Statistical mechanics of complex neural systems and high dimensional data

Madhu Advani, Subhaneil Lahiri, and Surya Ganguli

Dept. of Applied Physics, Stanford University, Stanford, CA

E-mail: sganguli@stanford.edu

Abstract.

Recent experimental advances in neuroscience have opened new vistas into the immense complexity of neuronal networks. This proliferation of data challenges us on two parallel fronts. First, how can we form adequate theoretical frameworks for understanding how dynamical network processes cooperate across widely disparate spatiotemporal scales to solve important computational problems? And second, how can we extract meaningful models of neuronal systems from high dimensional datasets? To aid in these challenges, we give a pedagogical review of a collection of ideas and theoretical methods arising at the intersection of statistical physics, computer science and neurobiology. We introduce the interrelated replica and cavity methods, which originated in statistical physics as powerful ways to quantitatively analyze large highly heterogeneous systems of many interacting degrees of freedom. We also introduce the closely related notion of message passing in graphical models, which originated in computer science as a distributed algorithm capable of solving large inference and optimization problems involving many coupled variables. We then show how both the statistical physics and computer science perspectives can be applied in a wide diversity of contexts to problems arising in theoretical neuroscience and data analysis. Along the way we discuss spin glasses, learning theory, illusions of structure in noise, random matrices, dimensionality reduction, and compressed sensing, all within the unified formalism of the replica method. Moreover, we review recent conceptual connections between message passing in graphical models, and neural computation and learning. Overall, these ideas illustrate how statistical physics and computer science might provide a lens through which we can uncover emergent computational functions buried deep within the dynamical complexities of neuronal networks.

PACS numbers: 87.19.L-, 87.10.Vg, 89.20.-a

Keywords: replica method, cavity method, message passing, neural networks, spin glasses, learning, random matrices, high dimensional data, random projections, compressed sensing

Contents

first illustration of the replica method?
relation??
classifier?
high dim? model-based data analysis?
relation? same model for diff implementation? more detailed model?
use of this??
Data->model? can you get interpretable model?
compare to vision study?

1. Introduction

Neuronal networks are highly complex dynamical systems consisting of large numbers of neurons interacting through synapses [1, 2, 3]. Such networks subserve dynamics over multiple time-scales. For example, on fast time scales, on the order of milliseconds, synaptic connectivity is approximately constant, and this connectivity directs the flow of electrical activity through neurons. On slower timescales, on the order of seconds to minutes and beyond, the synaptic connectivity itself can change through synaptic plasticity induced by the statistical structure of experience, which itself can stay constant over even longer timescales. These synaptic changes are thought to underly our ability to learn from experience. To the extent that such separations of timescale hold, one can exploit powerful tools from the statistical physics of disordered systems to obtain a remarkably precise understanding of neuronal dynamics and synaptic learning in basic models. For example, the replica method and the cavity method, which we introduce and review below, become relevant because they allow us to understand the statistical properties of many interacting degrees of freedom that are coupled to each other through some fixed, or quenched, interactions that may be highly heterogenous, or disordered.

However, such networks of neurons and synapses, as well as the dynamical processes that occur on them, are not simply tangled webs of complexity that exist for their own sake. Instead they have been sculpted over time, through the processes of evolution, learning and adaptation, to solve important computational problems necessary for survival. Thus biological neuronal networks serve a *function* that is useful for an organism in terms of improving its evolutionary fitness. The very concept of function does not of-course arise in statistical physics, as large disordered statistical mechanical systems, like glasses or non-biological polymers do not arise through evolutionary processes. In general, the function that a biological network performs (which may not always be a-priori obvious) can provide a powerful way to understand both its structure and the details of its complex dynamics [4]. As the functions performed by neuronal networks are often computational in nature, it can be useful to turn to ideas from distributed computing algorithms in computer science for sources of insight into how networks of neurons may learn and compute in a distributed manner. In this review we also focus on distributed message passing algorithms whose goal is to compute the marginal probability distribution of a single degree of freedom in a large interacting system. Many problems in computer science, including error correcting codes and constraint satisfaction, can be formulated as message passing problems [5]. As we shall review below, message passing is intimately related to the replica and cavity methods of statistical physics, and can serve as a framework for thinking about how specific dynamical processes of neuronal plasticity and network dynamics may solve computational problems like learning and inference.

