Title
Bounds on the Entropy of a Binary System with Known Mean and Pairwise Constraints

Author
Albanna, Badr Faisal

Publication Date
2013

Peer reviewed|Thesis/dissertation

Bounds on the Entropy of a Binary System with Known Mean and Pairwise Constraints

by

Badr Faisal Albanna

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael R. DeWeese, Chair
Professor Ahmet Yildiz
Professor David Presti

Fall 2013

Bounds on the Entropy of a Binary System with Known Mean and Pairwise Constraints

Abstract

Bounds on the Entropy of a Binary System with Known Mean and Pairwise Constraints

by

Badr Faisal Albanna

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Michael R. DeWeese, Chair

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. In this dissertation, I provide lower and upper bounds on the entropy for both the minimum and maximum entropy distributions over binary units with any fixed set of mean values and pairwise correlations, and we construct distributions for several relevant cases. Surprisingly, the minimum entropy solution has entropy scaling logarithmically with system size, unlike the possible linear behavior of the maximum entropy solution, for any set of first- and second-order statistics consistent with arbitrarily large systems.

I find the following bounds on the maximum and minimum entropies for fixed values of $\{\mu_i\}$ and $\{\nu_{ij}\}$. For the *maximum entropy*:

$$xN - \mathcal{O}(\log_2 N) \leq S_2 \leq N.$$

In the case of uniform constraints, $x = 4(\mu - \nu)$ if $\nu \geq 1/2\,\mu$ and $\nu \geq 3/2\,\mu - 1/2$; otherwise $x = \frac{\nu - \mu^2}{1/4 - (\mu - \nu)}$. For the *minimum entropy*:

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

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 comparable to the scale of the maximum entropy.

I also demonstrate that some sets of low-order statistics can only be realized by small systems. My results show how only small amounts of randomness are needed to mimic low-order statistical properties of highly entropic distributions, and I discuss some applications for engineered and biological information transmission systems.

To Mama, Baba, and Bud,

Words cannot express the debt of gratiude I have for all the love and support you have given me over these many years. I hope you take this dissertation as a small tribute to all that you have taught me.

With great love and affection,
Your Son & Brother

To my partner in life, science, and crime, Michèle,

We did it!

Love,
Your Partner

Contents

List of Figures

Acknowledgments

Countless thanks go to my advisor, Mike DeWeese, for all his help in making this dissertation a reality and being a model for the kind of academic I would like to be. I hope I am able to maintain a fraction of the love and curiosity for the world that you have. Now that the dissertation is done, next stop: solving the brain!

I'd also like to thank my other collaborators on the work presented in this dissertation, Jascha Sohl-Dickstein and Christopher Hillar. In addition to being collaborators on this research they have contributed many hours to improving this text. If this work is lucid to any readers out there, it is they and Mike who should get the credit. In addition, I am so grateful for my years working with everyone at the Redwood Center. The Redwood Center is as stimulating and collaborative a research environment as I could have hoped for, and I feel fortunate to have spent my graduate years there.

I could not have produced this work without the support of the community of the Berkeley Compass Project. There are far too many people involved to thank each one here, but I would like to acknowledge the unwavering support of our faculty advisor, Bernard Sadoulet, over the years. Without his help we could not have grown into the vibrant community we are today. I am also very grateful to Nicci Nunes, former director of CalTEACH, for allowing me to share my love of physics and physics education with a group of excited undergraduate students.

Finally, a huge thank you to the staff of the Physics Department for their help over the years and all the work they do to make the Physics Department what it is. A special shout out to Anne Takizawa and Donna Sakima for being the first people to welcome me to Berkeley when I was visiting graduate school. On that first day, I literally could not have found my way to Berkeley without you, and I could not have found my way out without your help on every day since.

Chapter 1

Introduction

Now is an exciting time to a be a physicist entering the neuroscience community. Neuroscience has made enormous strides in the past half century from understanding relatively simple neural systems (*C. elegans*) to the extremely complex (a primate brain) at levels spanning the individual protein to the entire nervous system. This understanding has come with enormous advances in experimental technique and theoretical ideas, but today we stand at a precipice that may radically push our understanding in new and unexpected directions. For the first time, we may soon be able to record the spiking activity from entire networks or sub-networks of neurons as they operate. If this activity is the sole characteristic that determines the larger scale function of these networks (as we currently believe), we will finally be able to begin the task of determining how the brain "works" as a whole.

This view of neural networks imagines them as a system for distributed computation. Each neuron acts as an individual information processing unit that integrates its inputs to create a binary output (spike or no spike) that will in turn act as the input for other neurons. In this view, any operation of the brain that we would identify at a macroscopic scale – a perception, decision, thought, etc. – would correspond to the binary activity of enormous numbers of neurons each issuing their binary signal in concert. Indeed, despite the tremendous progress of localizing many brain functions in specialized regions, these regions are still quite large when compared to the number of neurons we can measure individually. This means that up until now, we have only been able to speculate how large neural networks may work together to create the functions and subjective experiences that the brain seems evolved to create. Appropriately, much work in theoretical neuroscience has focused on predicting the response properties of individual neurons or small collections of neurons. The mesoscopic picture bridging these two scales remains mostly a mystery.

Since the 1960s, our ability to measure larger and larger groups of neurons simultaneously has obeyed a Moore's law type behavior. Instead of the two years required for computing power to double, the number of neurons accessible to our experiments has doubled every seven years [1]. These changes have been facilitated by a host of advances in probe design and recording technology as well as the computational tools needed to isolate individual cells using a set of recorded channels. State of the art technology currently affords us the ability

to measure simultaneously in the low hundreds of cells and if these trends continue (a näive estimate), we would expect to record from an entire human brain in a few short centuries [1]. Of course, the transition from viewing neurons as discrete units to participants in a larger concert of activity will likely happen at scales far smaller than the entire human brain. In 1957, V.B. Mountcastle observed that small regions spanning the depth of cortex seemed to react in concert to particular stimuli. He dubbed these structures "cortical columns" and proposed that they may represent a basic functional unit for cortex [2]. These observations have since been strengthened by findings that cortex primarily develops by neuronal migration radially outward from the depths of cortex to the most superficial layers [3–5]. These cortical columns are small - on the order of half a millimeter by half a millimeter and spanning the whole depth of cortex (a few millimeters). We will soon be in a position to examine these developmentally, anatomically, and functionally defined structures as an entire network, and we may be able to derive rules that conclusively explain its function as a unit.

Along with the growth in electrophysiological recordings has come a host of recent techniques for mapping the complete anatomical connectivity of a neural system confined within a given volume. These methods are broadly categorized as finding the "connectome" of a neural network – a term that can refer to many different scales of observed functional connectivity, but here I take as the complete "wiring diagram" for a set of neurons. Methods employing Single Slice Electron Microscopy (SSEM) have now been used to map each and every neuron and synapse in small regions (10^{-2} mm^3 or about 1000 cell bodies) of the fly visual system [6], mouse retina [7, 8], and mouse visual cortex [9]. These maps were created after functional imaging allowing the authors to connect the connectivity of individual neurons to their observed receptive field properties. Such work holds out hope that we may not only understand how connectivity results in individual neuron activity, but that we may begin to ask questions about how entire networks shape themselves to act productively as a whole. It also makes it possible to imagine the discovery of neural wiring principles which may be applied even when such detailed connectomes are not available.

If we accept that the power of these techniques will continue to grow, we are quickly faced with a daunting set of theoretical questions.

- What are the appropriate set of larger scale "cognitive variables" that we should attempt to link to underlying neural activity?
- How do we determine which features in the underlying activity connect to said variables?
- What are the dynamical principles that govern this neural system in time and how do these dynamics relate to the observed "dynamics" of the cognitive variables of the brain?

The first two issues are no different then those for any other complex system where activity crosses scales (such as the global climate). However, unlike the global climate there

is a clear sense in which complex biological systems have *functions* and *purposes* that underlie their behavior (this issue is discussed in [10]); moreover, the behavior that accompany these volitional activities are accompanied by subjective experience (Chalmer's so-called "hard problem" [11]). This may complicate the pursuit of the final issue as now we are faced with difficult philosophical issues in addition to the formidable mathematical and technical challenges.

While one may be inclined to brush these issues aside and pursue a purely physical understanding of biological systems as machines, I think this view is mistaken. The volitional and subjective aspects of (at least some) biological systems is clearly important to the very idea of what makes them distinct from non-biological systems. Moreover, in practice these aspects provide fodder for thinking about how we should theorize a biological system. Across neuroscience, our own subjective experience is a constant source of intuition and inspiration in pursuing experiments and constructing models of neural network function. Although these intuitive clues must be tempered and submitted to experimental definition and verification, they are often the seed for ideas about which "cognitive variables" we should consider. Although this thesis does not explore these philosophical issues, these thoughts do underly my outlook on neural systems, and my choice to explore uses of statistical mechanics to think about these systems.

1.1 One way forward

Statistical mechanics is the physical theory which connects macroscopic quantities to underlying microscopic states, as such, it is poised to help. Although it remains to be seen what connection there is (if any) between statistical mechanics and the volitional and subjective qualities of neural systems, its ability to connect macroscopic behavior to microscopic dynamics and its connections to information theory make it a good place to start for those of us trained in the physical approach to neural systems (although there are hints of deeper connections as in [12, 13]). However, in applying the tools of statistical mechanics to neural systems we must get our feet wet by (1) demonstrating that it is capable of accurately describing underlying neural states using only a few macroscopic observables, (2) showing that there are good reasons to use statistical mechanics in a setting where many of the assumptions of statistical mechanics may not hold, and (3) showing that the applications of statistical mechanics will not trivially succeed.

Recent research has already demonstrated the applicability of the maximum entropy models used in statistical mechanics (hereafter referred to as *MaxEnt*) in a wide variety of biological settings from protein folding [14, 15], antibody diversity [16], and neural population activity [17–21]. In each of these cases, MaxEnt has done quite well as measured by a number of different metrics (*cf.*, [22]). In part due to this success, these types of models have also been used to infer functional connectivity in complex neural circuits [23] and to model collective phenomena of systems of organisms, such as flock behavior [24].

This broad application of MaxEnt 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 such models have been justified in terms of imposing no structure beyond what is explicitly measured as we have already discussed [17, 25]. I explore this point in the next chapter by discussing how Shannon's work in information theory allowed for a more general understanding of the canonical ensemble in physics. In addition to explaining Jaynes' original formulation of the MaxEnt principle, I present a general overview of the various critiques of MaxEnt in order to ground the reader in the challenges of applying statistical mechanics outside of its original domain. I conclude the second chapter by presenting my own views on how to properly apply MaxEnt which ultimately supports Jaynes' original proposal in the limit of large data.

Interestingly, in all of these studies mean and pairwise statistics form the basis of the chosen model just as in the case of the familiar Ising spin-glass model from physics. The reason for the success of MaxEnt models utilizing only pairwise statistics is unknown although there are recent results which show these MaxEnt distributions include distributions which maximize a specific definition of the multi-information [26], defined as the difference between the sum of the entropy of individual variables and the total entropy or

$$I_N(X_1, \ldots, X_N) \equiv \sum_i^N H(X_i) - H(X_1, \ldots, X_N). \tag{1.1}$$

This hints that MaxEnt involving only pairwise interactions may be sufficient to describe maximizing other relevant informational quantities; However, the precise reason why this would ensure a level of universality for these distributions remains a very open question.

It is currently not clear whether the success of MaxEnt is simply a product of the restrictions on the model provided by the experiment or whether the MaxEnt procedure plays an essential role. In other words, choosing a set of constraints restricts the set of consistent probability distributions generally — the MaxEnt solution being only one of all possible consistent probability distributions. If the space of distributions were sufficiently constrained by observations, then agreement would be an unavoidable consequence of the constraints rather than a consequence of the unique suitability of the MaxEnt model for the data set in question. Consequently in the field of systems neuroscience, understanding the range of allowed entropies for given constraints is an area of active interest. There is controversy [17, 27–30] over the notion that small pairwise correlations can conspire to constrain the behavior of large neural ensembles, and it has been shown [27, 28] that pairwise models do not always allow accurate extrapolation from small populations to large ensembles. Recent work [31] has also examined specific classes of biologically-plausible neural models whose entropy grows linearly with system size. These authors point out that entropy can be sub-extensive, at least for one special distribution originally constructed when entropy was a new concept [32]. Understanding the possible scaling properties of the entropy in a more general setting is of particular importance to neuroscience because of the entropy's interpretation as a measure of the amount of information a neural system could potentially relay to downstream neurons.

Previous authors have studied these issues with MaxEnt models expanded to second- [17], third- [28], and fourth-order [30]. In this dissertation, I 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 construct explicit low and high entropy models for the full range of possible uniform first- and second-order constraints (Eqs. (3.1)-(3.7); Figs. 3.1, 7.1). Interestingly, I 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 MaxEnt models in biology [14–21, 23, 24], when it occurs for large enough systems [28], represents a real triumph of the MaxEnt approach.

My results also have relevance for engineered information transmission systems. I 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 [33], do guarantee information transmission near channel capacity [34, 35], 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 [36] – a form of optimal model selection based on an interpretation of Occam's Razor complementary to that of Jaynes [25]. Intuitively, maximum entropy models impose no structure beyond that needed to fit the measured properties of a system, whereas minimum entropy models in a sense require the fewest "moving parts" in order to fit the data. In addition, my results have implications for computer science as algorithms for generating binary random variables with low entropy have found many applications (*e.g.*, [37–51]).

Chapter 2

Information theory as a bridge from statistical mechanics to neuroscience

It is difficult to understate the impact that Claude Shannon's articulation of the new field of information theory in 1948. The work of Shannon and those that have followed in his footsteps have formed the foundation of an age moved by information technology. Although some of the ideas of the yet-to-be-born theory were hinted at in earlier works of Nyquist, Hartley, and Gabor [52], Shannon's work gave these ideas clarity, precision, and most importantly he made the results general and compelling. It is worth noting that the "bit" and "byte" are perhaps two of the the most popularly recognizable units falling just behind the ubiquitous units of length, time, and mass formulated centuries earlier.

In one fell swoop, "A mathematical theory of communication," both formulated a clear conceptual framework for information and solved the key problems that this framework suggested. As others have pointed out, the key insight was to see that to quantify information, we needed less rather than more. The semantic character of a particular message had to be separated from the fact that ultimately a message - any message - is as much a statement about what is not as what is. Given an observer's uncertainty about a fixed set of choices (as quantified by a probability distribution over those choices p_i), Shannon showed that the observer should quantify their uncertainty about such a source using the *entropy*

$$H[\mathbf{p}] = -\sum_i p_i \log_2 p_i \tag{2.1}$$

Originally, Shannon did not recognize the connection between his measure of information and the H-function of Boltzmann and the entropy of Gibbs despite the identical expressions. It was John von Neumann who pointed out the similarity and suggested Shannon name his quantity "entropy" as well. "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage" [53]. Despite the decades between us and them, von Neumann's observation still rings true; the connections

between thermodynamic entropy, statistical mechanical entropy, and information theoretic entropy remain opaque despite the opening of a number of fruitful avenues. After all these years, Shannon still has the advantage.

