arXiv:1209.3744v1 [physics.bio-ph] 17 Sep 2012

Minimum and maximum entropy distributions for binary systems with known means and pairwise correlations

Badr F. Albanna[ab1], Christopher Hillar [bd2], Jascha Sohl-Dickstein[be3], and Michael R. DeWeese[abc4]
[a]Department of Physics, [b]Redwood Center for Theoretical Neuroscience
[c]Helen Wills Neuroscience Institute, [e]Biophysics Graduate Group
University of California, Berkeley, CA 94720
[d]Mathematical Sciences Research Institute, Berkeley, CA 94720-5070
[1]badr_albanna@berkeley.edu, [2]chillar@msri.edu, [3]jascha@berkeley.edu, [4]deweese@berkeley.edu

(Dated: September 18, 2012)

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 lower and upper bounds on the entropy for both the minimum and maximum entropy distributions over binary units with fixed mean and pairwise correlation, and we construct distributions for several relevant cases. Surprisingly, the minimum entropy solution has entropy scaling logarithmically with system size, unlike the linear behavior of the maximum entropy solution, resolving an open question in neuroscience. 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.

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], neural population activity [5–11], and even flock behavior [12] quite well (*cf.*, [13]). This is perhaps surprising since the usual physical arguments involving ergodicity or equality among energetically accessible states are not obviously applicable for such systems, though such models have been justified in terms of imposing no structure beyond what is explicitly measured [5, 14]. Conversely, it is not clear to what extent this good agreement was inevitable. If the space of distributions were sufficiently constrained by observations, then agreement is an unavoidable consequence of the constraints rather than a consequence of the unique suitability of the maximum entropy model. In neuroscience, there is also controversy [5, 9, 15–17] over the notion that small pairwise correlations can conspire to constrain the behavior of large neural ensembles, and it has been shown [9, 15] that pairwise models do not always allow accurate extrapolation from small populations to large ensembles.

Previous authors have studied these issues with maximum entropy models expanded to second- [5], third- [15], and fourth-order [17]. Here we use non-perturbative methods to derive rigorous upper and lower bounds on the entropy of the *minimum* entropy distribution for fixed means and pairwise correlations, and we construct explicit low and high entropy models for the full range of possible uniform first- and second-order constraints (Eqs. (3)-(6); Figs. 1, 2). 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. Thus, entropy is only weakly constrained by these statistics, and the success of maximum entropy models in biology [2–12], when it occurs for large enough systems [15], represents a real triumph of the maximum entropy approach.

Our results demonstrate that empirically measured first-, second-, and third-order statistics are essentially inconsequential for testing coding optimality in a broad class of engineered information transmission systems, whereas the existence of other statistical properties, such as finite exchangeability [18], do guarantee information transmission near channel capacity [19, 20], 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 [21] – a form of optimal model selection based on an interpretation of Occam's Razor complementary to that of Jaynes [14]. Our results also have implications for computer science as algorithms for generating binary random variables with low entropy have found many applications (e.g., [22–36]).

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 $\mathbf{p} \in [0,1]^{2^N}$, $\sum_{\vec{s}} p(\vec{s}) = 1$.

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 [37]. This "inverse" problem of inferring the interaction and magnetic field terms in an

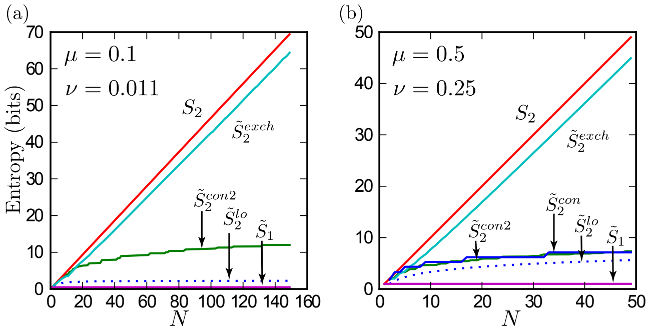

FIG. 1. The minimum entropy grows no faster than logarithmically with the system size N for any mean activity level μ and pairwise correlation strength ν. (a) In a parameter regime relevant for neural population activity in the retina [5, 6] ($\mu = 0.1$, $\nu = 0.011$), we can construct an explicit low entropy solution ($\tilde{S}_2^{con2}$) that grows logarithmically with N, unlike the linear behavior of the maximum entropy solution (S_2). (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.

Ising spin glass Hamiltonian that produce the measured means and correlations is nontrivial, but there has been progress [17, 38–42]. The maximum entropy distribution is not the only one consistent with these observed statistics, however. In fact, there are many others, 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 until, at least in principle, 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 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 any observed first- or second-order statistics for any system size N. We find that the gap grows linearly with N, up to a logarithmic correction.

We will restrict ourselves here to symmetric constraints; that is, values of *mean firing rates* and *pairwise correlations* are uniform:

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

$$\nu_{ij} = \nu, \quad \text{for all } i \neq j. \tag{2}$$

Given symmetric constraints, we find the following bounds on the maximum and minimum entropies for fixed values of μ and ν. For the *maximum entropy*:

$$(1 - x)N - C_1(\mu, \nu) \leq S_2 \leq N, \tag{3}$$

where $x = \frac{\nu - \mu}{\mu(1-\mu)}$ and C_1 is a constant that only depends on μ and ν. For the *minimum entropy*:

$$\log_2\left(\frac{N}{1 + (N-1)\alpha(\mu,\nu)}\right) \leq \tilde{S}_2 \leq \log_2(N(2N-1)), \tag{4}$$

where $\alpha(\mu, \nu) = (4(\nu - \mu) + 1)^2$. In most cases, the lower bound in Eq. (4) asymptotes to a constant for large N, but in the special case where μ and ν have values consistent with *independent neurons* ($\mu = 1/2$ and $\nu = 1/4$), we can give the tighter bound:

$$\log_2(N) \leq \tilde{S}_2 \leq \log_2(N) + 2. \tag{5}$$

An important class of probability distributions are the *exchangeable distributions* [18], which 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 that we conjecture is a minimum entropy exchangeable solution with entropy $\tilde{S}_2^{exch}$ that scales linearly with N:

$$C_2(\mu, \nu)N - \mathcal{O}(\log_2 N) \leq \tilde{S}_2^{exch} \leq C_3(\mu, \nu)N, \tag{6}$$

where C_2 and C_3 are constants that only depend on μ and ν. We have empirically confirmed that this is indeed a minimum entropy exchangeable solution for $N \leq 200$.