This combination of ideas from statistical physics and computer science are not only useful in thinking about how network dynamics and plasticity may mediate computation, but also for thinking about ways to analyze large scale datasets arising from high

throughput experiments in neuroscience. Consider a data set consisting of P points in an N dimensional feature space. Much of the edifice of classical statistics and machine learning has been tailored to the situation in which P is large and N is small. This is the low dimensional data scenario in which we have large amounts of data. In such situations, many classical unsupervised machine learning algorithms can easily find structures or patterns in data, when they exist. However, the advent of high throughput techniques in neuroscience has pushed us into a high dimensional data scenario in which both P and N are large, but their ratio is $O(1)$. For example, we can simultaneously measure the activity of $O(100)$ neurons but often only under a limited number of trials (i.e. also $O(100)$) for any given experimental condition. Also, we can measure the single cell gene expression levels of $O(100)$ genes but only in a limited number of cells. In such a high dimensional scenario, it can be difficult to find statistically significant patterns in the data, as often classical unsupervised machine learning algorithms yield illusory structures. The statistical physics of disordered systems again provides a powerful tool to understand high dimensional data, because many machine learning algorithms can be formulated as the minimization of a data dependent energy function on a set of parameters. We review below how statistical physics plays a useful role in understanding possible illusions of structure in high dimensional data, as well as approaches like random projections and compressed sensing, which are tailored to the high dimensional data limit.

We give an outline and summary of this review as follows. In section 2 we introduce the fundamental techniques of the replica method and cavity method within the context of a paradigmatic example, the Sherrington-Kirkpatrick (SK) model [6] of a spin glass [7, 8, 9]. In a neuronal network interpretation, such a system qualitatively models a large network in which the heterogenous synaptic connectivity is fixed and plays the role of quenched disorder. On the otherhand, neuronal activity can fluctuate and we are interested in understanding the statistical properties of the neuronal activity. We will find that certain statistical properties, termed self-averaging properties, do not depend on the detailed realization of the disordered connectivity matrix. This is a recurring theme in these notes; in large random systems with microscopic heterogeneity, striking levels of almost deterministic macroscopic order can arise in ways that do not depend on the details of the heterogeneity. Such order can govern dynamics and learning in neuronal networks, as well as the performance of machine learning algorithms in analyzing data, and moreover, this order can be understood theoretically through the replica and cavity methods.

learning/inference?

We end section 2 by introducing message passing which provides an algorithmic perspective on the replica and cavity methods. Many models in equilibrium statistical physics are essentially equivalent to joint probability distributions over many variables, which are equivalently known and described as graphical models in computer science [10]. Moreover, many computations in statistical physics involve computing marginal probabilities of a single variable in such graphical models. Message passing, also known in special cases as belief propagation [11], involves a class of algorithms that

really?? try to draw the connection!

yield dynamical systems whose fixed points are designed to approximate marginal probabilities in graphical models. Another recurring theme in these notes is that certain aspects of neuronal dynamics may profitably be viewed through the lens of message passing; in essence, these neuronal (and also synaptic) dynamics can be viewed as approximate versions of message passing in a suitably defined graphical model. This correspondence between neuronal dynamics and message passing allows for the possibility of both understanding the computational significance of existing neuronal dynamics, as well as deriving hypotheses for new forms of neuronal dynamics from a computational perspective.