For any physicist familiar with Shannon's work, the correspondence between his expression for the informational entropy and the familiar entropy of Boltzmann and Gibbs is irresistible (as the von Neumann quote suggests). Information theory furnishes the hope that the paradoxes and difficulties of interpretation that have surrounded statistical mechanics since its inception may have a satisfactory resolution when viewed in a new light. But if information theory is to make good on its promise it is clear that a radical epistemological shift is necessary. The entropy of Clausius is a firmly physical quantity derived from spontaneous energy transfer. It is a function of the physical system just like any other state functions, and can be measured and manipulated in a lab by an otherwise detached experimenter. However, the entropy of Shannon is more subjective. It is usually interpreted as the proper measure of an observer's uncertainty about a random variable. Viewed in this way it is a quantity measuring something about the experimenter – not the physical system. A statistical mechanics built from Shannon's entropy then seems to be a theory about an observer trying to understand an experiment with access to limited information rather than simply a multi-body extension of dynamics as the kinetic theory of Maxwell or the ergodic theory of Boltzmann. While this approach may be difficult to swallow for some, it allows us to use the tools developed for understanding ideal gases and solid state systems to understand any system about which we have limited information.

In this section, I will briefly summarize the development of statistical mechanics as a theory of inference by contrasting it to other ergodic formulations of the theory, and then describing Jaynes' MaxEnt proposal in detail. I will then summarize various critiques to MaxEnt and provide a modest proposal for an application of MaxEnt that neutralizes many of these critiques while mostly leaving Jaynes' formulation intact. This section will provide the foundation for discussing applications of statistical mechanics outside of physics before delving into my own specific results.

2.1 Two roads to statistical mechanics

First it is useful to have a simplistic version of the standard accounts of statistical mechanics as a basis for comparison. This summary draws on the work of both the inventors of these fields and researchers who have spent a lifetime teasing apart the intent of those inventors, but it is not in any way meant to be a comprehensive account. Essentially, I will caricature the early origins of statistical mechanics by separating it into philosophical threads that would presage later work. The first thread is the *ergodic* approach to statistical mechanics which tries to explain the properties of equilibrium systems by understanding the mathematical properties of the underlying dynamical system. This thread begins the kinetic theory of Maxwell and passes through the work of Kolmogrov, Sinai, and Arnold, and Moser (to reference a few highlights) [54]. Alternatively, the *inferential* approach to

statistical mechanics attempts to explain the results of statistical mechanics not from the "ground up" but instead as a rational result of reasoning in the face of limited information.

It's important to understand that very often both of these threads are present in any particular body of work to greater or lesser degrees. Indeed, Boltzmann can properly be situated as a key figure in establishing both lines of thinking [55, 56], and even Jaynes - the champion of the inferential approach often had kind words to say about the insights of ergodic theory. Therefore, the two approaches should not be seen as mutually exclusive but as complementary ways of viewing the same theory. The inferential approach can be seen as providing a lens to understand what is essential in the ergodic formulation; meanwhile, the ergodic approach can often be used to understand the bounds of applicability of inferential methods. The key for our purposes is that the inferential method is philosophically portable to new fields where ergodic methods are currently unjustifiable - namely neuroscience.

What precisely do I mean by the *ergodic* and *inferential* approaches to statistical mechanics? The ergodic approach is the main thread in most undergraduate approaches to statistical mechanics. Imagine we have a classical finite system with a large number of degrees of freedom. We use a Hamiltonian description of the system on its phase space. The *microstates* of the system are simply the individual configurations of the system represented by points in the phase space. The *macrostates* of the system are collections of microstates (regions in phase space) and are intended to represent regions which appear in some way identical to an external observer of the system. Typically, the macrostates are defined by functions which map each microstate to a real value; in this scheme all microstates that take the same value are considered a part of the same macrostate. Of course, the energy is most often used in this way to define macrostates of fixed energy.

Now for the central question in statistical mechanics: If we only have access to a set of macrostates for a given system, how should we predict average values of other, unobserved, functions of the microstates? The approach of Boltzmann and Gibbs was to define an *ensemble* which is a collection of microstates consistent given observed macrostate along with a probability distribution over these microstates. Using this probability distribution, we can compute the ensemble average of a new function to make our prediction. But of course, how should we pick this probability distribution? The ergodic approach has a seemingly natural answer: pick the distribution that is left invariant by the dynamics. But in general there is no unique solution to this proposition. We must further specify properties of the dynamics in order to reach a result; specifically, we require that the system is *ergodic* - that the long-term time averages of a system starting at a point in phase space are equal to the average which weights each microstate equally (the so called *microcanonical* ensemble). The Birkhoff-Khinchin theorem [57] establishes that the microcanonical distribution is essentially unique for systems if and only if the system is ergodic. Any other ensembles which are also invariant have zero measure in the phase plane and are considered "unphysical" leaving the microcanonical ensemble as the only reasonable choice.

The Boltzmann distribution (or *canonical* ensemble) is then derived by dividing an isolated system into two subsystems that can exchange energy: a large *reservoir* and a much smaller system of interest. The entire isolated system is assumed to be ergodic and there-

fore be best described by the microcanonical ensemble. How then should we weight the microstates of the smaller system of interest? The answer now is simple, the relative probability of a microstate of the smaller system is proportional to the number of microstates of the reservoir that are consistent with the microstate of the smaller system. In the limit of a small system (now we are in a position to say exactly what this means: a system with a small fraction of the total energy) the resulting distribution is the familiar *Boltzmann distribution*

$$p_x = \frac{\mathrm{e}^{-E_x/\tau}}{\mathcal{Z}} \tag{2.2}$$

There are a number of common critiques of this approach which the inferential line of thinking addresses. This derivation of the canonical ensemble relies on the use of the microcanonical ensemble for the entire system which in turn assumes ergodicity. The set of circumstances under which systems are ergodic is an ongoing matter of research in its own right and very few systems have been proved to have this property [58, 59]. Moreover, the Kolmogorov-Arnold-Moser Theorem shows that under limited non-linear perturbations many trajectories of integrable systems will continue to confine their motion to a fraction of the constant energy hyper-surface which means that for those trajectories ergodicity is violated. Therefore, if we want to argue that ergodicity is in some sense generic, we must be able to state the conditions for ergodicity under certain perturbations, but as of now such a comprehensive picture is lacking. Moreover, the ergodic justification in no way explains why non-equilibrium systems would evolve towards an ergodic trajectories. This fact was recognized by Gibbs himself [59], and in modern parlance we would say that we need a system that exhibits *mixing* if we want ergodicity to be a generic long-term outcome for a system. Given the success of statistical mechanics in describing a wide variety of physical systems, these arguments are not arguments against statistical mechanics per se, but they should convince us that our confidence in any particular application of the canonical ensemble are ultimately grounded in the method's past success and experimental verification in the current instance rather than an air-tight mathematical foundation.

Another important observation about the ergodic approach relates to the reservoir itself. First, we note that the above result is only exactly true in the limit where the amount of energy contained in the reservoir is infinite. For any real reservoir the Boltzmann distribution is an approximation. More fundamentally, in this derivation the Boltzmann distribution can only be seen as valid in the presence of the reservoir - after all, it merely reflects the degeneracy present in the reservoir for a given state of the smaller system. However, for many of the applications of statistical mechanics outside of physics no such reservoir is present. In the cases, the success of the ensemble cannot be attributed to the reasoning above.

Enter the inferential approach to statistical mechanics. In this approach we do not attempt to justify the tools of statistical mechanics in terms of the dynamics of a special class of systems, rather, we view statistical mechanics as a set of tools for reasoning about a system with a large number of degrees of freedom in the face of limited information. Just as in the ergodic approach we wish to find a probabilistic ensemble over the microstates of the system given what we know about the macrostates; however, unlike the ergodic approach

this connection is not justified in terms of a distribution selected by the dynamics, but by a method of inference limited by incomplete data. Often this reasoning is mentioned in parallel with ergodic reasoning above, and it is simple to state: We must use the microcanonical ensemble simply because we have no information that would allow us to weight one microstate as any more probable than any other. Here the microcanonical ensemble is nothing more than an extension of Laplace's "principle of insufficient reason" which states that one should weight all possible outcomes of a random variable equally in the absence of other data. The obvious complaint here is that now we have no reason to accept the results produced by this ensemble, but of course the ergodic argument provides little more cover - both methods are interlopers until exposed to the light of empirical measurement. In the next section we will continue to unspool this thread by explaining how Jaynes extended such reasoning to reproduce all the machinery of statistical mechanics - not simply the microcanonical distribution.

2.2 Jaynes' maximum entropy principle (MaxEnt)

In 1957, Edwin Thompson Jaynes published two papers in *Physical Review* that provided a radically different foundation for statistical mechanics [60, 61]. In these papers, Jaynes attempted to take the inferential thread outlined above seriously, stating that "[i]f one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum entropy principle." [60] Just as Bayseian statistics tried to reinterpret probabilities as quantities reflecting an optimal observers beliefs about a set of uncertain outcomes, Jaynes reinterpreted the canonical and microcanonical ensembles as outcomes of a process of inference from the data.

The MaxEnt algorithm is deceptively easy to describe. We are a set of experimenters interested in understanding the average properties of a system with known structure. By structure I mean that we can enumerate the N possible states of the system $\mathcal{S}$ (for simplicity we assume here that the sumber of states is finite) and we know the relationship between these states and any observable quantities of interest ($f_i(x)$ of the state $x \in \mathcal{S}$). In the end, we want to know the probability of finding the system in a state $p(x)$ given some set of observations about the systems. How should we proceed?

First it is important to note that Jaynes' method does not operate any possible set of observations. Generally speaking, we could measure a time series of some subset of our observable quantites where measurements are taken at some regular interval

$$(f_1(i_0), f_2(i_0), \dots), (f_1(i_1), f_2(i_1), \dots), \dots,$$

where i_T is the state of the system at time step T. Assuming that the probability of finding the system in a given state does not change from moment to moment, we could imagine some inference process that translates these observations into a set of probabilities - perhaps even a distribution of probabilities that captures our uncertainty about our estimates. Jaynes'

MaxEnt, however, only operates on a specific statistic of these observed quantities, namely their average values,

$$\bar{f}_1, \bar{f}_2, \ldots.$$

where $\bar{\cdot}$ is used to denote averages over experimental data. After taking our measurements, all that matters is are the mean values of the quantities of interest - we can throw the rest of the data away as far as MaxEnt is concerned. This may seem like a drastic (and perhaps negligent) reduction of the system, but it is important to remember that *any* function of the state can be included as an observable including the indicator function which counts the number of instances of a certain outcome.

Now we proceed with a seemingly simple premise that links these observations to our desired distribution. This rule can be thought of as playing the role of the *ergodic hypothesis* in the ergodic approach. Using $\langle\cdot\rangle$ to denote ensemble averages (as opposed to averages over measured quantities),

$$\langle f_1 \rangle \equiv \sum_{x \in \mathcal{S}} p_x f_i(x).$$

Jaynes suggests that we require that the average values of our observed quantities produced by our desired probability distribution $\mathbf{p}$ match the experimentally measured values,

$$\bar{f}_1 = \langle f_1 \rangle \,, \;\; \bar{f}_2 = \langle f_2 \rangle \,, \;\; \ldots. \tag{2.3}$$

In [62] this rule is referred to as the *constraint rule*, and I will use the same convention here.

The constraint rule restricts the allowed probabilities to a convex set of probabilities which agree with the measured statistics of the system, but there are still a potentially infinite number of probability distributions within this restricted space. This is where Jaynes invokes his updated version of the "Principle of Insufficient Reason". In Laplace's version of this principle, if one has no information about the likelihood of any given outcome of a random system, one should assign equal probability to all of the possible outcomes. In other words,

$$p_x = \frac{1}{N} \quad \forall x, \tag{2.4}$$

if there are N possible outcomes. One could justify this choice by saying that it maximizes our uncertainty about the possible outcomes. With each state having equal probability we cannot claim that any outcome is even the least bit more likely than any other.

But if we restrict ourselves to the distributions allowed by the constraint rule, the probability distribution given by 2.4 may not be allowed, and we need a more sophisticated definition of what it means to "maximize our uncertainty". Luckily, Shannon provides an answer; if we allow our measure of uncertainty to be Shannon's entropy function then the above argument implies that we should maximize

$$H[\mathbf{p}] = -\sum_{x \in \mathcal{S}} p_x \log_2(p_x),$$

given

$$\bar{f} = \langle f \rangle .$$

For simplicity, I have restricted our analysis to one observed quantity, but the argument is trivially generalizable if one wishes to measure more than one. Because the entropy functional is globally convex and we restrict the maximization to a convex subset of the probability space, this procedure yields a single maxima which is our distribution on interest. If there are no constraints, it is easy to show that this method reproduces the distribution in Eq. 2.4.

We can analytically solve for the form for this solution; moreover, when we choose the energy as our constraint function the distribution that this method returns is the familiar Boltzmann distribution. Maximizing the entropy subject to the constraint rule is accomplished by introducing the Lagrange multiplier β and maximizing the appropriately augmented functional $\mathcal{H}$ over the probability vector $\mathbf{p} = \{p_x\}$. Of course in addition to the constraint on the means we must also ensure that the probability vector is normalized,

$$\mathcal{H}[\mathbf{p}, \beta, \lambda] = -\sum_{x \in \mathcal{S}} p_x \log_2(p_x) + \beta \left(\bar{f} - \sum_{x \in \mathcal{S}} p_x f(x) \right) + \lambda \left(1 - \sum_{x \in \mathcal{S}} p_x \right).$$

Finding

$$\underset{\mathbf{p}}{\operatorname{argmax}}\, \mathcal{H}[\mathbf{p}] \equiv \mathbf{p}^*,$$

is then equivalent to finding the $\mathbf{p}^*$ that satisfies

$$\frac{\partial \mathcal{H}}{\partial p_x} = 0 \quad \forall p_x, \tag{2.5}$$

and is subject to

$$\frac{\partial \mathcal{H}}{\partial \beta} = 0,$$
$$\frac{\partial \mathcal{H}}{\partial \lambda} = 0,$$

which simply correspond to our constraints

$$\bar{f} = \langle f \rangle , \tag{2.6}$$
$$\sum_{x \in \mathcal{S}} p_x = 1. \tag{2.7}$$

Eq. 2.5 gives us the relation,

$$-1 - \log_2(p_x^*) - \beta f(x) - \lambda = 0. \tag{2.8}$$

Solving for p_x we find

$$p_x^* = \exp(-(1 + \lambda) - \beta f(x)). \tag{2.9}$$

Eqs. 2.6 and 2.7 are satisfied by appropriately selecting β and λ respectively. Alternatively, we can normalize the distribution by ignoring the λ term in Eq. 2.9 and simply expressing p_x as

$$p_x^* = \frac{\exp(-\beta f(x))}{\mathcal{Z}}, \tag{2.10}$$

with $\mathcal{Z}$ defined as the *partition function*

$$\mathcal{Z} \equiv \sum_{x \in \mathcal{S}} \exp(-\beta f(x)). \tag{2.11}$$

Of course, setting the observed equality to be the energy we recover the familiar Boltzmann distribution with the Lagrange multiplier now interpreted as the inverse temperature for the system.