We obtained these bounds by constructing families of low entropy distributions and exploiting the geometry of the entropy function. Entropy is a strictly concave function of the probabilities and therefore has a unique maximum that can be identified using standard methods [43], at least for sufficiently small or symmetric systems. Indeed, it is easy to show (see Appendix B) that the maximum entropy S_2 for any system with specified mean and pairwise correlation scales linearly with N (Eq. (3), Fig. 1).

By contrast, the minimum entropy distribution exists at a vertex of the allowed space, where most states have probability zero ([44]; see Appendix C). Our challenge then is to determine in which vertex (or vertices) a minimum resides. The entropy function is nonlinear, precluding 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 bound $\tilde{S}_2^{lo}$ on the entropy of the minimum entropy solution for all N (Fig. 1), and we have constructed two families of explicit solutions with low entropies ($\tilde{S}_2^{con}$ and $\tilde{S}_2^{con2}$; Figs. 1,2) for a broad parameter regime covering all allowed values for μ and ν.

Using the concavity of the entropy function together with Jensen's inequality, one can derive an upper bound on the entropy [20], but similar methods also allow us to obtain a lower bound $\tilde{S}_2^{lo}$ as in Eq. (4) on the entropy (see Appendix H):

$$\tilde{S}_2(N,\mu,\nu) \geq \tilde{S}_2^{lo} = \log_2\left(\frac{N}{1+(N-1)\alpha(\mu,\nu)}\right), \qquad (7)$$

where $\alpha(\mu,\nu) = (4(\nu-\mu)+1)^2$, and $\tilde{S}_2(N,\mu,\nu)$ is the minimum entropy given a network of size N with constraints μ and ν. This lower bound asymptotes to the constant value $\log_2(1/\alpha(\mu,\nu))$ as N becomes large except for the special case:

$$\nu = \mu - {}^1\!/\!_4, \qquad (8)$$

where α vanishes. In the large N limit, we have the inequality $\mu \geq \nu \geq \mu^2$ (see Appendix A), so the only values of μ and ν satisfying Eq. (8) are

$$\mu = {}^1\!/\!_2, \quad \nu = \mu^2 = {}^1\!/\!_4. \qquad (9)$$

In this particular case, the lower bound Eq. (7) scales logarithmically with N, rather than as a constant, but for 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 variables).

In addition to this lower bound, we can also construct probability distributions that provide upper bounds on the entropy of a minimum entropy solution. Each of these solutions has an entropy that grows logarithmically with N (see Appendices E, **??**, Eqs. (4)-(5)):

$$\begin{aligned}\tilde{S}_2^{con2} &\leq \log_2\Big(\lceil N\rceil_p\left(\lceil N\rceil_p - 1\right)\Big) - 1 \\ &\leq \log_2\left(N(2N-1)\right), \qquad (10)\\ \tilde{S}_2^{con} &= \lceil \log_2(N) + 1\rceil \\ &\leq \log_2(N) + 2, \qquad (11)\end{aligned}$$

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

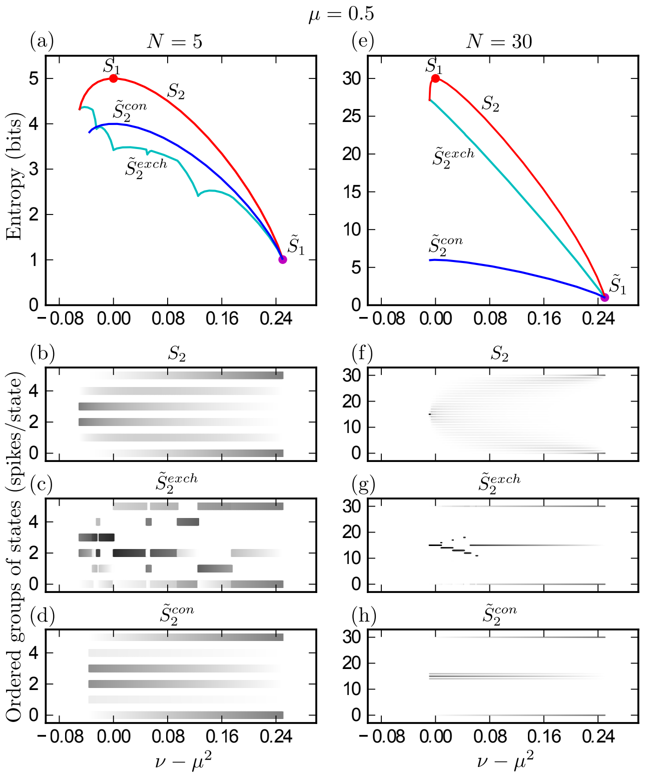

FIG. 2. Minimum and maximum entropy models for symmetric constraints. (a) Entropy as a function of the strength of pairwise correlations for the maximum entropy model (S_2), finitely exchangeable minimum entropy model ($\tilde{S}_2^{exch}$), and a constructed low entropy solution ($\tilde{S}_2^{con}$), 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.

Consider the special case of first- and second-order constraints (Eq. 9) 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 (Fig. 1a). In fact, we have constructed an explicit solution (Eq. (11); Figs. 1b,2a,d,e,h; Appendix F), whose entropy $\tilde{S}_2^{con}$ is never more than two bits above our lower bound (Eq. (7)) 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 covering most of the allowed region (see Appendix G) while still remaining a distribution whose entropy grows only logarithmically with N.

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 the 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 is in a sense reversed: Observed statistics obeying a symmetry (the observations about the system) are produced by an underlying model that does not.