In section 3 we apply the ideas of replicas, cavities and messages introduced in section 2 to the problem of learning in neuronal networks as well as machine learning (see [12] for a beautiful book length review of this topic). In this context, training examples, or data play the role of quenched disorder, and the synaptic weights of a network, or the learning parameters of a machine learning algorithm, play the role of fluctuating statistical mechanical degrees of freedom. In the zero temperature limit, these degrees of freedom are optimized, or learned, by minimizing an energy function. The learning error, as well as aspects of the learned structure, can be described by macroscopic order parameters that do not depend on the detailed realization of the training examples, or data. We show how to compute these order parameters for the classical perceptron [13, 14], thereby computing its storage capacity. Also we compute these order parameters for classical learning algorithms, including Hebbian learning, principal components analysis (PCA), and K-means clustering, revealing that all of these algorithms are prone to discovering illusory structures that reliably arise in random realizations of high dimensional noise. Finally, we end section 3 by discussing an application of message passing to learning with binary valued synapses, known to be an NP-complete problem [15, 16]. The authors of [17, 18] derived a biologically plausible learning algorithm capable of solving random instantiations of this problem by approximating message passing in a joint probability distribution over synaptic weights determined by the training examples.

In section 4, we discuss the eigenvalue spectrum of random matrices. Matrices from many random matrix ensembles have eigenvalue spectra whose probability distributions display fascinating macroscopic structures that do not depend on the detailed realization of the matrix elements. These spectral distributions play a central role in a wide variety of fields [19, 20]; within the context of neural networks for example, they play a role in understanding the stability of linear neural networks, the transition to chaos in nonlinear networks [21], and the analysis of high dimensional data. We begin section 4 by showing how replica theory can also provide a general framework for computing the typical eigenvalue distribution of a variety of random matrix ensembles. Then we focus on understanding an ensemble of random empirical covariance matrices (the Wishart ensemble [22]) whose eigenvalue distribution, known as the Marcenko-Pastur distribution [23], provides a null model for the outcome of PCA applied to high dimensional data. Moreover, we review how the eigenvalues of many random matrix

ensembles can be thought of as Coulomb charges living in the complex plane, and the distribution of these eigenvalues can be thought of as the thermally equilibriated charge density of this Coulomb gas, which is stabilized via the competing effects of a repulsive two dimensional Coulomb interaction and an attractive confining external potential. Moreover we review how the statistics of the largest eigenvalue, which obeys the Tracy-Widom distribution [24, 25], can be understood simply in terms of thermal fluctuations of this Coulomb gas [26, 27]. The statistics of this largest eigenvalue will make an appearance later in section 5 when we discuss how random projections distort the geometry of manifolds. Overall, section 4 illustrates the power of the replica formalism, and plays a role in connecting the statistical physics of two dimensional Coulomb gases to PCA in section 3.5 and geometric distortions induced by dimensionality reduction in section 5.3.

In section 5 we discuss the notion of random dimensionality reduction. High dimensional data can be difficult to both model and process. One approach to circumvent such difficulties is to reduce the dimensionality of the data; indeed many machine learning algorithms search for optimal directions on which to project the data. As discussed in section 3.5, such algorithms yield projected data distributions that reveal low dimensional, illusory structures that do not exist in the data. An alternate approach is to simply project the data onto a random subspace. As the dimensionality of this subspace is lower than the ambient dimensionality of the feature space in which the data resides, features of the data will necessarily be lost. However, it is often the case that interesting data sets lie along low dimensional submanifolds in their ambient feature space. In such situations, a random projection above a critical dimension, that is more closely related to the dimensionality of the submanifold than to the dimensionality of the ambient feature space, often preserves a surprising amount of structure of the submanifold. In section 5 we review the theory of random projections and their ability to preserve the geometry of data submanifolds. We end section 5 by introducing a statistical mechanics approach to random dimensionality reduction of simple random submanifolds, like point clouds and hyperplanes. This analysis connects random dimensionality reduction to extremal fluctuations of 2D Coulomb gases discussed in sections 4.3 and 4.4.

