Ising problem. *arXiv* **2007**, arXiv:0712.2437.
49. Sohl-Dickstein, J.; Battaglino, P.B.; DeWeese, M.R. New method for parameter estimation in probabilistic models: Minimum probability flow. *Phys. Rev. Lett.* **2011**, *107*, 220601.
50. Tkačik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, M.J., II; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech. Theory Exp.* **2013**, *2013*, P03011.
51. Toulouse, G. The frustration model. In *Modern Trends in the Theory of Condensed Matter*; Pekalski, A., Przystawa, J.A., Eds.; Springer: Berlin/Heidelberg, Germany, 1980; Volume 115, pp. 195–203.
52. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
53. Rosen, J.B. Global Minimization of a Linearly Constrained Function by Partition of Feasible Domain. *Math. Oper. Res.* **1983**, *8*, 215–230.

54. Candes, E.; Tao, T. Decoding by linear programming. *IEEE Trans. Inf. Theory* **2005**, *51*, 4203–4215.
55. Donoho, D. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306.
56. Donoho, D.L. Compressed Sensing. Available online: https://statweb.stanford.edu/~donoho/Reports/2004/CompressedSensing091604.pdf (accessed on 17 August 2017).
57. Sarvotham, S.; Baron, D.; Baraniuk, R.G. Measurements vs. Bits: Compressed Sensing meets Information Theory. Available online: https://scholarship.rice.edu/handle/1911/20323 (accessed on 17 August 2017).
58. Hao, B. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*; World Scientific: Singapore, 1989.
59. Luby, M.; Wigderson, A. *Pairwise Independence and Derandomization*; Now Publishers Inc.: Breda, The Netherlands, 2006; Volume 4.
60. Joffe, A. On a set of almost deterministic k-independent random variables. *Ann. Probab.* **1974**, *2*, 161–162.
61. MacWilliams, F.; Sloane, N. *Error Correcting Codes*; North Holland: New York, NY, USA, 1977.
62. Hedayat, A.; Sloane, N.; Stufken, J. *Orthogonal Arrays: Theory and Applications*; Springer: New York, NY, USA, 1999.
63. Hall, M. *Combinatorial Theory*; Blaisdell Publishing Company: London, UK, 1967.
64. Lancaster, H. Pairwise statistical independence. *Ann. Math. Stat.* **1965**, *36*, 1313–1317.
65. Rieke, F.; Warland, D.; van Steveninck, R.d.R.; Bialek, W. *Spikes: Exploring the Neural Code*; The MIT Press: Cambridge, MA, USA ;1999; p. 416.
66. Advani, M.; Lahiri, S.; Ganguli, S. Statistical mechanics of complex neural systems and high dimensional data. *J. Stat. Mech. Theory Exp.* **2013**, *2013*, P03014.
67. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.* **2007**, *98*, 1064–1072.
68. Rolls, E.T.; Treves, A. The neuronal encoding of information in the brain. *Prog. Neurobiol.* **2011**, *95*, 448–490.
69. Crumiller, M.; Knight, B.; Yu, Y.; Kaplan, E. Estimating the amount of information conveyed by a population of neurons. *Front. Neurosci.* **2011**, *5*, 90.
70. Strong, S.P.; Koberle, R.; van Steveninck, R.d.R.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *80*, 197.
71. Nemenman, I.; Bialek, W.; van Steveninck, R.d.R. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E* **2004**, *69*, 056111.
72. Borst, A.; Theunissen, F.E. Information theory and neural coding. *Nat. Neurosci.* **1999**, *2*, 947–957.
73. Quian Quiroga, R.; Panzeri, S. Extracting information from neuronal populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.* **2009**, *10*, 173–185.
74. Gale, D.; Kuhn, H.W.; Tucker, A.W. *Linear Programming and the Theory of Games*; Koopmans, T.C., Ed.; Wiley: New York, NY, USA, 1951; pp. 317–329.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).