We now examine consequences for engineered communication systems. Specifically, consider a device such as a digital camera that exploits compressed sensing [45, 46] to reduce the dimensionality of its image representations. A compressed sensing scheme involves 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 [45, 47]. Without such knowledge, it would be desirable to know which empirically measured output statistics could tell us 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 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, measuring the extent to which other statistical properties are obeyed can yield strong guarantees of system performance. In particular, exchangeability is one such constraint. Fig. 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 experimentally find that any exchangeable distribution has as much entropy as the maximum entropy solution, up to terms of order $\log_2(N)$ (see Appendix D).

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 [48]. 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 [27, 28, 49–51], combinatorial block designs [26, 52], Hadamard matrix theory [36, 53], and linear programming [35], among others. Our construction here of a pairwise independent distribution with entropy $\tilde{S}_2^{con}$ adds to this literature and is completely elementary.

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. In some settings, minimum entropy models can also provide a floor on information transmission, complementary to channel capacity, which provides a ceiling on system performance.

Acknowledgments

The authors would like to thank 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. MRD is grateful for support from the Hellman Foundation, the McDonnell Foundation, the McKnight Foundation, and the Mary Elizabeth Rennie Endowment for Epilepsy Research. CH was supported under an NSF All-Institutes Postdoctoral Fellowship administered by the Mathematical Sciences Research Institute through its core grant DMS-0441170.

APPENDIX A: ALLOWED RANGE OF ν GIVEN μ ACROSS ALL DISTRIBUTIONS FOR LARGE N

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 for large N 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 [43, 54], 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* [18], 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

$$p(100) = p(010) = p(001), \tag{A.1}$$

$$p(110) = p(101) = p(011). \tag{A.2}$$

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 (Eqs. (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:

$$p_e(\vec{s}) \equiv \sum_{\sigma} \frac{p(\sigma(\vec{s}))}{N!}, \tag{A.3}$$

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 symmetric 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

$$1 = \sum_{i=0}^{N} \binom{N}{i} p(i). \tag{A.4}$$

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:

$$\mu = \sum_{i=1}^{N} \binom{N-1}{i-1} p(i). \tag{A.5}$$

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:

$$\nu = \sum_{i=2}^{N} \binom{N-2}{i-2} p(i). \tag{A.6}$$

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

$$-\nu = -\sum_{i=2}^{N} \binom{N-2}{i-2} p(i) \tag{A.7}$$

subject to

$$1 = \sum_{i=0}^{N} \binom{N}{i} p(i), \tag{A.8}$$

$$\mu = \sum_{i=1}^{N} \binom{N-1}{i-1} p(i), \tag{A.9}$$

$$p(i) \geq 0, \quad \text{for all } i. \tag{A.10}$$

This gives us the following dual problem: Minimize

$$\mathcal{E} \equiv \lambda_0 + \mu\lambda_1, \tag{A.11}$$

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

$$\binom{N}{i}\lambda_0 + \binom{N-1}{i-1}\lambda_1 \geq -\binom{N-2}{i-2}, \quad N \geq i \geq 0, \tag{A.12}$$

where $\binom{a}{b}$ is taken to be zero for $b < 0$. The principle of strong duality [43] 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 (A.1). The minimum of our dual objective generically occurs at a vertex of the boundary of the allowed region. It can be shown that this occurs where Eq. (A.12) is an equality for two adjacent values of i. Calling the first of these two values i_0, we then have the following two equations that allow us to

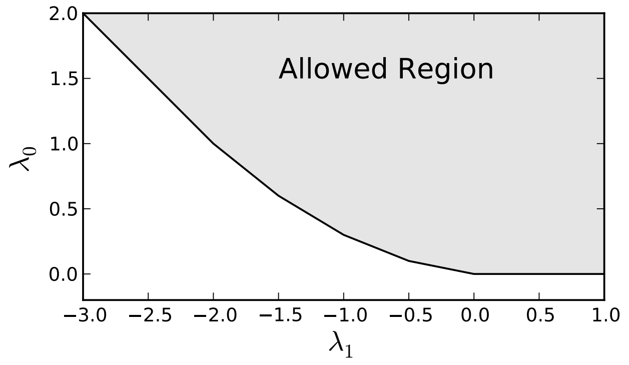

FIG. A.1. An example of the allowed values of λ_0 and λ_1 for the dual problem ($N = 5$).

determine the optimal values of λ_0 and λ_1 (λ_0^* and λ_1^*, respectively) as a function of i_0

$$\binom{N}{i}\lambda_0^* + \binom{N-1}{i-1}\lambda_1^* = -\binom{N-2}{i_0-2} \tag{A.13}$$

$$\binom{N}{i_0+1}\lambda_0^* + \binom{N-1}{i_0}\lambda_1^* = -\binom{N-2}{i_0-1}. \tag{A.14}$$

Solving for λ_0^* and λ_1^*, we find

$$\lambda_0^* = \frac{i_0(i_0+1)}{N(N-1)} \tag{A.15}$$

$$\lambda_1^* = \frac{-2i_0}{(N-1)}. \tag{A.16}$$

Plugging this into Eq. (A.11) we find the optimal value $\mathcal{E}^*$ is

$$\begin{aligned} \mathcal{E}^* &= \lambda_0^* + \mu\lambda_1^* && \text{(A.17)}\\ &= \frac{i_0(i_0+1)}{N(N-1)} - \mu\frac{2i_0}{(N-1)} && \text{(A.18)}\\ &= \frac{i_0(i_0+1-2\mu N)}{N(N-1)}. && \text{(A.19)} \end{aligned}$$

Now all that is left it 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 Eq. (A.11) and Fig. A.1 that at the solution, i_0 satisfies

$$-m(i_0) \leq \mu \leq -m(i_0+1), \tag{A.20}$$

where $m(i)$ is the slope, $d\lambda_0/d\lambda_1$, of constraint i. The expression for $m(i)$ is determined from Eq. (A.12),

$$\begin{aligned} m(i) &= -\frac{\binom{N-1}{i-1}}{\binom{N}{i}} && \text{(A.21)}\\ &= -\frac{i}{N}. && \text{(A.22)} \end{aligned}$$