The manifold of sparse signals forms a ubiquitous and interesting low dimensional structure that accurately captures many types of data. The field of compressed sensing [28, 29], discussed in section 6, rests upon the central observation that a sparse high dimensional signal can be recovered from a random projection down to a surprisingly low dimension by solving a computationally tractable convex optimization problem, known as L_1 minimization. In section 6 we focus mainly on the analysis of L_1 minimization based on statistical mechanics and message passing. For readers who are more interested in applications of random projections, compressed sensing and L_1 minimization to neuronal information processing and data analysis, we refer them to [30]. There, diverse applications of how the techniques in sections 5 and 6 can be used to acquire and analyze high dimensional neuronal data are discussed, including, magnetic resonance

imaging [31, 32, 33], compressed gene expression arrays [34], compressed connectomics [35, 36], receptive field measurements, and fluorescence microscopy [37, 38] of multiple molecular species at high spatiotemporal resolution [39] using single pixel camera [40, 41] technology. Also diverse applications of these same techniques to neuronal information processing are discussed in [30], including semantic information processing [42, 43, 44], short-term memory [45, 46], neural circuits for L_1 minimization [47], learning sparse representations [48, 49], regularized learning of high dimensional synaptic weights from limited examples [50], and axonally efficient long range brain communication through random projections [51, 52, 53, 54].

After introducing CS in 6.1, we show how replica theory can be used to analyze its performance in section 6.2. Remarkably, the performance of CS, unlike other algorithms discussed in section 3.5, displays a phase transition. For any given level of signal sparsity, there is a critical lower bound on the dimensionality of a random projection which is required to accurately recover the signal; this critical dimension decreases with increasing sparsity. Also, in section 6.3 we review how the L_1 minimization problem can be formulated as a message passing problem [55]. This formulation yields a message passing dynamical system that qualitatively mimics neural network dynamics with a crucial history dependence terms. L_1 minimization via gradient descent has been proposed as a framework for neuronal dynamics underlying sparse coding in both vision [56] and olfaction [57]. On the otherhand, the efficiency of message passing in solving L_1 minimization, demonstrated in [55] may motivate revisiting the issue of sparse coding in neuroscience, and the role of history dependence in sparse coding network dynamics.

Finally, the appendix in section 8 provides an overview of the replica method, in a general form that is immediately applicable to spin glasses, perceptron learning, unsupervised learning, random matrices and compressed sensing. Overall, the replica method is a powerful, if non-rigorous, method for analyzing the statistical mechanics of systems with quenched disorder. We hope that this exposition of the replica method, combined with the cavity and message passing methods discussed in this review within a wide variety of disparate contexts, will help enable students and researchers in both theoretical neuroscience and physics to learn about exciting interdisciplinary advances made in the last few decades at the intersection of statistical physics, computer science, and neurobiology.

2. Spin Glass Models of Neural Networks

The SK Model [6] is a prototypical example of a disordered statistical mechanical system. It has been employed as a simple model of spin glasses [7, 8], as well as neural networks [58], and has made a recent resurgence in neuroscience within the context of maximum entropy modeling of spike trains [59, 60]. It is defined by the energy function

$$H(\mathbf{s}, \mathbf{J}) = -\frac{1}{2} \sum_{ij} \mathbf{J}_{ij} s_i s_j, \tag{1}$$

where the s_i are N spin degrees of freedom taking the values ± 1. In a neural network interpretation, $\mathbf{s}_i$ represents the activity state of a neuron and $\mathbf{J}$ is the synaptic connectivity matrix of the network. This Hamiltonian yields an equilibrium Gibbs distribution of neural activity given by

$$P_{\mathbf{J}}(\mathbf{s}) = \frac{1}{Z[\mathbf{J}]} e^{-\beta H(\mathbf{S},\mathbf{J})} \tag{2}$$

where

$$Z[\mathbf{J}] = \sum_{\mathbf{S}} e^{-\beta H(\mathbf{S},\mathbf{J})} \tag{3}$$

is the partition function, and β is an inverse temperature reflecting sources of noise. The connectivity matrix is chosen to be random, where each $\mathbf{J}_{ij}$ is an independent, identically distributed (i.i.d) zero mean Gaussian with variance $1/N$.