Here we have recovered the microcanonical ensemble without reference to ergodicity, the microscopic details or the system, or (significantly) reference to a reservoir of any kind. The relative probabilities we describe are no longer no be interpreted as the multiplicities of a larger system that we ignore. Instead, we justify the probabilities as a *rational* choice given the available data (or at least a particular statistic of the data). The conclusion to Jaynes' 1957 paper does not shy away form this interpretation:

> The essential point in the arguments presented above is that we accept the von-Neumann-Shannon expression for entropy, very literally, as the measure of the amount of uncertainty represented by a probability distribution; thus entropy becomes the primitive concept which we work, more fundamental even than energy. If in addition we interpret the prediction problem of statistical mechanics in the subjective sense, we can derive the usual relations in a very elementary way without any consideration of ensembles or appeal to the usual arguments concerting ergodicity or equal *a priori* probabilities. The principles and mathematical methods of statsitical mechanics are seen to be of much more general applicability than conventional arguments would lead one to suppose. In the problem of prediction, the maximization of entropy is not an application of a lae of physics, but merely a method of resasoning which ensures that no unconscious arbitrary assumptions have been introduced. [60]

It is this interpretation of the canonical ensemble which allows us to take the tools of statistical mechanics to a very different realm where many of the familiar assumptions may no longer hold. We have the freedom to try; whether or not the prediction succeeds is a matter for experiment to decide.

2.3 Statistical errors

This argument takes statistical mechanics out of its familiar place as a physical theory and places it squarely in the domain of statistical learning. It is in this arena that it faced

much of its early criticism. In this section, I will outline some of the major critiques of the principle and the rebuttals that attempt to shore up its validity. I hope to demonstrate to the reader that despite these critiques, MaxEnt remains a powerful and useful method with a broad applicability.

Before proceeding it is useful for us to have in mind a model system when discussing the various critiques of Jaynes' theory. The example most often used is the so-called "Laplace dice problem," and I will continue that tradition here. The problem is simple: We are presented with a die with each of the faces numbered with consecutive integers. Usually the traditional six sided die is considered, but we will further simplify the situation by considering a three-sided die with sides marked one, two, and three. We do not know whether the die we have in our hand is "honest" or whether it is "rigged"; in other words, we assume that the relative number of face counts will converge to a set of values with probability approaching 1 as number of rolls becomes infinite, but we do not know what those relative numbers are. This is a very frequentist way of phrasing the situation, but I believe this is the closest conceptually to what we mean by the terminology of "rigged" and "honest" since these adjectives apply to the die itself and not the obersver of that die. For a die where each roll is independent of all others, we can show that if we wish to optimally assign probabilities (the the Baysian dutch-book sense) to each of the outcomes for the next roll, we would be best served by interpreting the long term count ratios as relative probabilities [63, 64]. Therefore, an "honest" die would have equal probabilities for each of the out comes whereas a "rigged" die might have the number three with a higher probability than the numbers one and two. Our task is to assign probabilities to the outcome of the next roll given some finite number of previous rolls where the outcome is known.

Following the procedure outlined in the previous section the result is easy to state. Using the average value of a side $\bar{d}$ as our constraint, using Eq. 2.10 we obtain that the probability of finding a die with side d is

$$p_d = \frac{\mathrm{e}^{-\beta d}}{\sum_{i=1}^{3} \mathrm{e}^{-\beta i}}. \tag{2.12}$$

β is chosen so that $\langle d \rangle = \bar{d}$. The relationship between $\langle d \rangle$ and β is plotted in Fig. 2.1

Obviously, there are many ways of doing this of which the MaxEnt proceedure is just one. The most common Baysian approach to the problem is to consider the three probabilities themselves as random variables with probabilities attached to each. Known as the *rule of sucession* and advocated by Laplace, De Fenetti, and Carnap [65, 66]. Bayes rule is combined with an assumption of *exchangability* [33] (which states that if each trial is independant then the joint probability of a finite set of outcomes should be equal to any other probability where the same outcomes occur in a different order) in order to predict the most likely probabilities. Using this method the most likely probability can be shown to be

$$p_d = \frac{n_d + 1}{n + 3},$$

where n_d is the number of times side d has appeared and n is the total number of rolls observed [67–69]. The generalization of this procedure is (so-called Carnap's so-called *continuum of*

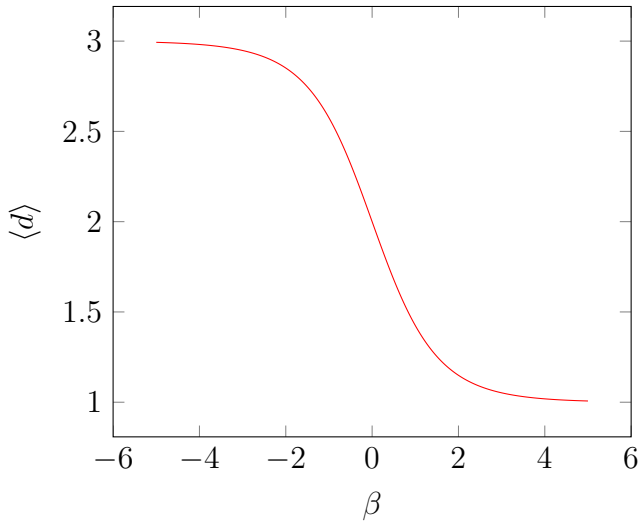

Figure 2.1: The relationship between the average dice roll versus the MaxEnt langrangian multiplier β.

inference) is also easy to state.

$$p_d = \frac{n_d + \alpha_d}{n + \alpha}$$
$$\alpha \equiv \sum_{i=1}^{3} \alpha_i,$$

In fact, this solution can be thought of simply as determining probabilities by averaging the observed counts with initial prior expectations about the various outcomes [69]. This becomes obvious if we rewrite the expression as

$$p_d = \left(\frac{n}{n+\alpha}\right)\frac{n_d}{n} + \left(\frac{\alpha}{n+\alpha}\right)\frac{\alpha_d}{\alpha}.$$

α and n are understood to weigh the contributions of prior knowledge and data respectively to our current probability estimates. The results produced by these methods are different than MaxEnt and these approaches can disagree even as the amount of data goes to infinity.

With our model dice in hand, we consider the first series of critiques and rebuttals that took place in the pages of the *Journal of Statistical Physics* from 1971 to 1973. In [70], Friedman and Shimoy attempted to show that the MaxEnt rule as formulated by Jaynes is in some sense inconsistent with the rules of probability as traditionally understood - specifically the rule for marginalizing over variables,

$$p(D|I) = \int_A p(D|AI)\, p(A|I). \tag{2.13}$$

Let us interpret the above formula by making the following associations: Take D the random variable describing the outcome of our die roll, I is the initial background information about our die (namely, that it has three sides and they produce the outcomes mentioned above), and finally A represents the average value of a set of observed die tosses (the key ingredient for MaxEnt). Note that there is a one-to-one correspondence between the averages a and the Lagrange multipliers β so we can rewrite Eq 2.13 as

$$p(D|I) = \int_B p(D|BI)\, p(B|I), \tag{2.14}$$

where B is the random variable over values of β.

Their argument notes that while MaxEnt clearly defines what is meant by $p(D|BI)$, it also seems to clearly indicate what $p(D|I)$ should be, and this puts perhaps unacceptable constraints on $p(B|I)$. After all, $p(D|I)$ is simply the probability of a die roll outcome given only the background information about the die. In this case, the MaxEnt prediction simply reduces to the principle of insufficient reason and we are given the prediction that all die outcomes are equally likely. This case also corresponds to the MaxEnt distribtion where $\beta = 0$ and the non-convexity of the one-parameter family of MaxEnt solutions in the space of all solutions ensures that this is the *only* distribution that can contribute to the RHS of Eq. 2.14. In other words, $p(B|I)$ must be a delta function at $\beta = 0$ which the authors claim is an unacceptable result.

> In other words, in this situation, a necessary condition for the consistency of [the above equations] is the inferrability with certainty from $[\beta]$ that evidence will be forthcoming which will fix the posterior expected value of $[A]$ to be $[a]$, the same as the prior expected value. Since one of the primary motivations of Jaynes's program is to find a probability assignment which "honestly describes what we know"..., he surely would not find this necessary condition acceptable. [70]

As [71–73] point out in their responses to this paper, there is nothing nefarious in the specificity of $p(B|I)$. Indeed, this is the proper result from a MaxEnt perspective. Without any knowledge of the die beyond the possible outcomes, we must weight each side equally which corresponds to setting $\beta = 0$. We can reinterpret their argument as saying that there is a missing conditioned variable in the original critique - namely, one that specifies the procedure for inference as MaxEnt. Without such information, the meaning of the two terms appearing in the integral on the RHS of Eq. 2.14 become ambiguous at best. Denoting this j in tribute to Jaynes, the conclusions of Eq. 2.14 become far less controversial. The marginalization condition

$$p(D|Ij) = \int_B p(D|BIj)\, p(B|Ij), \tag{2.15}$$

implies $p(B|Ij) = \delta(\beta)$ which is not a surprising result at all as the authors above argue.

The second substantial line of critique focuses on the crucial and seemingly innocuous constraint rule equating ensemble averages to measured averages. The primary problem with Eq. 2.3 is that it can lead to fairly radical conclusions when the sample size is small [62]. Taking our fictional three-sided die in hand and giving it a toss, if we were to find a three on our first roll, Jaynes' prescription would lead us to conclude that

$$(p_1, p_2, p_3) = (0, 0, 1),$$

because this is the only probability distribution consistent with the observation that $\bar{d} = 3$. By contrast, the *rule of succession* produces the result

$$(p_1, p_2, p_3) = (1/4, 1/4, 1/2).$$

This result appears to incorporate the new observation with our previous assumption of equal probabilities nicely.

The objection to this determination can be pulled into two pieces: (1) Weighting a side with probability 1 and (2) doing so with a finite set of data (in this case, only one sample). We might be tempted to see (1) as simply a consequence of choosing average face number as a our statistic. For example, if we had observed a two rather than a three than we would have found

$$(p_1, p_2, p_3) = (1/3, 1/3, 1/3),$$

because two happens to also be the average value if each side is weighted with equal probability (which is the overall maximum entropy state). Our problem then appears to be that we happened to see an extremal value on the die. Perhaps we should chose our function of the die face to respect the apparent symmetry of the die by choosing our constraint functions to be the indicator function for each side.

In that case the MaxEnt solution is trivial because the constraint rule in this case reduces to

$$\sum_{i=1}^{3} p_i \mathbb{I}(d) = \bar{\mathbb{I}}(d)$$

giving us the result

$$p_d = \frac{n_d}{n}.$$

Although this eliminated the asymmetry between the different outcomes it does not alleviate problem (1) at all. It seems when with improved symmetry we are still left without room for a prior.

Rather than attempt to fix this problem at this point in the story, it is helpful to touch on another question related to the problem above: What is the relationship between the type of "inference" performed by MaxEnt and the "inference" accomplished by Bayes' Theorem? This is an issue which has has been a matter of interest for some time, with a variety of opinions on offer (e.g. they are incompatible [74], they are both special cases of yet another inference principle [75], they agree for certain observables [76, 77], and they represent two distinct types of "inference" appropriate in distinct circumstances [78, 79]). Settling in on one of these stances is necessary to even entertain a solution to the problems above.

2.4 A modest proposal

To address these concerns, I will take what I consider to be a pragmatic approach:

1. MaxEnt is to be considered a model for the data like any other - it is a procedure for generating a probability distribution given certain model parameters.

2. In this case, the model parameters are the averages the chosen functions of the system state.

3. Instead of using the *constraint rule* to connect ensemble averages to data averages, we will use Bayes' Rule (defined below).

It is simple to show that given a reasonable choice of prior, these assumptions lead to Jaynes' constraint rule when the amount of data becomes large; Moreover, they allow for a learning process that is dependent on the data and smoothly progresses from a set of prior knowledge to a distribution consistent with Jaynes' constraint rule. Let's review Bayes rule. If we have a model with parameters θ and a prior distribution over the model parameters $p(\theta)$, Bayes' Rule gives us a way to generate a new prior that incorporates the knowledge obtained by observing the data D. Mathematically, it can be seen as nothing more than an application of the fact that $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$, but the implications are profound. So we have

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)},$$

where $p(D) = \int p(D|\theta)p(\theta)\, d\theta$.

Applying this to our example, we begin by assuming that our data comes in the form of the series of dice rolls (the result is unaffected if we only have access to the averages). For example we might find the series of rolls (d_i labels the ith roll, $\{d_i\}$ refers to the entire sequence.)

$$\{d_i\} = \{3, 2, 3, 3, 1, 3, 1, 2, 1, 2, 3, 3, \dots\}. \tag{2.16}$$

Our distribution for each roll is simply the MaxEnt distribution conditioned on the supposed average $\langle d \rangle$. The dependence on $\langle d \rangle$ comes through the choice of the Lagrange multiplier β so perhaps it is better to write $\beta(\langle d \rangle)$. The precise relationship is simply the inverse of the function plotted in Fig. 2.1. Since each roll is independent, $p(\{d_i\}|\, \langle d \rangle)$ simply becomes the

product of the MaxEnt distributions for each roll derived in Eq. 2.12

$$\begin{aligned} p(\{d_i\}|\langle d\rangle) &= \prod_{i=1}^{n} p(d_i|\beta(\langle d\rangle)) \\ &= \prod_{i=1}^{n} \frac{\mathrm{e}^{-\beta(\langle d\rangle)d_i}}{\mathcal{Z}} \\ &= \frac{\mathrm{e}^{-\beta(\langle d\rangle)\sum_{i=1}^{n} d_i}}{\mathcal{Z}^n} \\ &= \exp\left(-n(\beta\bar{d}_n + \ln(\mathcal{Z}))\right). \end{aligned}$$

Multiplying likelihood by the prior $p(\langle d\rangle)$ we find for the posterior $p(\langle d\rangle\,|\{d_i\})$

$$\begin{aligned} p(\langle d\rangle\,|\{d_n\}) &\propto p(\{d_i\}|\langle d\rangle)p(\langle d\rangle) \\ &= \exp\left(-n(\beta\bar{d}_n + \ln(\mathcal{Z})) + \ln(p(\langle d\rangle))\right). \end{aligned} \tag{2.17}$$

Assuming the $\ln(p(\langle d\rangle))$ is sufficiently smooth, we can find the peak of this distribution by determining when the derivative of argument is equal to zero (Remember, the normalization constant does not depend on $\langle d\rangle$). In the following, I will write p instead of $p(\langle d\rangle)$ to make the notation less cumbersome.