Substituting Eq. (A.22) into Eq. (A.20), we find

$$\frac{i_0}{N} \leq \mu \leq \frac{i_0+1}{N}. \tag{A.23}$$

This allows us to write

$$\mu = \frac{i_0 + b(N)}{N}. \tag{A.24}$$

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

$$i_0 = N\mu - b(N) \tag{A.25}$$

Substituting Eq. (A.25) into Eq. (A.19), we find

$$\begin{aligned} \mathcal{E}^* &= \frac{(N\mu - b(N))(N\mu - b(N) + 1 - 2N\mu)}{N(N-1)} && \text{(A.26)}\\ &= \frac{(N\mu - b(N))(-N\mu - b(N) + 1)}{N(N-1)} && \text{(A.27)}\\ &= -\mu^2 + \mathcal{O}\left(\frac{1}{N}\right) && \text{(A.28)} \end{aligned}$$

Taking the large N limit we find that $\mathcal{E}^* = -\mu^2$ and by the principle of strong duality [43] the maximum value of $-\nu$ is $-\mu^2$. Therefore we have shown that for large N, the region of satisfiable constraints is simply

$$\mu^2 \leq \nu \leq \mu, \tag{A.29}$$

as illustrated in Fig. A.2.

APPENDIX B: THE MAXIMUM ENTROPY SOLUTION

We begin by stating the general form for the solution for known mean firing rate and pairwise constraints and impose the symmetry that all statistics are equal across neurons and pairs of neurons. We will then demonstrate that for arbitrary fixed values for μ and ν, the maximum entropy must scale linearly with N as $N \to \infty$.

In general, the constraints can be written

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

$$\nu = \langle s_i s_j \rangle = \sum_{\vec{s}} p(\vec{s}) s_i s_j, \quad i \neq j, \tag{B.2}$$

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

$$\begin{aligned} \mathcal{S}(p(\vec{s})) = &-\sum_{\vec{s}} p(\vec{s}) \log_2 p(\vec{s}) \\ &-\sum_i \lambda_i \left(\sum_{\vec{s}} p(\vec{s}) s_i - \mu\right) && \text{(B.3)}\\ &-\sum_{i<j} \gamma_{ij} \left(\sum_{\vec{s}} p(\vec{s}) s_i s_j - \nu\right). \end{aligned}$$

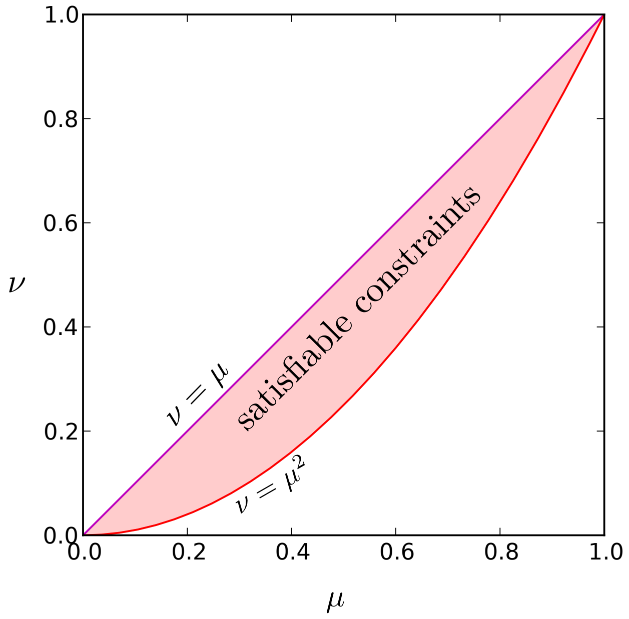

FIG. A.2. The red shaded region is the set of values for μ and ν that can be satisfied for at least one probability distribution in the $N \to \infty$ limit. The purple line along the diagonal where $\nu = \mu$ is the distribution for which only the all active and all inactive states have non-zero probability. 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.

Maximizing this function gives us the Boltzmann distribution for an Ising model

$$p(\vec{s}) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_i \lambda_i s_i - \sum_{i<j} \gamma_{ij} s_i s_j\right), \tag{B.4}$$

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 μ and ν.

It can be shown that for symmetric constraints the Lagrange multipliers are uniform. In other words,

$$\lambda_i = \lambda, \quad \forall i, \tag{B.5}$$

$$\gamma_{ij} = \gamma, \quad \forall i < j. \tag{B.6}$$

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

$$p(\vec{s}) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda \sum_i s_i - \gamma \sum_{i<j} s_i s_j\right). \tag{B.7}$$

If there are k neurons active, this becomes

$$p(k) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda k - \gamma \frac{k(k-1)}{2}\right). \tag{B.8}$$

Note that there are $\binom{N}{k}$ states with probability $p(k)$.

Using expression (B.8), we find the maximum entropy by using the `fsolve` function from the `SciPy` package of Python subject to constraints (B.1) and (B.2).

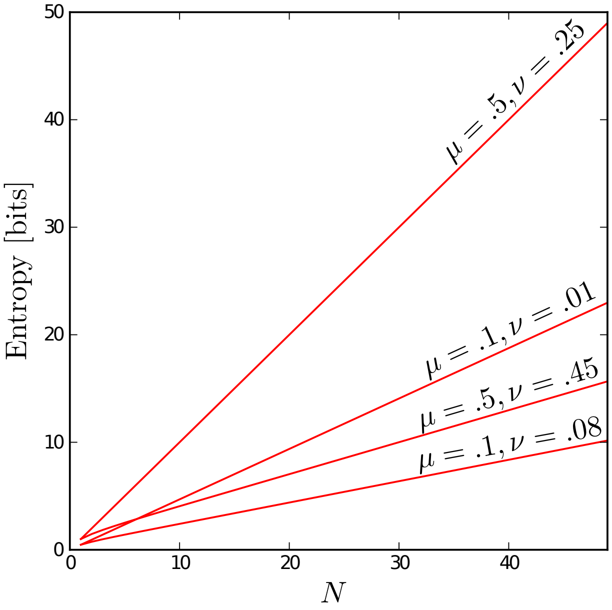

FIG. B.1. The maximum possible entropy scales linearly with system size, N, as shown here for various values of μ and ν. Note that this linear scaling holds even for large correlations.