The main property of interest is the statistical structure of high probability (low energy) activity patterns. Much progress in spin glass theory [7] has revealed a physical picture in which the Gibbs distribution in (2) decomposes at low temperature (large β) into many "lumps" of probability mass (more rigorously, pure states [61]) concentrated on subsets of activity patterns. Equivalently, these lumps can be thought of as concentrated on the minima of a free energy landscape with many valleys. Each lump, indexed by a, is characterized by a mean activity pattern $m_i^a = \langle s_i \rangle_a$, where $\langle \cdot \rangle_a$ is an average over configurations belonging to the free energy valley a, and a probability mass P_a (the probability that a random activity pattern belongs to valley a). In the large N limit, free energy barriers between valleys diverge, so that in dynamical versions of this model, if an activity pattern starts in one valley, it will stay in that valley for infinite time. Thus ergodicity is broken, as time average activity patterns are not equal to the full Gibbs average activity pattern. The network can thus maintain multiple steady states, and we are interested in understanding the structure of these steady states.

what does a single neuron state(s_i) look like given the whole neuron array state are in the minimum — or lumps of local minimum, aka valley

Now the detailed activity pattern in any free energy minimum a (i.e the mean pattern m_i^a) depends on the detailed realization of the connectivity $\mathbf{J}$, and is hard to compute. However, many interesting quantities, that involve averages over all neurons, are self-averaging, which by definition means that their fluctuations across different realizations of $\mathbf{J}$ vanish in the large N limit. As we see below, typical values of such quantities, for any given realization of $\mathbf{J}$, can be computed theoretically by computing their average over all $\mathbf{J}$. One interesting quantity that probes the geometry of free energy minima is the distribution of overlaps between all pairs of activity patterns. If the activity patterns belong to two valleys, a and b, then the overlap is

$$q_{ab} = \frac{1}{N} \sum_i m_i^a m_i^b. \tag{4}$$

not like mi^a - mi^b? why times each other? — if +-1, then times make sense. But you are taking average, still +-1?

Now since P_a is the probability a randomly chosen activity pattern belongs to valley a, the distribution of overlaps between any two pairs of activity patterns independently chosen from (2) is given by

$$P_{\mathbf{J}}(q) = \sum_{ab} P_a \, P_b \, \delta(q - q_{ab}). \tag{5}$$

This distribution turns out not to be self-averaging (it fluctuates across realizations of $\mathbf{J}$), unless there is only one valley, or state (modulo the reflection symmetry $\mathbf{s}_i \to -\mathbf{s}_i$), in which case the distribution becomes concentrated at a single number q, which is the self-overlap of the state, $q = \frac{1}{N}\sum_i m_i^2$. If there is indeed one state, then q does not depend on the detailed realization of $\mathbf{J}$ and provides a measure of the variability of mean activity across neurons due to the quenched disorder in the connectivity. In the case of multiple valleys, one can also compute the disorder averaged overlap distribution $\langle\langle\, P_{\mathbf{J}}(q)\,\rangle\rangle_{\mathbf{J}}$; despite the fact that the overlap distribution $P_{\mathbf{J}}(q)$ may not be self-averaging, its average over $\mathbf{J}$ can still yield a wealth of information about the geometric organization of free energy minima in neural activity space. This can be done using the replica method, which we now introduce.