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\langle d\rangle}\left(-n(\beta\bar{d}_n + \ln(\mathcal{Z})) + \ln(p(\langle d\rangle)\right) = 0 \\ \Rightarrow\; &-n\frac{\mathrm{d}\beta}{\mathrm{d}\langle d\rangle}\left[\bar{d}_n + \frac{\mathrm{d}}{\mathrm{d}\beta}\ln(\mathcal{Z})) - \frac{1}{np}\frac{\mathrm{d}p}{\mathrm{d}\beta}\right] = 0. \end{aligned} \tag{2.18}$$

As we know from statistical mechanics,

$$\frac{\mathrm{d}\ln(\mathcal{Z})}{\mathrm{d}\beta} = -\langle d\rangle\,, \tag{2.19}$$

and using Eq. 2.19 in Eq. 2.18 we find that at the peak

$$\langle d\rangle = \bar{d}_i - \frac{1}{nP}\frac{\mathrm{d}p}{\mathrm{d}\beta}. \tag{2.20}$$

In the large data limit ($n \to \infty$), the part which depends on the initial prior disappears and the peak approaches

$$\bar{d}_\infty \equiv \bar{d} = \langle d\rangle\,. \tag{2.21}$$

Moreover, the distribution will become narrower as the amount of data points increases by the central limit theorem. In the infinite data limit, we then have a distribution that is sharply peaked around the true average (as demonstrated in Fig. 2.2).

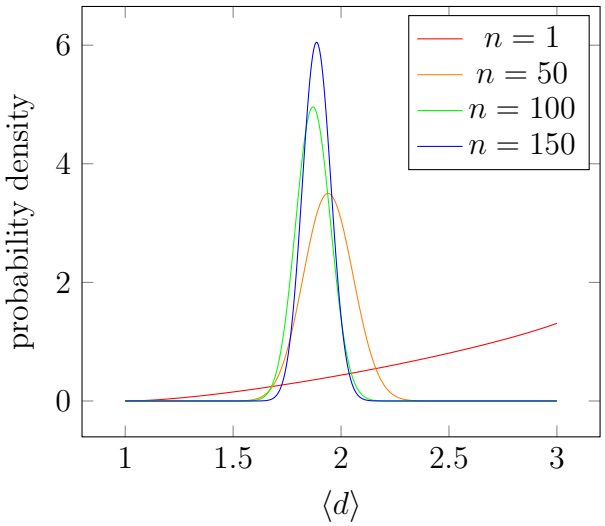

Figure 2.2: The probability of average face value as the amount of data n increases (using a flat prior). In this case the true die had equal probability on each side implying a long term average of 2. Actual averages for the samples used are $\bar{d}_1 = 3$, $\bar{d}_{50} = 1.94$, $\bar{d}_{100} = 1.87$, $\bar{d}_{150} = 1.89$. Note that the peak of each distribution matches the empirically measured average, and it approaches 2 as the empirical average approaches 2. Moreover, as the amount of data increases the distribution narrows around the true value.

How do we choose our prior $p(\langle d \rangle)$? Since almost all reasonable priors reproduce the *constraint rule* for large data samples I will not delve into the widely debated issue of how to choose a prior except to state a few of the common approaches. One natural choice is simply a flat prior that weights each average value equally or one that weights the logarithm of the average equally. MaxEnt itself can also provide a prior if we find the distribution $p(\langle d \rangle)$ that maximizes the entropy of the joint distribution $p(E, \langle d \rangle)$ where $p(d|\langle d \rangle)$ is constrained to have a mean value of $\langle d \rangle$. In this case, the prior can be shown to be

$$p(\langle d \rangle) \propto \exp(nH(\langle d \rangle)), \tag{2.22}$$

where n are the number of measurements that will be used to update the prior and $H(\langle d \rangle)$ is the entropy of the distribution $p(d|\langle d \rangle)$ for a given value of $\langle d \rangle$. This is perhaps not a surprising result because according to the Asymptotic Equipartition Theorem [35] this is equal to the number of typical sequences of length n drawn from $p(E|\langle d \rangle)$. In other words, this prior treats every possible sequence of length n equally and then assigns weights on $\langle d \rangle$ based on the fraction of those sequences that it typically produces.

Of course neither of these methods are invariant to reparameterizations of the average. Only Jeffrey's prior has that unique distinction because it is the Haar measure on the space of distributions with given parameters. In practice, it is proportional to the square root of the determinant of the fisher information. In our example, the Jeffreys prior can easily be

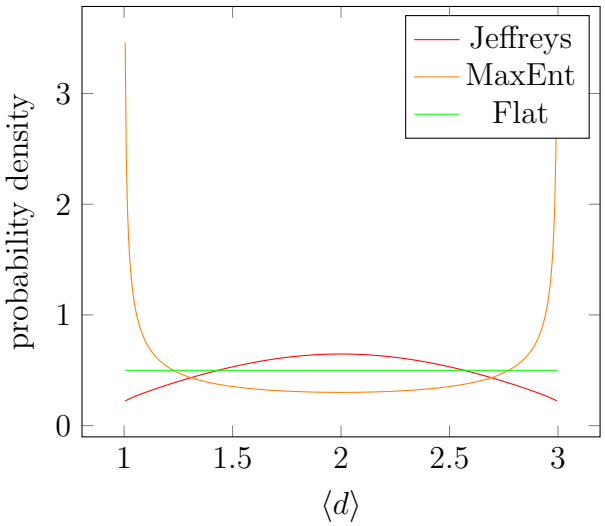

Figure 2.3: Various possible non-informative priors for the average side value. In the limit of large amounts of data, each of these is consistent with Jaynes' constraint rule.

shown to be

$$p(\langle d \rangle) = \left| \frac{\mathrm{d}\beta}{\mathrm{d}\,\langle d \rangle} \Delta d \right|,$$

where Δd is the standard deviation of the MaxEnt distribution with mean $\langle d \rangle$. In Fig. 2.3 we see how different each of these priors is for our dice example. The important takeaway for our purposes is that despite their major differences, each is consistent with Jaynes' constraint rule in the limit of large amounts of data.

Chapter 3

Summary of results

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 [80]. 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 [30, 81–85]. The maximum entropy distribution is not the only one consistent with these observed statistics, however. In fact, there are many others, and I 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 [17] 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 I 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. I find that the gap grows linearly with N, up to a logarithmic correction.

I find the following bounds on the maximum and minimum entropies for fixed values of

$\{\mu_i\}$ and $\{\nu_{ij}\}$. For the *maximum entropy*:

$$xN - \mathcal{O}(\log_2 N) \leq S_2 \leq N. \tag{3.1}$$

In the case of uniform constraints, $x = 4(\mu - \nu)$ if $\nu \geq 1/2\,\mu$ and $\nu \geq 3/2\,\mu - 1/2$; otherwise $x = \frac{\nu - \mu^2}{1/4 - (\mu - \nu)}$. For the *minimum entropy*:

$$\log_2\left(\frac{N}{1 + (N-1)\bar{\alpha}}\right) \leq \tilde{S}_2 \leq \log_2\left(1 + \frac{N(N+1)}{2}\right), \tag{3.2}$$

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 comparable to the scale of the maximum entropy.

I emphasize that these results are valid for arbitrary sets of mean firing rates and pairwise correlations, but for clarity I will often focus on the special class of distributions with *uniform constraints* — those with uniform mean firing rates and pairwise correlations:

$$\mu_i = \mu, \quad \text{for all } i = 1, \ldots, N \tag{3.3}$$

$$\nu_{ij} = \nu, \quad \text{for all } i \neq j. \tag{3.4}$$

For uniform constraints, Eq. (3.2) reduces to

$$\log_2\left(\frac{N}{1 + (N-1)\alpha}\right) \leq \tilde{S}_2 \leq \log_2\left(1 + \frac{N(N+1)}{2}\right), \tag{3.5}$$

where $\alpha(\mu, \nu) = (4(\nu - \mu) + 1)^2$. In most cases, the lower bound in Eq. (3.2) 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 tighter bound:

$$\log_2(N) \leq \tilde{S}_2 \leq \log_2(N) + 2. \tag{3.6}$$

An important class of probability distributions are the *exchangeable distributions* [33], 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. I have constructed a family of exchangeable distributions that I conjecture is a minimum entropy exchangeable solution with entropy $\tilde{S}_2^{exch}$. Our solution scales linearly with N:

$$C_1 N - \mathcal{O}(\log_2 N) \leq \tilde{S}_2^{exch} \leq C_2 N, \tag{3.7}$$

where C_1 and C_2 are constants that only depend on μ and ν. I have empirically confirmed that this is indeed a minimum entropy exchangeable solution for $N \leq 200$.

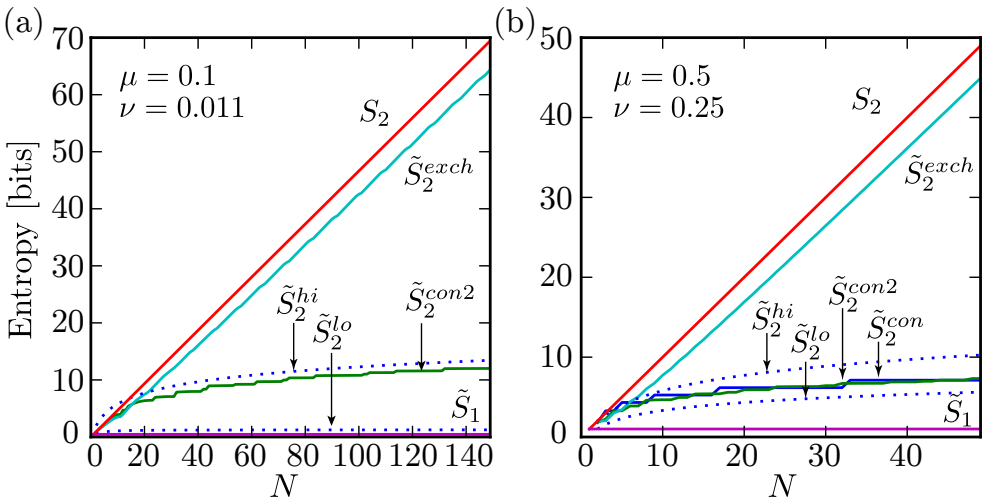

Figure 3.1: Minimum and maximum entropy for fixed uniform constraints as a function of N. The minimum entropy grows no faster than logarithmically with the system size N for any mean activity level μ and pairwise correlation strength ν. (a) In a parameter regime relevant for neural population activity in the retina [17, 18] ($\mu = 0.1$, $\nu = 0.011$), I 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; $\mu = 1/2$, $\nu = 1/4$), I 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.

Chapter 4

Allowed range of ν given μ across all distributions for large N

In this chapter I will only consider distributions satisfying uniform constraints

$$\mu_i = \mu, \quad \forall i = 1, \ldots, N \tag{4.1}$$

$$\nu_{ij} = \nu, \quad \forall i \neq j, \tag{4.2}$$

and I will show that

$$\mu^2 \leq \nu \leq \mu \tag{4.3}$$

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

I 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 I 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, $p_1 = \mu = \nu$. I use the term "active" to refer to neurons that are spiking, and thus equal to one, in a given time bin, and I also refer to "active" states in a distribution, which are those with non-zero probabilities.

I 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. I can find the lower bound by viewing this as a linear programming problem [86, 87], 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* [33], 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

$$p(100) = p(010) = p(001), \tag{4.4}$$
$$p(110) = p(101) = p(011). \tag{4.5}$$

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, I need only consider exchangeable distributions. If there exists a probability distribution that satisfies our constraints, I 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 I construct an exchangeable distribution $p_e(w)$ with the same constraints as follows:

$$p_e(\vec{s}) \equiv \sum_{\sigma} \frac{p(\sigma(\vec{s}))}{N!}, \tag{4.6}$$

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 chapter I 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

$$1 = \sum_{i=0}^{N} \binom{N}{i} p(i). \tag{4.7}$$

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:

$$\mu = \sum_{i=1}^{N} \binom{N-1}{i-1} p(i). \tag{4.8}$$

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:

$$\nu = \sum_{i=2}^{N} \binom{N-2}{i-2} p(i). \tag{4.9}$$

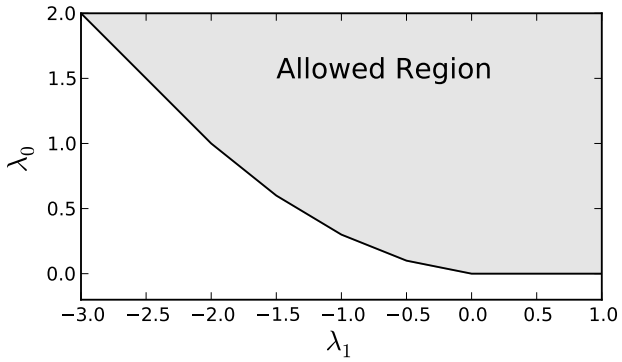

Figure 4.1: An example of the allowed values of λ_0 and λ_1 for the dual problem ($N = 5$).

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

$$-\nu = -\sum_{i=2}^{N} \binom{N-2}{i-2} p(i) \tag{4.10}$$

subject to

$$1 = \sum_{i=0}^{N} \binom{N}{i} p(i), \tag{4.11}$$

$$\mu = \sum_{i=1}^{N} \binom{N-1}{i-1} p(i), \tag{4.12}$$

$$p(i) \geq 0, \quad \text{for all } i. \tag{4.13}$$

This gives us the following dual problem: Minimize

$$\mathcal{E} \equiv \lambda_0 + \mu\lambda_1, \tag{4.14}$$

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

$$\binom{N}{i}\lambda_0 + \binom{N-1}{i-1}\lambda_1 \geq -\binom{N-2}{i-2}, \quad N \geq i \geq 0, \tag{4.15}$$

where $\binom{a}{b}$ is taken to be zero for $b < 0$. The principle of strong duality [87] 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 (4.1). 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. (4.15) 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 determine the optimal values of λ_0 and λ_1 (λ_0^* and λ_1^*, respectively) as a function of i_0

$$\binom{N}{i}\lambda_0^* + \binom{N-1}{i-1}\lambda_1^* = -\binom{N-2}{i_0-2} \tag{4.16}$$

$$\binom{N}{i_0+1}\lambda_0^* + \binom{N-1}{i_0}\lambda_1^* = -\binom{N-2}{i_0-1}. \tag{4.17}$$

Solving for λ_0^* and λ_1^*, we find

$$\lambda_0^* = \frac{i_0(i_0+1)}{N(N-1)} \tag{4.18}$$

$$\lambda_1^* = \frac{-2i_0}{(N-1)}. \tag{4.19}$$

Plugging this into Eq. (4.14) we find the optimal value $\mathcal{E}^*$ is

$$\mathcal{E}^* = \lambda_0^* + \mu\lambda_1^* \tag{4.20}$$

$$= \frac{i_0(i_0+1)}{N(N-1)} - \mu\frac{2i_0}{(N-1)} \tag{4.21}$$

$$= \frac{i_0(i_0+1-2\mu N)}{N(N-1)}. \tag{4.22}$$

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. (4.14) and Fig. 4.1 that at the solution, i_0 satisfies

$$-m(i_0) \leq \mu \leq -m(i_0+1), \tag{4.23}$$

where $m(i)$ is the slope, $d\lambda_0/d\lambda_1$, of constraint i. The expression for $m(i)$ is determined from Eq. (4.15),

$$m(i) = -\frac{\binom{N-1}{i-1}}{\binom{N}{i}} \tag{4.24}$$

$$= -\frac{i}{N}. \tag{4.25}$$

Substituting Eq. (4.25) into Eq. (4.23), we find

$$\frac{i_0}{N} \leq \mu \leq \frac{i_0+1}{N}. \tag{4.26}$$

This allows us to write

$$\mu = \frac{i_0 + b(N)}{N}. \tag{4.27}$$

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

$$i_0 = N\mu - b(N) \tag{4.28}$$

Substituting Eq. (4.28) into Eq. (4.22), we find

$$\mathcal{E}^* = \frac{(N\mu - b(N))(N\mu - b(N) + 1 - 2N\mu)}{N(N-1)} \tag{4.29}$$

$$= \frac{(N\mu - b(N))(-N\mu - b(N) + 1)}{N(N-1)} \tag{4.30}$$

$$= -\mu^2 + \mathcal{O}\left(\frac{1}{N}\right) \tag{4.31}$$

Taking the large N limit we find that $\mathcal{E}^* = -\mu^2$ and by the principle of strong duality [87] the maximum value of $-\nu$ is $-\mu^2$. Therefore I have shown that for large N, the region of satisfiable constraints is simply

$$\mu^2 \leq \nu \leq \mu, \tag{4.32}$$

as illustrated in Fig. 4.2.