As Fig. B.1 shows, the entropy scales linearly as a function of N, even in cases where the correlations between all pairs of neurons (ν) are quite large. While this is perhaps a surprising result, we can see that this must be the case for independent neurons, the maximum entropy solution with $\nu = \mu^2$. Because each neuron is independent, the entropy of this system must certainly scale linearly with N.

Moreover, we can construct a distribution that has entropy with linear scaling for any allowed values of μ and ν using this solution. Recall that the vector $\mathbf{p}$ represents the full distribution over all 2^N states. Consider the following probability distribution $\mathbf{p}_{pop}$, which we will call the "population spike" model. This model contains only two states with non-zero probability: The state with all neurons active (p_1) and the state with no active neurons (p_0). They are weighted so that the firing rate of this model matches that of the independence model:

$$p_1 = \mu, \tag{B.9}$$

$$p_0 = 1 - \mu. \tag{B.10}$$

As mentioned above, in this model ν is equal to its maximum allowed value, μ. Because the independent model has the smallest allowed value of ν (in the large N limit), we can combine these two models to create a one-parameter family of distributions that have fixed μ value and cover all allowed values for ν. The independent part of this distribution will guarantee that the entire family has an entropy that scales linearly with N; thus, the true maximum must grow *at least* linearly with N as well.

Our new distribution $\mathbf{p}_{mix}$ is simply

$$\mathbf{p}_{mix} = (1-x)\mathbf{p}_{ind} + x\mathbf{p}_{pop}, \quad \text{where } 0 \leq x \leq 1 \quad \text{(B.11)}$$

$\mathbf{p}_{mix}$ has firing rate μ (just like both $\mathbf{p}_{ind}$ and $\mathbf{p}_{pop}$) and $\nu = (1-x)\mu^2 + x\mu$.

Because entropy is a concave function [20], by Jensen's inequality the entropy of $\mathbf{p}_{mix}$ is bounded below by

$$S[\mathbf{p}_{mix}] \geq (1-x)S[\mathbf{p}_{ind}] + xS[\mathbf{p}_{pop}]. \quad \text{(B.12)}$$

For fixed μ and ν the second term is a constant in N, whereas the first term grows linearly with N. This implies that the true maximum entropy must grow at least linearly with N for any fixed values of μ and ν.

We note that there is a simple upper bound on the entropy that also scales linearly with N. The maximum possible entropy for fixed N is obtained by setting all probabilities equal to one another yielding an entropy of exactly N (in fact, this is the entropy of the independence model with $\mu = 1/2$). Considering that both the upper bound and lower bound for the maximum entropy for fixed μ and ν scale linearly, the maximum entropy itself must also scale linearly for large N, consistent with our computations (Fig. B.1).

APPENDIX C: MINIMUM ENTROPY OCCURS AT SMALL SUPPORT

Our goal is to minimize the entropy function

$$S(\mathbf{p}) = \sum_{i=0}^{n_s} -p_i \log_2 p_i, \quad \text{(C.1)}$$

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, $n_s = 2^N$ and normalization, mean firing rates and pairwise firing rates give $n_c = 1 + N + N(N-1)/2$. For the exchangeable case with symmetric constraints, $n_s = N+1$ and $n_c = 3$.

Our task is therefore to minimize a globally concave function over a $d = n_s - n_c$ dimensional linear (affine) space L contained in the (compact) simplex $\{p : \sum_{i\in 1}^{n_s} p_i = 1,\ p_i \geq 0\}$. It is well known that the minima of such a problem occur at the vertices of the boundary of the space [44], which necessarily have some p_i equal to zero, unless L intersects the simplex in a single point. 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 [35]). Together, these two facts imply that there are minimum entropy distributions with a maximum of n_c non-zero p_i (and can occasionally have fewer). 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. In fact, we know that for certain values of μ and ν solutions can have a far smaller support because the construction shown in Appendix F has a support size that scales only linearly with 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:

$$n_c = 1 + \frac{N(N+1)}{2}. \quad \text{(D.1)}$$

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 Eqs. (A.4), (A.5), and (A.6), 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.

For the 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 zeroes 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

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

which corresponds to an entropy of

$$\begin{aligned}\tilde{S}_2^{exch} &= \frac{\log_2(2N)}{N} \\ &+ \frac{N-1}{N}\log_2\left(\frac{NN!}{(N/2)!^2(N-1)}\right) \quad \text{(D.3)}\\ &\approx N - 1/2\log_2(N) - 1/2\log_2(2\pi) \\ &+ O\left[\frac{\log_2(N)}{N}\right]. \quad \text{(D.4)}\end{aligned}$$

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 Fig. D.1. 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 Eq. (D.3) 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.

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

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

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 [49] as explained by Luby in [27].

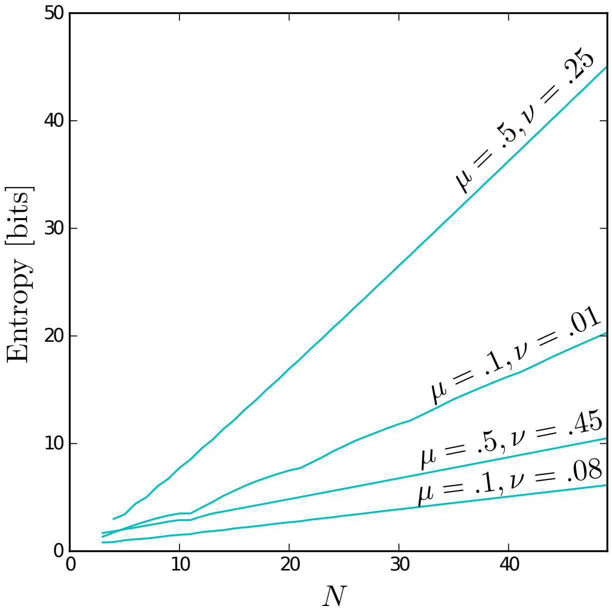

FIG. D.1. The minimum entropy for exchangeable distributions versus N for various values of μ and ν. Note that, like the maximum entropy, the exchangeable minimum entropy scales linearly with N as $N \to \infty$, albeit with a smaller slope for $\nu \neq \mu^2$. We can calculate the entropy exactly for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