2.1. Replica Solution

To understand the statistical properties of the Gibbs distribution in (2), it is useful to compute its free energy $-\beta F[\mathbf{J}] = \ln Z[\mathbf{J}]$. Correlations between neurons can then be computed via suitable derivatives of the free energy. Fortunately, the free energy is self-averaging, which means that to understand the free energy for any realization of $\mathbf{J}$, it suffices to compute its average over all $\mathbf{J}$:

$$\langle\langle\, -\beta F[\mathbf{J}]\,\rangle\rangle_{\mathbf{J}} = \langle\langle\, \ln Z[\mathbf{J}]\,\rangle\rangle_{\mathbf{J}}, \tag{6}$$

where $\langle\langle\, \cdot\,\rangle\rangle_{\mathbf{J}}$ denotes an average over the disorder $\mathbf{J}$. This average is difficult to do because the logarithm appears inside the average. The replica trick circumvents this difficulty by exploiting the identity

$$\ln Z = \lim_{n\to 0} \frac{Z^n - 1}{n} = \lim_{n\to 0} \frac{\partial}{\partial n} Z^n. \tag{7}$$

This identity is useful because it allows us to first average over an integer power of $Z[\mathbf{J}]$, which can be done more easily, and then take the $n \to 0$ limit. Appendix 8 provides a general outline of the replica approach that can be used for many problems. Basically to compute the average over Z^n it is useful to introduce n replicated neuronal activity patterns $\mathbf{s}^a$, for $a = 1, \ldots, n$, yielding

$$\langle\langle\, Z^n\,\rangle\rangle_{\mathbf{J}} = \left\langle\!\!\left\langle\, \sum_{\{\mathbf{S}^a\}} e^{\beta \sum_{a=1}^n \sum_{ij} J_{ij} s_i^a s_j^a} \,\right\rangle\!\!\right\rangle_{\mathbf{J}}. \tag{8}$$

Now the average over $\mathbf{J}$ can be performed because it is reduced to a set of Gaussian integrals. To do so, we use the fundamental identity

$$\langle e^{zx}\rangle_z = e^{\frac{1}{2}\sigma^2 x^2}, \tag{9}$$

where z is a zero mean Gaussian random variable with variance σ^2. Applying this to (8) with $z = J_{ij}$, $\sigma^2 = \frac{1}{N}$, and $x = \beta \sum_a s_i^a s_j^a$ yields

$$\langle\langle\, Z^n\,\rangle\rangle_{\mathbf{J}} = \sum_{\{\mathbf{S}^a\}} e^{\frac{1}{4N}\sum_{ij}(\sum_{a=1}^n s_i^a s_j^a)^2} = \sum_{\{\mathbf{S}^a\}} e^{N\frac{\beta^2}{4}\sum_{ab} Q_{ab}^2}, \tag{10}$$

where

$$Q_{ab} = \frac{1}{N} \sum_{i=1}^{N} s_i^a s_i^b \tag{11}$$

is the overlap matrix between replicated activity patterns.

Thus although for any fixed realization of the quenched disorder $\mathbf{J}$, the replicated activity patterns $\mathbf{s}^a$ were independent, marginalizing over, or integrating out the disorder introduces attractive interactions between the replicas. Consistent with the general framework, presented in section 8, the interaction between replicas depends only on the overlap matrix Q, and we have in (120), $E(Q) = -\frac{\beta^2}{4} \sum_{ab} Q_{ab}^2$. Thus minimization of this energy function promotes alignment of the replicas. The intuition is that for