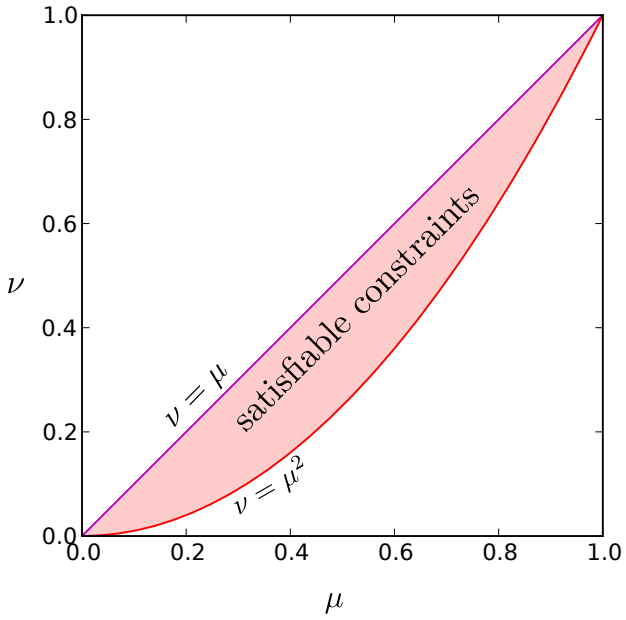

Figure 4.2: The red shaded region is the set of values for μ and ν that can be satisfied for at least one probability distribution in the $N \to \infty$ limit. The purple line along the diagonal where $\nu = \mu$ is the distribution for which only the all active and all inactive states have non-zero 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.

Chapter 5

Bounds on the maximum entropy

I 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. I 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

$$\mu_i = \langle s_i \rangle = \sum_{\vec{s}} p(\vec{s}) s_i, \quad i = 1, \ldots, N, \tag{5.1}$$

$$\nu_{ij} = \langle s_i s_j \rangle = \sum_{\vec{s}} p(\vec{s}) s_i s_j, \quad i \neq j, \tag{5.2}$$

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

$$\begin{aligned} \mathcal{S}(p(\vec{s})) = &- \sum_{\vec{s}} p(\vec{s}) \log_2 p(\vec{s}) \\ &- \sum_i \lambda_i \left(\sum_{\vec{s}} p(\vec{s}) s_i - \mu_i \right) \\ &- \sum_{i<j} \gamma_{ij} \left(\sum_{\vec{s}} p(\vec{s}) s_i s_j - \nu_{ij} \right). \end{aligned} \tag{5.3}$$

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

$$p(\vec{s}) = \frac{1}{\mathcal{Z}} \exp\left(- \sum_i \lambda_i s_i - \sum_{i<j} \gamma_{ij} s_i s_j \right), \tag{5.4}$$

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.

This "inverse" problem of inferring Lagrange multipliers (the interaction and magnetic field terms) in the exponent (Ising spin glass Hamiltonian) that produce the measured means and correlations is nontrivial, but there has been progress [30, 81–85].

Here I consider the symmetric case of uniform constraints to make the problem computationally tractable while deriving bounds that apply more generally. For uniform constraints the Lagrange multipliers are themselves uniform. In other words,

$$\lambda_i = \lambda, \quad \forall i, \tag{5.5}$$

$$\gamma_{ij} = \gamma, \quad \forall i < j. \tag{5.6}$$

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

$$p(\vec{s}) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda \sum_i s_i - \gamma \sum_{i<j} s_i s_j\right). \tag{5.7}$$

If there are k neurons active, this becomes

$$p(k) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda k - \gamma \frac{k(k-1)}{2}\right). \tag{5.8}$$

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

Using expression (5.8), we find the maximum entropy by using the `fsolve` function from the `SciPy` package of Python subject to constraints (5.1) and (5.2).

As Fig. 5.1 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 $\nu_{ij} = \mu_i \mu_j$. 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 values of $\{\mu_i\}$ and $\{\nu_{ij}\}$ that remain away from the boundary of allowed values in the infinite neuron limit by combining this independent solution with a low entropy solution. Let us call the independent distribution for a set of N neurons with mean firing rates of $\{\mu_i\}$ $\mathbf{p}_{ind}$. Moreover, let us designate a low entropy solution (with a support size of n_c or smaller) that has constraint values $\{\mu_i\}$ and $\{\tilde{\nu}_{ij}\}$ as $\mathbf{p}_{low}$.

Before I begin, I must specify our procedure for "growing" the size of the network. In order to speak about scaling behavior, we must be able to scale the network to infinite size - a non-trivial task given that there is a constrained range of allowed values for each statistic of the system given the others. I consider a procedure whereby I add a neuron to a finite sized system (of size n), by specifying its mean (μ_{n+1}) firing rate and the pairwsie firing rate with each of the preexisting neurons ($\nu_{i\,n+1}$). Of course, not all values are allowed, and specifying the range of values given the previous statistics is a difficult task. Here, I simply assert that such bounds exist (after all, we know there exist statistics which allow for scaling to infinite size).

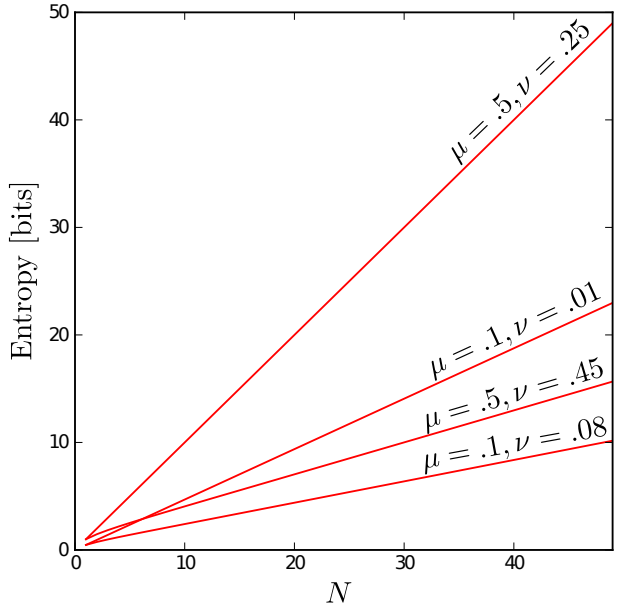

Figure 5.1: The maximum possible entropy scales linearly with system size, N, as shown here for various values of μ and ν. Note that this linear scaling holds even for large correlations.

Our goal is to construct a distribution which is a mixture of an maximum entropy distribution ($\mu_i^{max} = 1/2$, $\nu_{ij}^{max} = 1/4$) and a low entropy distribution ($\tilde{\mu}_i$, $\tilde{\nu}_{ij}$) such that this mixture has the same statistics as our distribution of interest. In other words, we wish to define the mixture as a one parameter family of distributions over x

$$\mathbf{p}_{mix} = x\mathbf{p}_{max} + (1-x)\mathbf{p}_{low}, \tag{5.9}$$

where $0 \leq x \leq 1$. This mixture will have constraint values equal to

$$\nu_{ij}^{mix} = 1/4\, x + (1-x)\tilde{\nu}_{ij}, \quad \forall i \neq j \tag{5.10}$$
$$\mu_i^{mix} = 1/2\, x + (1-x)\tilde{\mu}_i. \quad \forall i \tag{5.11}$$

If we have chosen $\mathbf{p}_{low}$ properly we can find an x such that such that

$$\nu_{ij}^{mix} = \nu_{ij}, \quad \forall i \neq j \tag{5.12}$$
$$\mu_i^{mix} = \mu_i. \quad \forall i \tag{5.13}$$

The difficulty is in choosing an x that remains independent of N as the system scales in size because as the system grows the range of allowed statistics for any finite set of statistics shrinks by necessity. Statistics which may have been achievable at smaller N may cease to be be so as the system grows in size. We may be tempted to simply adjust which statistics

we use as the "walls" close in, but this is unacceptable because doing so would change our x as a value of N. Therefore, we must be careful to choose the statistics of our low entropy distribution such that they will remain available as the system grows in size and x will remain a constant.

In other words, the statistics of our low entropy distribution must remain a finite distance from the boundary of allowed statistics even as system size goes to infinity. Let us make this more precise by imagining the space of allowed statistics $\{\mu_i\}$ and $\{\nu_{ij}\}$ values as a $d = N(N+1)/2$ dimensional euclidean space with coordinates given by each of the μ_i and ν_{ij} (I will refer to this space as $\mathcal{S}$). First I note that this space of allowed values is convex because the space of probability distributions that create these linear statistics is itself convex. If we have two allowed distributions labeled a and b with statistics $\{\mu_i^a\}$, $\{\nu_{ij}^a\}$ and $\{\mu_i^b\}$, $\{\nu_{ij}^b\}$ then

$$\mu_i^c = \lambda\mu_i^a + (1-\lambda)\mu_i^b, \tag{5.14}$$
$$\nu_{ij}^c = \lambda\nu_{ij}^a + (1-\lambda)\nu_{ij}^b, \tag{5.15}$$

are also allowed statistics for $0 \leq \lambda \leq 1$.

Recognizing that the pairwise constraint space is convex allows us to note that points on the interior of the space are also in the *relative interior* meaning if $\{\nu_{ij}\}$ is a set of statistics is this space there exists a $\lambda > 1$ such that

$$z(\lambda) = \lambda x + (1-\lambda)y \in \mathcal{S} \quad \forall y \in \mathcal{S}. \tag{5.16}$$

In plain language, this captures the idea that if you are not at boundary, you are still free to "move" in any direction without immediately leaving the space. In our case, we pick y to be the maximum entropy distribution, and select the largest value of λ (λ^*) such that $z(\lambda^*)$ remains an element of $\mathcal{S}$.

Imagine λ^* as a function of N; as we grow the system, λ^* can only decrease. If it were to increase when increasing the size of the system from N to $N+1$, we could always marginalize over the last neuron without affecting the statistics of the original N. This would imply that there is a λ larger than the previously determined λ^* for N neurons which is of course a contradiction.

Therefore, λ_n^* is a monotonically decreasing sequence bounded from below. Now we must restrict ourselves to networks where our distribution of interest stays a finite distance from the boundary. In other words, $\lim_{n\to\infty}\lambda_n^* = \lambda_\infty^* > 1$. In this case, we can use λ_∞^* to find the statistics we will use for $\{\nu_{ij}^{low}\}$ as the system size grows because we can be assured that

$$\bar{\mu}_i = \lambda_\infty^*\mu_i + (1-\lambda_\infty^*)\,1/2, \tag{5.17}$$
$$\bar{\nu}_{ij} = \lambda_\infty^*\nu_{ij} + (1-\lambda_\infty^*)\,1/4, \tag{5.18}$$

will be within the space of allowed statistics as $N \to \infty$. If we invert this to solve for ν_{ij},

$$\mu_i = x\,1/2 + (1-x)\bar{\nu}_{ij}, \tag{5.19}$$
$$\nu_{ij} = x\,1/4 + (1-x)\bar{\nu}_{ij}, \tag{5.20}$$

where $x = (\lambda^*_\infty - 1)/\lambda^*_\infty$.

We now use this x to create a distribution $\mathbf{p}_{mix}$ as in Eq. (5.9) where x does not depend on N, and $\mathbf{p}_{mix}$ satisfies Eqs. (5.12) and (5.13). The entropy of $\mathbf{p}_{mix}$ can now act as a lower bound for the maximum entropy with identical statistics,

$$S_2 \geq S(\mathbf{p}_{mix}), \tag{5.21}$$

and because entropy is a concave function [35], by Jensen's inequality the entropy of $\mathbf{p}_{mix}$ is bounded below by

$$S[\mathbf{p}_{mix}] \geq xS[\mathbf{p}_{ind}] + (1-x)S[\mathbf{p}_{low}]. \tag{5.22}$$

The entropy of the first term is simply N so if x is constant in N and non-zero, the first term must grow linearly in N which means that the true maximum entropy must grow at least linearly with N as well.

In the case where the statistics are symmetric, we can apply this method generally to find x. In this case there are two cases (shown in Fig. 5.2): the first is depicted in blue when $\nu \geq 3/2\,\mu - 1/2$ and $\nu \geq 1/2\,\mu$. In this case the low entropy distribution is taken to be along $\nu = \mu$ and we find

$$x = 4(\mu - \nu). \tag{5.23}$$

Otherwise, we must pick a low-entropy distribution in the red region and find

$$x = \frac{\nu - \mu^2}{1/4 - (\mu - \nu)}. \tag{5.24}$$

I note that at this stage I know of no easy method for characterizing which growing systems will end up away from the boundary as described above, near a boundary in the infinite system limit, or on a boundary for some finite N. This is a matter of considerable interest for the authors as there are certainly cases where the maximum entropy scales sub-linearly as a function of N as well as possible cases where no scaling to infinity is possible. These latter categories can give radically different answers to the question of how much a set of mean and pairwise constraints also constrains the amount of entropy in a system.

Finally, with the lower bound in hand, I note that there is also 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. 5.1).