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. Wighting each state with equal probability gives us the values for μ and ν stated in Eq. (E.1)

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 $s=1$):

$$l = (s \cdot i) \bmod N, \tag{E.2}$$

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,

$$\begin{aligned} (s \cdot i_1) \bmod N &= (s \cdot i_2) \bmod N & \text{(E.3)}\\ \Rightarrow (s \cdot (i_1 - i_2)) \bmod N &= 0. & \text{(E.4)} \end{aligned}$$

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

$$l = (s \cdot i + d) \bmod N, \tag{E.5}$$

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 Eq. (E.5) 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

$$\mu = \frac{n}{N}. \tag{E.6}$$

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

$$\begin{aligned} l_1 &= (s \cdot i_1 + d) \bmod N, & \text{(E.7)}\\ l_2 &= (s \cdot i_2 + d) \bmod N. & \text{(E.8)} \end{aligned}$$

Taking the difference between these two equations, we find

$$\Delta l = (s \cdot (i_2 - i_1)) \bmod N. \tag{E.9}$$

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 Eqs. (E.7) and (E.8) 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

$$\nu = \frac{n(n-1)}{N(N-1)}, \tag{E.10}$$

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

$$\tilde{S}_2^{con2} \le \log_2\left(\frac{N(N-1)}{2}\right), \tag{E.11}$$

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 Figs. 1,2 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 $n > 1$:

$$\tilde{S}_2^{con2} \le \log_2\left(N(2N-1)\right), \tag{E.12}$$

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, $\mu = 1/2$ & $\nu = 1/4$

We have found another low entropy construction in the regime most relevant for 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. The algorithm to determine the states needed is recursive in that the 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.

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:

$$\begin{array}{cc} 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{array} \tag{F.1}$$

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

$$\begin{array}{cc} 1 & 1 \\ 0 & 1 \end{array} \tag{F.2}$$

and each 0 by

$$\begin{array}{cc} 0 & 0 \\ 1 & 0 \end{array} \tag{F.3}$$

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

$$\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{array} \tag{F.4}$$

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 Eqs. (F.2) and (F.3) that we obtain the following two relations,

$$\mu_1^{(q+1)} = \mu^{(q)}, \tag{F.5}$$

$$\mu_2^{(q+1)} = 1/2. \tag{F.6}$$

$$\tag{F.7}$$

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

$$\tilde{S}_2^{con} = \log_2(2N), \quad N = 2^q, \ q = 2, 3, 4, \ldots. \tag{F.8}$$

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 Eq. (F.8) 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:

$$\tilde{S}_2^{con} = \lceil \log_2(2N) \rceil \leq \log_2(N) + 2, \quad N \geq 2. \tag{F.9}$$

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 [25, 28] provide a lower bound of $\Omega(N)$ 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 Fig. A.2; 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

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

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

$$\mu' = (1 - p_0 - p_1)\mu + p_1 \tag{G.2}$$

$$= (1 - p_0)\mu + p_1(1 - \mu), \tag{G.3}$$

$$\nu' = (1 - p_0)\nu + p_1(1 - \nu). \tag{G.4}$$

These values span a triangular region in the μ-ν plane that covers the majority of satisfiable constraints. Fig. G.1 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

$$\tilde{S}_2^{con2'} = p_n \tilde{S}_2^{con2} - \sum_{i \in \{0,1,n\}} p_i \log_2(p_i). \tag{G.5}$$

For given values of μ' and ν', the p_i are fixed and only the first term depends on N. This means that, 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 Fig. G.1).

Similarly, we can extend the range of validity for the construction described in Appendix **??** to the triangular region shown in Fig. A.2 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 (Eq. (F.9)), since the same number of states are active, but now they are not weighted equally, so this remains a low entropy distribution.

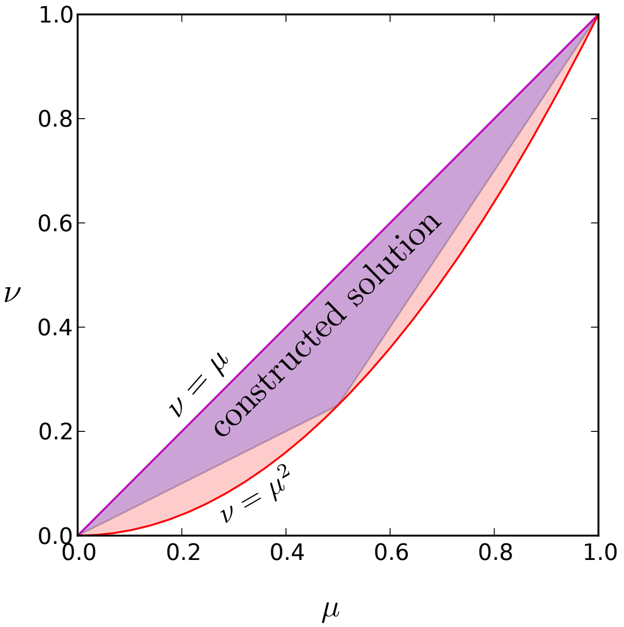

FIG. G.1. The full shaded region includes all allowed values for the constraints μ and ν for all possible probability distributions, replotted from Fig. A.2. The triangular blue shaded region includes all possible values for the constraints beginning with either of our constructed solutions with $\mu = 1/2$ and $\nu = 1/4$. Choosing other values of μ and ν for the construction described in Appendix E would move the vertex to any desired location on the $\nu = \mu^2$ boundary. Note that even with this solution alone, we can cover most of the allowed region.

APPENDIX H: PROOF OF THE LOWER BOUND ON ENTROPY FOR ANY DISTRIBUTION CONSISTENT WITH GIVEN μ & ν

Using the concavity of the entropy function, we can derive a *lower* bound on the minimum entropy. Our lower bound asymptotes to a constant except for the special case $\mu = 1/2$ and $\nu = 1/4$, which is especially relevant for communication systems since it matches the low order statistics of the maximum entropy solution.

We begin by re-expressing the entropy as follows:

$$S(\mathbf{p}) = -\sum_w p(w) \log_2 p(w) \tag{H.1}$$

$$= \sum_w \frac{p(w)^2}{p(w)} \log_2 \frac{1}{p(w)} \tag{H.2}$$

$$= \mathbf{p}^2 \sum_w \frac{p(w)^2}{\mathbf{p}^2} \frac{1}{p(w)} \log_2 \frac{1}{p(w)}, \tag{H.3}$$

where $\mathbf{p}$ represents the full vector of all 2^N state probabilities. Note that $p(w)^2/\mathbf{p}^2$ can be thought of as a probability distribution over w since its elements are non-negative and they sum to one. In this form, we can take advantage of the convexity of $x \log_2 x$ by using Jensen's inequality to obtain a *lower* bound on the entropy:

$$S(\mathbf{p}) \geq \mathbf{p}^2 \sum_w \left(\frac{p(w)^2}{\mathbf{p}^2} \frac{1}{p(w)} \right) \times \log_2 \sum_{w'} \left(\frac{p(w')^2}{\mathbf{p}^2} \frac{1}{p(w')} \right) \tag{H.4}$$

$$= \mathbf{p}^2 \sum_w \left(\frac{p(w)}{\mathbf{p}^2} \right) \log_2 \sum_{w'} \left(\frac{p(w')}{\mathbf{p}^2} \right) \tag{H.5}$$

$$= -\log_2 \mathbf{p}^2. \tag{H.6}$$

In the final step we use the fact that $p(w)$ is normalized.

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 binary column vectors, where i labels the state and each of the N components is either 1 or 0):

$$C = \left\langle \vec{s}\vec{s}^T \right\rangle \tag{H.7}$$

$$= \sum_i p(s_i) \vec{s}_i \vec{s}_i^T, \tag{H.8}$$

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}. Of course, in the symmetric problem we consider here, all diagonal entries are the same, and all off-diagonal entries are the same. We will take this to be the case from this point on.

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

$$\vec{\bar{s}} = 2\vec{s} - \vec{1}, \tag{H.9}$$

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

$$\bar{C} = \left\langle \vec{\bar{s}}\vec{\bar{s}}^T \right\rangle \tag{H.10}$$

$$= \left\langle (2\vec{s} - \vec{1})(2\vec{s}^T - \vec{1}^T) \right\rangle \tag{H.11}$$

$$= 4 \left\langle \vec{s}\vec{s}^T \right\rangle - 2 \left\langle \vec{s}\vec{1}^T \right\rangle - 2 \left\langle \vec{1}\vec{s}^T \right\rangle + \vec{1}\vec{1}^T, \tag{H.12}$$

$$\bar{c}_{mn} = 4c_{mn} - 2c_{mm} - 2c_{nn} + 1. \tag{H.13}$$

For our symmetric case, this reduces to

$$\bar{c}_{mm} = 1, \tag{H.14}$$

$$\bar{c}_{mn} = 4(\nu - \mu) + 1, \quad m \neq n. \tag{H.15}$$

Returning to Eq. (H.8) to find an upper bound on $\mathbf{p}^2$, we take the square of the Frobenius norm of $\bar{C}$:

$$\|\bar{C}\|_F^2 = \mathrm{Tr}(\bar{C}^T\bar{C}) \tag{H.16}$$

$$\begin{aligned} &= \mathrm{Tr}\left(\left(\sum_i p(\vec{\bar{s}}_i)\vec{\bar{s}}_i\vec{\bar{s}}_i^T\right)\right. \\ &\quad \times \left.\left(\sum_j p(\vec{\bar{s}}_j)\vec{\bar{s}}_j\vec{\bar{s}}_j^T\right)\right) \end{aligned} \tag{H.17}$$

$$= \mathrm{Tr}\left(\sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\vec{\bar{s}}_i\vec{\bar{s}}_i^T\vec{\bar{s}}_j\vec{\bar{s}}_j^T\right) \tag{H.18}$$

$$= \sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\,\mathrm{Tr}\left(\vec{\bar{s}}_i\vec{\bar{s}}_i^T\vec{\bar{s}}_j\vec{\bar{s}}_j^T\right) \tag{H.19}$$

$$= \sum_{i,j} p(\vec{\bar{s}}_i)p(\vec{\bar{s}}_j)\left(\vec{\bar{s}}_i\cdot\vec{\bar{s}}_j\right)^2 \tag{H.20}$$

$$\geq \sum_i p(\vec{\bar{s}}_i)^2\left(\vec{\bar{s}}_i\cdot\vec{\bar{s}}_i\right)^2 \tag{H.21}$$

$$\geq N^2\mathbf{p}^2. \tag{H.22}$$

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 $\mathbf{p}^2$:

$$\frac{\|\bar{C}\|_F^2}{N^2} \geq \mathbf{p}^2. \tag{H.23}$$

Using Eqs. (H.16), (H.14), and (H.15), we can express $\|\bar{C}\|_F^2$ in terms of μ and ν:

$$\|\bar{C}\|_F^2 = \sum_m \bar{c}_{mm}^2 + \sum_{m\neq n}\bar{c}_{mn}^2 \tag{H.24}$$

$$= N + N(N-1)\left(4(\nu-\mu)+1\right)^2. \tag{H.25}$$

Combining this result with Eqs. (H.23) and (H.6), we obtain a lower bound for the entropy for any distribution consistent with any given pair of values for μ and ν:

$$S(\mathbf{p}) \geq \tilde{S}_2^{lo} = \log_2\left(\frac{N}{1+(N-1)\alpha(\mu,\nu)}\right), \tag{H.26}$$

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

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

$$\lim_{N\to\infty}\tilde{S}_2^{lo} = \log_2(1/\alpha) \tag{H.27}$$

unless $\mu = 1/2$ and $\nu = 1/4$, in which case

$$\tilde{S}_2^{lo} = \log_2(N). \tag{H.28}$$

[1] R. Pathria, *Statistical Mechanics* (Butterworth Heinemann, 1972).