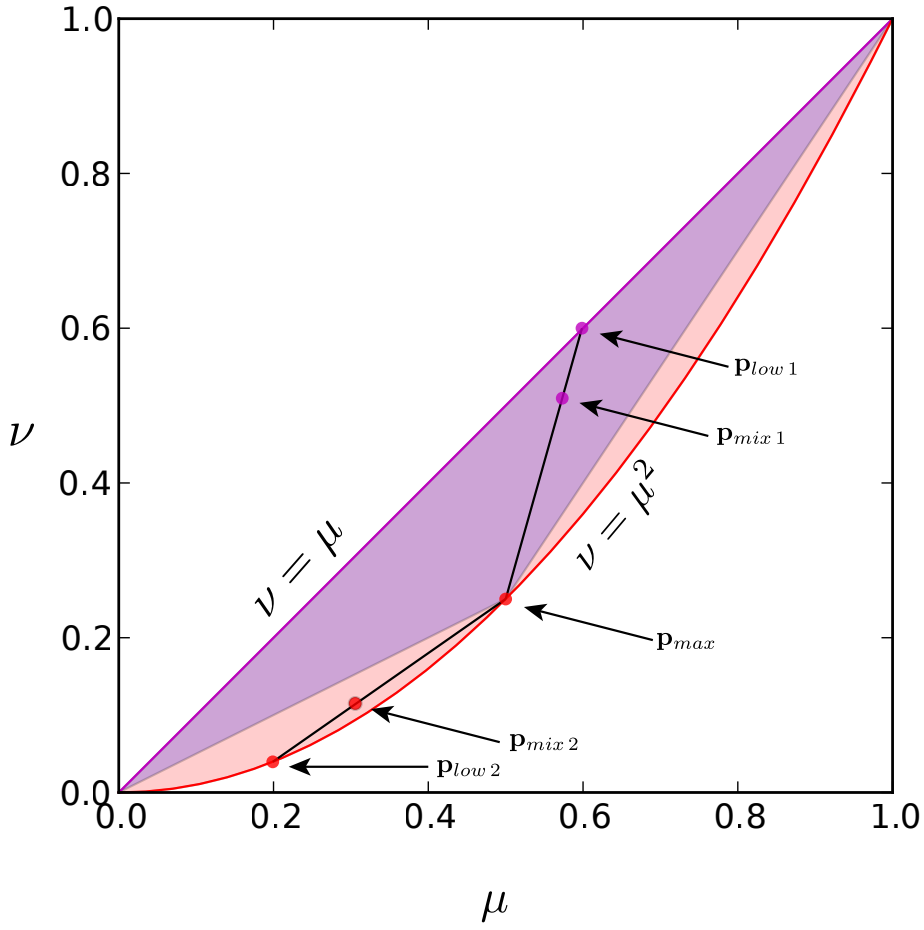

Figure 5.2: A plot showing the allowed values for μ and ν with the two cases for constructing maximum entropy lower bound mixture distributions overlaid using purple and red. One example distribution is shown in each region ($\mathbf{p}_{mix}$) along with the maximum entropy distribution ($\mathbf{p}_{max}$) and the low entropy distribution used ($\mathbf{p}_{low}$). Distributions in the purple region (subscript 1) use lwo-entropy distributions along $\nu = \mu$ ($\mathbf{p}_{low\,1}$) while distributions in the red region (subscript 2) use low-entropy distributions along $\nu = \mu^2$ ($\mathbf{p}_{loq\,2}$). The blue region is defined by the lines from independent statistics ($\mu = 1/2$, $\nu = 1/4$) to ($\mu = 0$, $\nu = 0$) and ($\mu = 1$, $\nu = 2$) giving us the constraints $\nu \geq 1/2\,\mu$ and $\nu \geq 3/2\,\mu - 1/2$ in a addition to $\nu < \mu$. The red region is what remains of the allowed space of statistics.

Chapter 6

Bounds on the minimum entropy

I obtain the bounds on the minimum entropy by 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 [87], at least for sufficiently small or symmetric systems. In Chapter 5 I will show that the maximum entropy S_2 for any system with specified mean and pairwise correlation scales linearly with N (Eq. (3.1), Fig. 3.1).

By contrast, we prove below that the minimum entropy distribution exists at a vertex of the allowed space of probabilities, where most states have probability zero [88]. 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 (Fig. 3.1), and we have constructed two families of explicit solutions with low entropies ($\tilde{S}_2^{con}$ and $\tilde{S}_2^{con2}$; Figs. 3.1,7.1) for a broad parameter regime covering all allowed values for μ and ν in the case of uniform constraints.

Our goal is to minimize the entropy function

$$S(\mathbf{p}) = \sum_{i=0}^{n_s} -p_i \log_2 p_i, \tag{6.1}$$

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, $n_s = 2^N$. The constraints for normalization, mean firing rates, and pairwise firing rates give

$$n_c = 1 + N + \frac{N(N-1)}{2}, \tag{6.2}$$

$$= 1 + \frac{N(N+1)}{2}. \tag{6.3}$$

In this chapter I will show that the minimum occurs at the verticies of the space of allowed probabilities. Moreover, these vertices correspond to probabilities of small support - specifically a support size equal to n_c. These two facts allow us to put an upper bound on the minimum entropy of

$$\tilde{S}_2 \leq \tilde{S}_2^{hi} \approx 2\log_2(N), \tag{6.4}$$

for large N.

I 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 simply 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 gives 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 soltuion space is a convex polytope of dimension d, which we refer to by $\mathcal{S}$. 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 $\mathcal{S}$ we may write

$$\mathbf{p} = \sum_i^{n_v} a_i \mathbf{v}_i, \tag{6.5}$$

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

Using the concavity of the entropy function, I 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, $\mathbf{v}_*$

$$S(\mathbf{p}) = S\left(\sum_i^{n_v} a_i \mathbf{v}_i\right) \tag{6.6}$$

$$\geq \sum_i^{v_s} a_i S(\mathbf{v}_i) \tag{6.7}$$

$$\geq \left(\sum_i^{v_s} a_i\right) S(\mathbf{v}_*) \tag{6.8}$$

$$= S(\mathbf{v}_*). \tag{6.9}$$

Therefore,

$$\tilde{S}_2 = S(\mathbf{v}_*) \tag{6.10}$$

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 [50]). 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.

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

$$\tilde{S}_2 \leq \tilde{S}_2^{hi} = \log_2(n_c) \tag{6.11}$$

$$= \log_2\left(1 + \frac{N(N+1)}{2}\right) \tag{6.12}$$

$$\approx 2\log_2(N), \tag{6.13}$$

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 $\{\nu_{ij}\}$.

Using the concavity of the logarithm function, we can also 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.

I begin by bounding the entropy from below as follows:

$$S(\mathbf{p}) = -\sum_w p(w)\log_2 p(w) \tag{6.14}$$

$$= \langle -\log_2 p(w)\rangle \tag{6.15}$$

$$\geq -\log_2 \langle p(w)\rangle \tag{6.16}$$

$$= -\log_2 \mathbf{p}^2, \tag{6.17}$$

where $\mathbf{p}$ represents the full vector of all 2^N state probabilities, and I 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 I seek an upper bound on $\mathbf{p}^2$. This can be obtained by starting with the matrix representation C of the constraints (for now, I 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):

$$C = \left\langle \vec{s}\vec{s}^T \right\rangle \tag{6.18}$$

$$= \sum_i p(s_i)\vec{s}_i\vec{s}_i^T, \tag{6.19}$$

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

$$\vec{\bar{s}} = 2\vec{s} - \vec{1}, \tag{6.20}$$

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

$$\bar{C} = \left\langle \vec{\bar{s}} \vec{\bar{s}}^T \right\rangle \tag{6.21}$$

$$= \left\langle (2\vec{s} - \vec{1})(2\vec{\bar{s}}^T - \vec{1}^T) \right\rangle \tag{6.22}$$

$$= 4 \left\langle \vec{s}\vec{\bar{s}}^T \right\rangle - 2 \left\langle \vec{s}\vec{1}^T \right\rangle - 2 \left\langle \vec{1}\vec{\bar{s}}^T \right\rangle + \vec{1}\vec{1}^T, \tag{6.23}$$

$$\bar{c}_{mn} = 4c_{mn} - 2c_{mm} - 2c_{nn} + 1. \tag{6.24}$$

This reduces to

$$\bar{c}_{mm} = 1, \tag{6.25}$$

$$\bar{c}_{mn} = 4\nu_{mn} - 2(\mu_m + \mu_n) + 1, \quad m \neq n. \tag{6.26}$$

Returning to Eq. (6.19) to find an upper bound on $\mathbf{p}^2$, we take the square of the Frobenius norm of $\bar{C}$:

$$\|\bar{C}\|_F^2 = \mathrm{Tr}(\bar{C}^T \bar{C}) \tag{6.27}$$

$$= \mathrm{Tr}\left(\left(\sum_i p(\vec{\bar{s}}_i) \vec{\bar{s}}_i \vec{\bar{s}}_i^T \right) \times \left(\sum_j p(\vec{\bar{s}}_j) \vec{\bar{s}}_j \vec{\bar{s}}_j^T \right) \right) \tag{6.28}$$

$$= \mathrm{Tr}\left(\sum_{i,j} p(\vec{\bar{s}}_i) p(\vec{\bar{s}}_j) \vec{\bar{s}}_i \vec{\bar{s}}_i^T \vec{\bar{s}}_j \vec{\bar{s}}_j^T \right) \tag{6.29}$$

$$= \sum_{i,j} p(\vec{\bar{s}}_i) p(\vec{\bar{s}}_j) \,\mathrm{Tr}\left(\vec{\bar{s}}_i \vec{\bar{s}}_i^T \vec{\bar{s}}_j \vec{\bar{s}}_j^T \right) \tag{6.30}$$

$$= \sum_{i,j} p(\vec{\bar{s}}_i) p(\vec{\bar{s}}_j) \,\mathrm{Tr}\left(\vec{\bar{s}}_j^T \vec{\bar{s}}_i \vec{\bar{s}}_i^T \vec{\bar{s}}_j \right) \tag{6.31}$$

$$= \sum_{i,j} p(\vec{\bar{s}}_i) p(\vec{\bar{s}}_j) \left(\vec{\bar{s}}_i \cdot \vec{\bar{s}}_j \right)^2 \tag{6.32}$$

$$\geq \sum_i p(\vec{\bar{s}}_i)^2 \left(\vec{\bar{s}}_i \cdot \vec{\bar{s}}_i \right)^2 \tag{6.33}$$

$$= N^2 \mathbf{p}^2. \tag{6.34}$$

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 $\mathbf{p}^2$:

$$\frac{\|\bar{C}\|_F^2}{N^2} \geq \mathbf{p}^2. \tag{6.35}$$

Using Eqs. (6.27), (6.25), and (6.26), we can express $\|\bar{C}\|_F^2$ in terms of μ and ν:

$$\|\bar{C}\|_F^2 = \sum_m \bar{c}_{mm}^2 + \sum_{m \neq n} \bar{c}_{mn}^2 \tag{6.36}$$

$$= N + (4\nu_{mn} - 2(\mu_m + \mu_n) + 1)^2. \tag{6.37}$$

Combining this result with Eqs. (6.35) and (6.17), we obtain a lower bound for the entropy for any distribution consistent with any given sets of values $\{\mu_i\}$ and $\{\nu_{ij}\}$:

$$S(\mathbf{p}) \geq \tilde{S}_2^{lo} = \log_2 \left(\frac{N^2}{N + \sum_{i \neq j} \alpha_{ij}} \right) \tag{6.38}$$

$$= \log_2 \left(\frac{N}{1 + (N-1)\bar{\alpha}} \right) \tag{6.39}$$

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

$$S(\mathbf{p}) \geq \tilde{S}_2^{lo} = \log_2 \left(\frac{N}{1 + (N-1)\alpha} \right), \tag{6.40}$$

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

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

$$\lim_{N \to \infty} \tilde{S}_2^{lo} = \log_2 (1/\bar{\alpha}). \tag{6.41}$$

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,

$$4\nu_{ij} - 2(\mu_i + \mu_j) + 1 = 0 \quad \forall i \neq j \tag{6.42}$$

But in the large N limit,

$$\mu_i \mu_j \leq \nu_{ij} \leq \min(\mu_i, \mu_j). \tag{6.43}$$

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

$$0 \leq \mu_i \leq \frac{1}{2} \tag{6.44}$$

$$\frac{1}{2} \leq \mu_j \leq \mu_i + \frac{1}{2}, \tag{6.45}$$

and

$$\nu_{ij} = \frac{\mu_i + \mu_j}{2} - \frac{1}{4}. \tag{6.46}$$

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 Eq. (6.46), 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,

$$\tilde{S}_2^{lo} = \log_2(N). \tag{6.47}$$

Although here the lower bound scales logarithmically with N, rather than as a constant, 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).

Chapter 7

Constructed low-entropy solutions

In addition to these bounds, we can also construct probability distributions between these minimum entropy boundaries for distributions with uniform constraints. These solutions provide concrete examples of distributions with qualitatively different scaling behavior than maximum entropy models. I include these so that the reader may gain a better intuition for what low entropy models look like in practice; These models are not intended as an improvement over maximum entropy models for a particular system. Nonetheless, these models are of practical importance to other fields as I discuss further in the conclusion.

Each of these solutions has an entropy that grows logarithmically with N (see Eqs. (3.2)-(3.6)):

$$\begin{aligned} \tilde{S}_2^{con2} &\leq \log_2\left(\lceil N\rceil_p\left(\lceil N\rceil_p - 1\right)\right) - 1 + \log_2(3) \\ &\leq \log_2\left(N(2N-1)\right) + \log_2(3), \qquad (7.1) \\ \tilde{S}_2^{con} &= \lceil \log_2(N) + 1\rceil \\ &\leq \log_2(N) + 2, \qquad (7.2) \end{aligned}$$

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. 3.1a, 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 ([17–21, 23, 27]; $\mu = 0.1$, $\nu = 0.011$), the range of possible entropies grows almost linearly with N, despite the highly symmetric constraints imposed (Eqs. (4.1) and (4.2)).

Consider the special case of first- and second-order constraints 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 I found for the neural regime (Fig. 3.1a). In fact, one can construct [31, 32] an explicit solution (Eq. (7.2); Figs. 3.1b,7.1a,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 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 while still remaining a distribution whose entropy grows only logarithmically with N.

7.1 First construction

I 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

$$\mu = \frac{n}{N}, \quad \nu = \frac{n(n-1)}{N(N-1)}, \tag{7.3}$$

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 Chapter 4). I 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 [89] as explained by Luby in [42].

I derive our construction by first assuming that N is a prime number, but this is not actually a limitation as I 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.