[2] W. Russ, D. Lowery, P. Mishra, M. Yaffe, and R. Ranganathan, Nature **437**, 579 (2005).

[3] M. Socolich, S. Lockless, W. Russ, H. Lee, K. Gardner, and et al., Nature **437**, 512 (2005).

[4] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc Nat'l Acad Sci (USA) **107**, 5405 (2010).

[5] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).

[6] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky, J. Neurosci. **26**, 8254 (2006).

[7] G. Tkacik, E. Schneidman, M. Berry, and W. Bialek, Arxiv preprint q-bio (2006).

[8] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. Smith, and et al., J Neurosci **28**, 505 (2008).

[9] M. Bethge and P. Berens, in *Advances in Neural Information Processing Systems*, Vol. 20, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Cambridge, MA: MIT Press, 2008) pp. 97–104.

[10] S. Yu, D. Huang, W. Singer, and D. Nikolic, Cereb Cortex **18**, 2891 (2008).

[11] J. Shlens, G. D. Field, J. L. Gauthier, M. Greschner, A. Sher, A. M. Litke, and E. J. Chichilnisky, Journal of Neuroscience **29**, 5022 (2009).

[12] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. Walczak, arXiv.org:1107.0604 [physics.bio-ph] (2011).

[13] E. Ganmor, R. Segev, and E. Schneidman, Proc Natl Acad Sci USA **108**, 9679 (2011).

[14] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[15] Y. Roudi, S. H. Nirenberg, and P. E. Latham, PLoS Computational Biology , 5:e1000380 (2009).

[16] S. H. Nirenberg and J. D. Victor, Curr Opin Neurobiol. **17(4)**, 397 (2007).

[17] F. Azhar and W. Bialek, arXiv.org:1012.5987 [q-bio.NC] (2010).

[18] P. Diaconis, Synthese **36** (1977).

[19] C. Shannon, Bell Syst. Tech. J **27**, 379 (1948).

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (1991).

[21] B. De Schutter, Journal of Computational and Applied Mathematics **121**, 331 (2000).

[22] J. Carter and M. Wegman, Journal of computer and system sciences **18**, 143 (1979).

[23] M. Sipser, in *Proceedings of the fifteenth annual ACM symposium on Theory of computing* (ACM, 1983) pp. 330–335.

[24] L. Stockmeyer, in *Proceedings of the fifteenth annual ACM symposium on Theory of computing* (ACM, 1983) pp. 118–126.

[25] B. Chor, O. Goldreich, J. Hasted, J. Freidmann, S. Rudich, and R. Smolensky, in *Foundations of Computer Science, 1985., 26th Annual Symposium on* (IEEE, 1985) pp. 396–407.

[26] R. Karp and A. Wigderson, Journal of the ACM (JACM) **32**, 762 (1985).

[27] M. Luby, SIAM J. Comput. **15**, 1036 (1986).

[28] N. Alon, L. Babai, and A. Itai, Journal of algorithms **7**, 567 (1986).

[29] W. Alexi, B. Chor, O. Goldreich, and C. Schnorr, SIAM J. Comput. **17**, 194 (1988).

[30] B. Chor and O. Goldreich, Journal of Complexity **5**, 96 (1989).

[31] B. Berger and J. Rompel, Journal of the ACM (JACM)

38, 1026 (1991).

[32] L. Schulman, in *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing* (ACM, 1992) pp. 17–25.

[33] M. Luby, Journal of Computer and System Sciences **47**, 250 (1993).

[34] R. Motwani, J. Naor, and M. Naor, Journal of Computer and System Sciences **49**, 478 (1994).

[35] D. Koller and N. Megiddo, in *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing* (ACM, 1993) pp. 268–277.

[36] H. Karloff and Y. Mansour, in *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing* (ACM, 1994) pp. 564–573.

[37] K. H. Fischer and J. A. Hertz, *Spin Glasses* (Cambridge University Press, 1991).

[38] T. Tanaka, Physical Review Letters E (1998).

[39] G. E. Hinton, S. Osindero, and Y.-W. Teh, Neural Computation **18**, 1527 (2006).

[40] A. Hyvärinen, IEEE Transactions on Neural Networks (2007).

[41] T. Broderick, M. Dudík, G. Tkačik, R. Schapire, and W. Bialek, E-print arXiv (2007).

[42] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, Phys Rev Lett **107(22)**, 220601 (2011).

[43] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, NY, USA, 2004).

[44] J. B. Rosen, Mathematics of Operations Research **8**, 215 (1983).

[45] D. L. Donoho, Stanford Technical Report (2004).

[46] E. J. Candés, in *Proceedings of the International Congress of Mathematicians* (Madrid, Spain: European Mathematical Society, 2006).

[47] S. Sarvotham, D. Baron, and R. G. Baraniuk, in *ECE Publications* (Rice University, 2006).

[48] M. Luby, M. Luby, and A. Wigderson, *Pairwise independence and derandomization*, Vol. 4 (Now Publishers Inc, 2006).

[49] A. Joffe, the Annals of Probability **2**, 161 (1974).

[50] F. MacWilliams and N. Sloane, *Error correcting codes* (North Holland, New York, 1977).

[51] A. Hedayat, N. Sloane, and J. Stufken, *Orthogonal arrays: theory and applications* (Springer Verlag, 1999).

[52] M. Hall and C. I. of Technology, *Combinatorial theory* (Wiley Online Library, 1967).

[53] H. Lancaster, The Annals of Mathematical Statistics **36**, 1313 (1965).

[54] D. Gale, H. W. Kuhn, and A. W. Tucker, Activity Analysis of Production and Allocation , 317 (1951).