I 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:

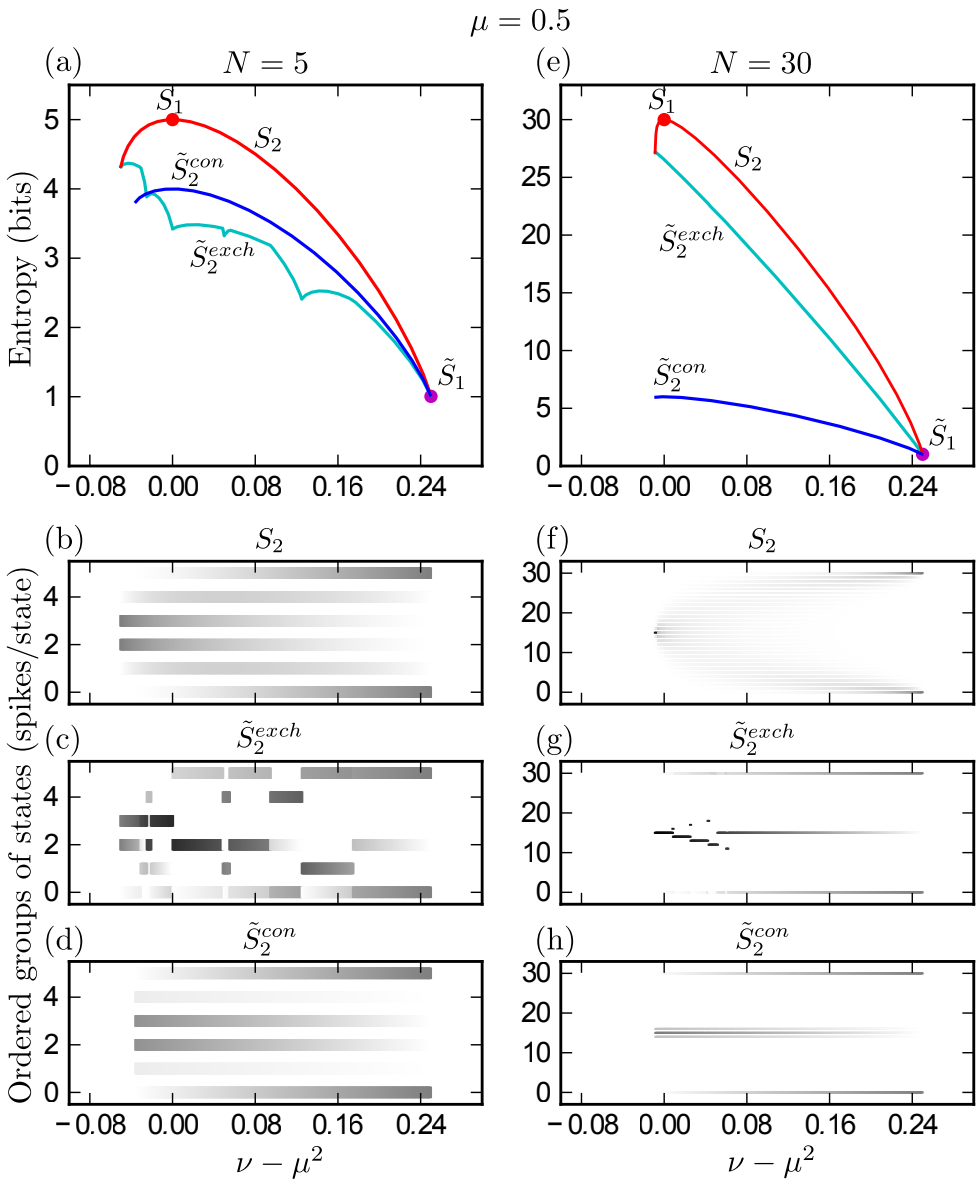

Figure 7.1: Minimum and maximum entropy models for uniform constraints. (a) Entropy as a function of the strength of pairwise correlations for the maximum entropy model (S_2), finitely exchangeable minimum entropy model ($\tilde{S}_2^{exch}$), and a constructed low entropy solution ($\tilde{S}_2^{con}$), 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.

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. (7.3)

Now I 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) I will neglect the issue of redundancy and consider the states produced through step 3 as distinct.

First I 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. I 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 $s = 1$):

$$l = (s \cdot i) \bmod N, \tag{7.4}$$

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,

$$(s \cdot i_1) \bmod N = (s \cdot i_2) \bmod N \tag{7.5}$$

$$\Rightarrow (s \cdot (i_1 - i_2)) \bmod N = 0. \tag{7.6}$$

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

$$l = (s \cdot i + d) \bmod N, \tag{7.7}$$

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.

I 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. (7.7) 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

$$\mu = \frac{n}{N}. \tag{7.8}$$

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

$$l_1 = (s \cdot i_1 + d) \bmod N, \tag{7.9}$$

$$l_2 = (s \cdot i_2 + d) \bmod N. \tag{7.10}$$

Taking the difference between these two equations, we find

$$\Delta l = (s \cdot (i_2 - i_1)) \bmod N. \tag{7.11}$$

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. (7.9) and (7.10) 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

$$\nu = \frac{n(n-1)}{N(N-1)}, \tag{7.12}$$

where N is prime.

Finally I 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 I 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

$$\tilde{S}_2^{con2} \leq \log_2 \left(\frac{N(N-1)}{2} \right), \tag{7.13}$$

where N is prime. As mentioned above, I 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.

I 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 $n > 1$:

$$\tilde{S}_2^{con2} \leq \log_2 \left(N(2N-1) \right), \tag{7.14}$$

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.

7.2 Second construction

In addition to the probability distribution described in the previous section, I 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 I 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 [32], and it was recently discussed in the context of neural modeling by Macke and colleagues [31].

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:

$$\begin{matrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{matrix} \tag{7.15}$$

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

$$\begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix} \tag{7.16}$$

and each 0 by

$$\begin{matrix} 0 & 0 \\ 1 & 0 \end{matrix} \tag{7.17}$$

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

$$\begin{matrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{matrix} \tag{7.18}$$

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. (7.16) and (7.17) that we obtain the following two relations,

$$\mu_1^{(q+1)} = \mu^{(q)}, \tag{7.19}$$

$$\mu_2^{(q+1)} = 1/2. \tag{7.20}$$

$$\tag{7.21}$$

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

$$\tilde{S}_2^{con} = \log_2(2N), \quad N = 2^q, \ q = 2, 3, 4, \ldots. \tag{7.22}$$

I 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."

I can extend Eq. (7.22) 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:

$$\tilde{S}_2^{con} = \lceil \log_2(2N) \rceil \leq \log_2(N) + 2, \quad N \geq 2. \tag{7.23}$$

By adding two other states I 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 I now describe.

I remark that the authors of [40, 43] 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.

7.3 Extending the range of validity for these constructions

I now show that each of these low entropy probability distributions can be generalized to cover much of the allowed region depicted in Fig. 4.2; in fact, the distribution derived in

Section 7.1 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

$$p_0 + p_n + p_1 = 1. \tag{7.24}$$

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

$$\mu' = (1 - p_0 - p_1)\mu + p_1 \tag{7.25}$$
$$= (1 - p_0)\mu + p_1(1 - \mu), \tag{7.26}$$
$$\nu' = (1 - p_0)\nu + p_1(1 - \nu). \tag{7.27}$$

These values span a triangular region in the μ-ν plane that covers the majority of satisfiable constraints. Fig. 7.2 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

$$\tilde{S}_2^{con2'} = p_n \tilde{S}_2^{con2} - \sum_{i \in \{0,1,n\}} p_i \log_2(p_i) \tag{7.28}$$

For given values of μ' and ν', the p_i are fixed and only the first term depends on N. Using Eqs. (7.25) and (7.27),

$$p_n = \frac{\mu' - \nu'}{\mu - \nu}. \tag{7.29}$$

This allows us to rewrite Eq. (7.28) as

$$\tilde{S}_2^{con2'} \leq \left(\frac{\mu' - \nu'}{\mu - \nu} \right) \tilde{S}_2^{con2} + \log_2(3). \tag{7.30}$$

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,

$$\tilde{S}_2^{con2'} \leq \tilde{S}_2^{con2} + \log_2(3). \tag{7.31}$$

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. 7.2).

Similarly, we can extend the range of validity for the construction described in section 7.2 to the triangular region shown in Fig. 4.2 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 $2N - 2$

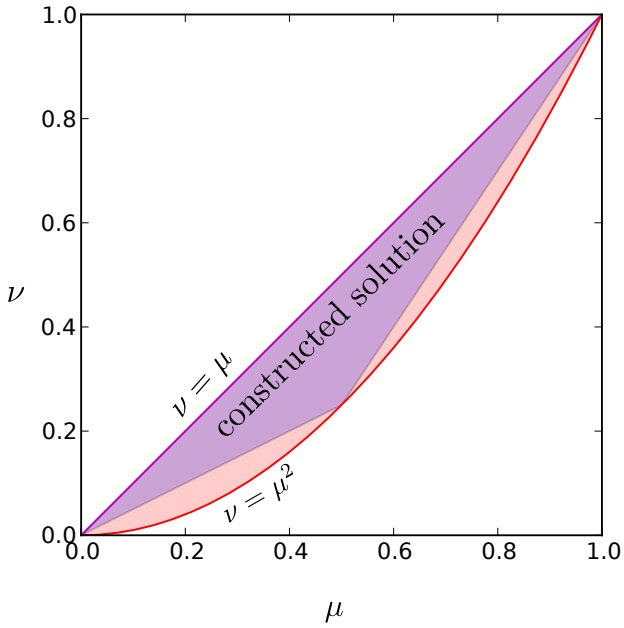

Figure 7.2: The full shaded region includes all allowed values for the constraints μ and ν for all possible probability distributions, replotted from Fig. 4.2. The triangular blue shaded region includes all possible values for the constraints beginning with either of our constructed solutions with $\mu = 1/2$ and $\nu = 1/4$. Choosing other values of μ and ν for the construction described in Appendix 7.2 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.

states of the original model, respectively. The entropy of this extended distribution must be no greater than the entropy of the original distribution (Eq. (7.23)), since the same number of states are active, but now they are not weighted equally, so this remains a low entropy distribution.

Chapter 8

Minimum entropy for exchangable 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 I have already shown. Indeed, as I 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:

$$n_c = 1 + \frac{N(N+1)}{2}. \tag{8.1}$$

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. (4.7), (4.8), and (4.9), 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 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 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

$$p(N/2) = \frac{N-1}{N}\frac{(N/2)!^2}{N!}, \quad N \text{ even}, \tag{8.2}$$

which corresponds to an entropy of

$$\begin{aligned} \tilde{S}_2^{exch} &= \frac{\log_2(2N)}{N} \\ &\quad + \frac{N-1}{N}\log_2\left(\frac{NN!}{(N/2)!^2(N-1)}\right) \qquad (8.3) \\ &\approx N - 1/2\log_2(N) - 1/2\log_2(2\pi) \\ &\quad + O\left[\frac{\log_2(N)}{N}\right]. \qquad (8.4) \end{aligned}$$

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, I find that the scaling behavior of the exchangeable minimum entropy is linear with N as shown in Fig. 8.1. I 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. (8.3) 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.

I 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 I have shown scales only 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 not surprising 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 Fig 3.1. The large gap between $\tilde{S}_2^{exch}$ and $\tilde{S}_2$

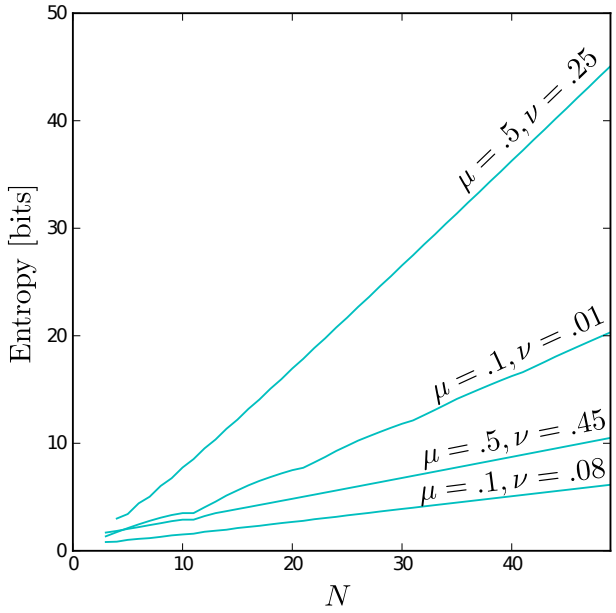

Figure 8.1: The minimum entropy for exchangeable distributions versus N for various values of μ and ν. Note that, like the maximum entropy, the exchangeable minimum entropy scales linearly with N as $N \to \infty$, albeit with a smaller slope for $\nu \neq \mu^2$. We can calculate the entropy exactly for $\mu = 0.5$ and $\nu = 0.25$ as $N \to \infty$, and we find that the leading term is indeed linear: $\tilde{S}_2^{exch} \approx N - 1/2 \log_2(N) - 1/2 \log_2(2\pi) + O[\log_2(N)/N]$.

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.

Chapter 9

Implications for communications 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 entropy as used in the context of information theory. I now examine consequences of our results for engineered communication systems. Specifically, consider a device such as a digital camera that exploits compressed sensing [90, 91] 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 I am unaware of any that do not depend on some knowledge about the input signal, such as its sparse structure [90, 92]. 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 I 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. I 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. Fig. 3.1 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)). I experimentally find that any exchangeable distribution has as much entropy as the maximum entropy solution, up to terms of order $\log_2(N)$.

Similarly, 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 [93, 94]. 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 pariwise 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 [94].

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 [95]. 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 [42, 43, 89, 96, 97], combinatorial block designs [41, 98], Hadamard matrix theory [51, 99], and linear programming [50], among others. Our construction here of two families of low entropy distributions fit to specified mean activities and pairwise statistics adds to this literature.

Chapter 10

Discussion

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 I have shown that second order statistics do not significantly constrain the range of allowed entropies, though other constraints, such as exchangeability, do guarantee that entropy be extensive.

I have shown that in order for the 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 have to agree with other, more direct, estimates of the entropy [100–106]. The fact that maximum entropy models chosen with relatively arbitrary selection of statistics are able to match the entropy of the system they model lends credence to the merits of this approach. I have also indicated how, 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.

Bibliography

[1] I. H. Stevenson and K. P. Kording. "How advances in neural recording affect data analysis." In: *Nature neuroscience* 14.2 (Feb. 2011), pp. 139–42.

[2] V. Mountcastle. "Modality and Topographic Properties of Single Neurons in Cat's Somatic Sensory Cortex". In: *J. neurophysiol* (1957).

[3] P. Rakic. "Guidance of neurons migrating to the fetal monkey neocortex". In: *Brain research* 33 (1971), pp. 471–476.

[4] P. Rakic. "Mode of cell migration to the superficial layers of fetal monkey neocortex." In: *The Journal of comparative neurology* 145.1 (May 1972), pp. 61–83.

[5] P. Rakic. "Commentary Radial versus tangential migration of neuronal clones in the developing cerebral cortex". In: *Proceedings of the National Academy of Sciences* 92.December (1995), pp. 11323–11327.

[6] D. B. Chklovskii, S. Vitaladevuni, and L. K. Scheffer. "Semi-automated reconstruction of neural circuits using electron microscopy." In: *Current opinion in neurobiology* 20.5 (Oct. 2010), pp. 667–75.

[7] K. L. Briggman, M. Helmstaedter, and W. Denk. "Wiring specificity in the direction-selectivity circuit of the retina." In: *Nature* 471.7337 (Mar. 2011), pp. 183–8.

[8] R. E. Marc, B. W. Jones, J. S. Lauritzen, C. B. Watt, and J. R. Anderson. "Building retinal connectomes." In: *Current opinion in neurobiology* 22.4 (Aug. 2012), pp. 568–74.

[9] D. Bock, W. Lee, and A. Kerlin. "Network anatomy and in vivo physiology of visual cortical neurons". In: *Nature* 471.7337 (2011), pp. 177–182.

[10] T. Deacon. *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company, 2011, p. 624.

[11] D. J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford Univ Press, 1996, p. 433.

[12] G. Tononi. "An information integration theory of consciousness." In: *BMC neuroscience* 5 (Nov. 2004), p. 42.

[13] G. Tononi. "Consciousness as integrated information: a provisional manifesto." In: *The Biological bulletin* 215.3 (Dec. 2008), pp. 216–42.

[14] W. Russ, D. Lowery, P. Mishra, M. Yaffe, and R. Ranganathan. "Natural-like function in artificial WW domains". In: *Nature* 437 (2005), pp. 579–583.

[15] M. Socolich, S. Lockless, W. Russ, H. Lee, K. Gardner, and et al. "Evolutionary information for specifying a protein fold". In: *Nature* 437 (2005), pp. 512–518.

[16] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan. "Maximum entropy models for antibody diversity". In: *Proc Nat'l Acad Sci (USA)* 107 (2010), pp. 5405–5410.

[17] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. "Weak pairwise correlations imply strongly correlated network states in a neural population". In: *Nature* 440.7087 (2006), pp. 1007–12.

[18] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky. "The structure of multi-neuron firing patterns in primate retina". In: *J. Neurosci.* 26.32 (2006), pp. 8254–66.

[19] G. Tkacik, E. Schneidman, M. Berry, and W. Bialek. "Ising models for networks of real neurons". In: *Arxiv preprint q-bio* (Jan. 2006).

[20] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. Smith, and et al. "A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro". In: *J Neurosci* 28 (2008), pp. 505–518.

[21] J. Shlens, G. D. Field, J. L. Gauthier, M. Greschner, A. Sher, A. M. Litke, and E. J. Chichilnisky. "The Structure of Large-Scale Synchronized Firing in Primate Retina". In: *Journal of Neuroscience* 29.15 (Apr. 2009), pp. 5022–5031.

[22] E. Ganmor, R. Segev, and E. Schneidman. "Sparse low-order interaction network underlies a highly correlated and learnable neural population code". In: *Proc Natl Acad Sci USA* 108.23 (2011), pp. 9679–84.

[23] S. Yu, D. Huang, W. Singer, and D. Nikolic. "A small world of neuronal synchrony". In: *Cereb Cortex* 18 (2008), pp. 2891–2901.

[24] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. Walczak. "Statistical mechanics for natural flocks of birds". In: *arXiv.org:1107.0604 [physics.bio-ph]* (2011).

[25] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Phys. Rev.* 106 (1957), pp. 620–630.

[26] N. Ay and A. Knauf. "Maximizing multi-information". In: *arXiv preprint math-ph/0702002* 40.1 (2007), pp. 1–23.

[27] M. Bethge and P. Berens. "Near-maximum entropy models for binary neural representations of natural images". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Cambridge, MA: MIT Press, 2008, pp. 97–104.

[28] Y. Roudi, S. H. Nirenberg, and P. E. Latham. "Pairwise maximum entropy models for studying large biological systems: when they can and when they can't work". In: *PLoS Computational Biology* (2009), 5:e1000380.

[29] S. H. Nirenberg and J. D. Victor. "Analyzing the activity of large populations of neurons: how tractable is the problem?" In: *Curr Opin Neurobiol.* 17(4) (2007), pp. 397–400.

[30] F. Azhar and W. Bialek. "When are correlations strong?" In: *arXiv.org:1012.5987 [q-bio.NC]* (2010).

[31] J. H. Macke, M. Opper, and M. Bethge. "Common Input Explains Higher-Order Correlations and Entropy in a Simple Model of Neural Population Activity". In: *Physical Review Letters* 106.20 (May 2011), p. 208102.

[32] J. Sylvester. "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." In: *Philosophical Magazine* 34 (1867), pp. 461–475.

[33] P. Diaconis. "Finite forms of de Finetti's theorem on exchangeability". In: *Synthese* 36 (1977).

[34] C. Shannon. "A mathematical theory of communications, I and II". In: *Bell Syst. Tech. J* 27 (1948), pp. 379–423.

[35] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 1991.

[36] B. De Schutter. "Minimal state-space realization in linear system theory: an overview". In: *Journal of Computational and Applied Mathematics* 121.1-2 (2000), pp. 331–354.

[37] J. Carter and M. Wegman. "Universal classes of hash functions". In: *Journal of computer and system sciences* 18.2 (1979), pp. 143–154.

[38] M. Sipser. "A complexity theoretic approach to randomness". In: *Proceedings of the fifteenth annual ACM symposium on Theory of computing*. ACM. 1983, pp. 330–335.

[39] L. Stockmeyer. "The complexity of approximate counting". In: *Proceedings of the fifteenth annual ACM symposium on Theory of computing*. ACM. 1983, pp. 118–126.

[40] B. Chor, O. Goldreich, J. Hasted, J. Freidmann, S. Rudich, and R. Smolensky. "The bit extraction problem or t-resilient functions". In: *Foundations of Computer Science, 1985., 26th Annual Symposium on*. IEEE. 1985, pp. 396–407.

[41] R. Karp and A. Wigderson. "A fast parallel algorithm for the maximal independent set problem". In: *Journal of the ACM (JACM)* 32.4 (1985), pp. 762–773.

[42] M. Luby. "A simple parallel algorithm for the maximal independent set problem". In: *SIAM J. Comput.* 15.4 (1986), pp. 1036–1053.

[43] N. Alon, L. Babai, and A. Itai. "A fast and simple randomized parallel algorithm for the maximal independent set problem". In: *Journal of algorithms* 7.4 (1986), pp. 567–583.

[44] W. Alexi, B. Chor, O. Goldreich, and C. Schnorr. "RSA and Rabin functions: Certain parts are as hard as the whole". In: *SIAM J. Comput.* 17.2 (1988), pp. 194–209.

[45] B. Chor and O. Goldreich. "On the power of two-point based sampling". In: *Journal of Complexity* 5.1 (1989), pp. 96–106.

[46] B. Berger and J. Rompel. "Simulating (log cn)-wise independence in NC". In: *Journal of the ACM (JACM)* 38.4 (1991), pp. 1026–1046.

[47] L. Schulman. "Sample spaces uniform on neighborhoods". In: *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*. ACM. 1992, pp. 17–25.

[48] M. Luby. "Removing randomness in parallel computation without a processor penalty". In: *Journal of Computer and System Sciences* 47.2 (1993), pp. 250–286.

[49] R. Motwani, J. Naor, and M. Naor. "The probabilistic method yields deterministic parallel algorithms". In: *Journal of Computer and System Sciences* 49.3 (1994), pp. 478–516.

[50] D. Koller and N. Megiddo. "Constructing small sample spaces satisfying given constraints". In: *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*. ACM. 1993, pp. 268–277.

[51] H. Karloff and Y. Mansour. "On construction of k-wise independent random variables". In: *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*. ACM. 1994, pp. 564–573.

[52] J. Pierce. "The Early Days of Information Theory". In: *Information Theory, IEEE Transactions on* I (1973).

[53] M. Tribus and E. McIrvine. "Energy and information". In: *Scientific American* (1971), p. 224.

[54] J. Uffink. "Compendium of the foundations of classical statistical physics". 2006.

[55] D. Lavis. "Boltzmann and Gibbs: An Attempted Reconciliation". In: *Studies in History and Philosophy of Science Part B: ...* June 2005 (2005), pp. 1–30.

[56] S. Goldstein. "Boltzmann ' s Approach to Statistical Mechanics". In: (2001), pp. 39–54.

[57] A. Kolmogorov. "A simplified proof of the Birkhoff-Khinchin ergodic theorem". In: *Uspekhi Matematicheskikh Nauk* 5 (1938), pp. 52–56.

[58] V. Arnold and A. Avez. *Ergodic problems of classical mechanics*. Benjamin, 1968, p. 286.

[59] J. Lebowitz and O. Penrose. "Modern ergodic theory". In: *Physics Today* February (1973), pp. 23–29.

[60] E. Jaynes. "Information theory and statistical mechanics". In: *Physical review* 106.4 (1957), pp. 620–630.

[61] E. Jaynes. "Information theory and statistical mechanics. II". In: *Physical review* 108.2 (1957), pp. 171–190.

[62] J. Uffink. "The Constraint Rule of the Maximum Entropy Principle". In: *Studies in History and Philosophy of Science Part B: ...* 27.1 (1996).

[63] J. Bernardo and A. Smith. *Baysian Theory*. New York: Wiley, 1994.

[64] I. Hacking. *An Introduction to Probability and Inductive Logic*. Cambridge, England: Cambridge University Press, 2001.

[65] B. D. Finetti. "Bayesianism: its unifying role for both the foundations and applications of statistics". In: *Statistical Review/Revue Internationale de Statistique* 42.2 (1974), pp. 117–130.

[66] B. de Finetti, M. C. Galavotti, H. Hosni, and A. Mura. "Exhangability". In: *Philosophical Lectures on Probability*. 2008, pp. 75–85.

[67] W. E. Johnson. *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press, 1924.

[68] W. E. Johnson. "Probability: The deductive and inductive problems". In: *Mind* 41 (1932), pp. 409–423.

[69] S. Zabell. "Carnap and the logic of inductive inference". In: *Handbook of the history of logic: Inductive logic* 10 (2004).

[70] K. Friedman and A. Shimony. "Jaynes's maximum entropy prescription and probability theory". In: *Journal of Statistical Physics* 3.4 (1971), pp. 381–384.

[71] M. Tribus and H. Motroni. "Comments on the Paper "Jaynes's Maximum Entropy Prescription and Probability Theory"". In: *Journal of Statistical Physics* 4.2 (1972), pp. 227–228.

[72] A. Hobson. "The interpretation of inductive probabilities". In: *Journal of Statistical Physics* 6.2-3 (1972), pp. 189–193.

[73] D. Gage and D. Hestenes. "Comment on the Paper "Jaynes's Maximum Entropy Prescription and Probability Theory"". In: *Journal of Statistical Physics* 7.4 (1973), pp. 89–90.

[74] P. Cheeseman and J. Stutz. "On the relationship between Bayesian and maximum entropy inference". In: *AIP Conference Proceedings*. 2004, pp. 445–461.

[75] A. Caticha. "Information and entropy". In: *arXiv preprint arXiv:0710.1068v1* (2007).

[76] T. Seidenfeld. "Entropy and Uncertainty". In: *Advances in the Statistical Sciences: Foundations of ...* 53.4 (1986), pp. 467–491.

[77] T. Seidenfeld. "Why I am not an Objective Bayesian; some reflections prompted by Rosenkrantz". In: *Theory and Decision* 11 (1979), pp. 413–440.

[78] B. Skyrms. "Maximum Entropy Inference as a special case of conditionalization". In: *Synthese* 63 (1985), pp. 55–74.

[79] B. Skyrms. "Updating, Supposing, and MaxEnt". In: *Theory and Decision* 22 (1987), pp. 225–246.

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

[81] T. Tanaka. "Mean-field theory of Boltzmann machine learning". In: *Physical Review Letters E* (Jan. 1998).

[82] G. E. Hinton, S. Osindero, and Y.-W. Teh. "A fast learning algorithm for deep belief nets." In: *Neural Computation* 18.7 (July 2006), pp. 1527–1554.

[83] A. Hyvärinen. "Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables". In: *IEEE Transactions on Neural Networks* (Jan. 2007).

[84] T. Broderick, M. Dudík, G. Tkačik, R. Schapire, and W. Bialek. "Faster solutions of the inverse pairwise Ising problem". In: *E-print arXiv* (Jan. 2007).

[85] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese. "New method for parameter estimation in probabilistic models: minimum probability flow". In: *Phys Rev Lett* 107(22) (2011), p. 220601.

[86] D. Gale, H. W. Kuhn, and A. W. Tucker. "Linear Programming and the Theory of Games". In: *Activity Analysis of Production and Allocation* (1951), pp. 317–329.

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

[88] J. B. Rosen. "Global Minimization of a Linearly Constrained Function by Partition of Feasible Domain". In: *Mathematics of Operations Research* 8.2 (1983), pp. 215–230.

[89] A. Joffe. "On a set of almost deterministic k-independent random variables". In: *the Annals of Probability* 2.1 (1974), pp. 161–162.

[90] D. L. Donoho. "Compressed sensing". In: *Stanford Technical Report* (2004).

[91] E. J. Candés. "Compressive Sampling". In: *Proceedings of the International Congress of Mathematicians*. Madrid, Spain: European Mathematical Society, 2006.

[92] S. Sarvotham, D. Baron, and R. G. Baraniuk. "Measurements vs. Bits: Compressed Sensing meets Information Theory". In: *ECE Publications*. Rice University, 2006.

[93] B. Hao. *Elementary symbolic dynamics and chaos in dissipative systems*. 1989.

[94] C. Shalizi and J. Crutchfield. "Computational mechanics: Pattern and Prediction, Structure and Simplicity". In: *Journal of statistical physics* 104.3-4 (2001), pp. 817–879.

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

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

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

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

[99] H. Lancaster. "Pairwise statistical independence". In: *The Annals of Mathematical Statistics* 36.4 (1965), pp. 1313–1317.

[100] S. Panzeri, R. Senatore, M. a. Montemurro, and R. S. Petersen. "Correcting for the sampling bias problem in spike train information measures." In: *Journal of neurophysiology* 98.3 (Sept. 2007), pp. 1064–72.

[101] E. T. Rolls and A. Treves. "The neuronal encoding of information in the brain." In: *Progress in neurobiology* 95.3 (Sept. 2011), pp. 448–490.

[102] M. Crumiller, B. Knight, Y. Yu, and E. Kaplan. "Estimating the amount of information conveyed by a population of neurons." In: *Frontiers in neuroscience* 5.July (Jan. 2011), p. 90.

[103] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. "Entropy and Information in Neural Spike Trains". In: *Physical Review Letters* 80.1 (Jan. 1998), p. 197.

[104] I. Nemenman, W. Bialek, and R. de Ruyter Van Steveninck. "Entropy and information in neural spike trains: Progress on the sampling problem". In: *Physical Review E* 69.5 (June 2004), p. 56111.

[105] A. Borst and F. E. Theunissen. "Information theory and neural coding." In: *Nature neuroscience* 2.11 (Nov. 1999), pp. 947–57.

[106] R. Quian Quiroga and S. Panzeri. "Extracting information from neuronal populations: information theory and decoding approaches." In: *Nature reviews. Neuroscience* 10.3 (Mar. 2009), pp. 173–85.