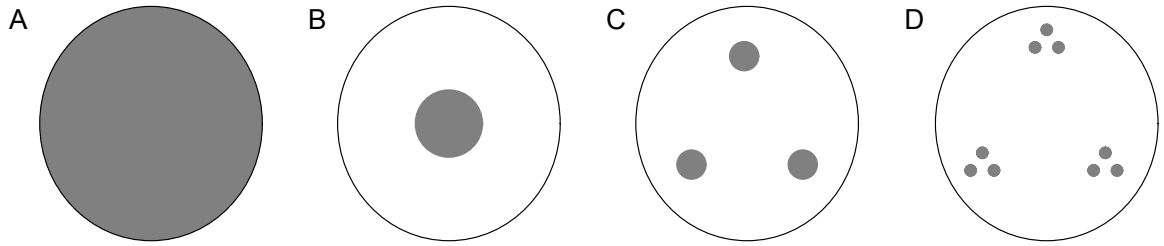

Figure 1. Probability lumps in free energy valleys. Schematic figures of the space of all possible neuronal or spin configurations (large circle) and the space of spin configurations with non-negligible probability under the Gibbs distribution in (2) (shaded areas). (A) At high temperature all spin configurations are explored by the Gibbs distribution. Thus the inner product between two random spins drawn from the Gibbs distribution will typically have 0 inner product, and so the replica order parameter q is 0. (B) The replica symmetric ansatz for a low temperature phase: the spins freeze into a small set of configurations (free energy valley), which can differ from realization to realization of the connectivity $\mathbf{J}$. However, the inner product between two random spins, and therefore also the replica order parameter, takes a nonzero value q that does not depend on the realization of $\mathbf{J}$. (C) One possible ansatz for replica symmetry breaking (RSB) in which the replica overlap matrix Q is characterized by two order parameters, $q_1 > q_2$. This ansatz, known as 1-step RSB, corresponds to a scenario in which the Gibbs distribution breaks into multiple lumps, with q_1 describing the typical inner product between two configurations chosen from the same lump, and q_2 describing the typical inner product between configurations from different lumps. (D) There exists a series of k-step RSB schemes describing scenarios in which the Gibbs distribution decomposes into a nested hierarchy of lumps of depth k. This figure describes a possible scenario for $k = 2$. The true low temperature phase of the SK model is thought to be described by a particular $k = \infty$ RSB ansatz [7].

any fixed realization of $\mathbf{J}$, the replicas will prefer certain patterns. Which patterns are preferred will vary across realizations of $\mathbf{J}$. However, for any fixed realization of $\mathbf{J}$, the preferred set of patterns will be similar across replicas since the fluctuations of each replicated neuronal activity pattern are controlled by the same quenched connectivity $\mathbf{J}$. Thus even after averaging over $\mathbf{J}$, we expect this similarity to survive, and hence we expect average overlaps between replicas to be nonzero.

However, minimization of the energy $E(Q)$ alone does not determine the overlap matrix Q_{ab}. One must still sum over $\mathbf{s}^a$ in (10) which yields an entropic term

corresponding to the number of replicated activity patterns with a given set of overlaps. While energy minimization drives overlaps to be large, entropy maximization drives overlaps to be small, since there are many more replicated configurations with small, rather than large overlaps. This competition between energy and entropy leads to a potentially nontrivial overlap matrix. After computing this entropic term, the most likely value of the overlap matrix can be computed via the saddle point method, yielding a set of self-consistent equations for Q (a special case of (126),(127)):

$$Q_{ab} = \langle s^a s^b \rangle_n, \tag{12}$$

where $\langle \cdot \rangle_n$ denotes an average with respect to the Gibbs distribution $P(s^1, \ldots, s^n) = \frac{1}{Z} e^{-\beta H_{\rm eff}}$, with $H_{\rm eff} = -\beta \sum_{ab} s^a Q_{ab} s^b$.

Now the physical meaning of the saddle point replica overlap matrix is explained in 8.2; it is simply related to the disorder averaged overlap distribution:

$$\langle\langle\, P_{\mathbf{J}}(q)\,\rangle\rangle_{\mathbf{J}} = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{a\neq b} \delta(q - Q_{ab}), \tag{13}$$

where $P_{\mathbf{J}}(q)$ is given by (5). So the distribution of overlaps between pairs of free energy minima m_i^a (weighted by their probability), is simply the distribution of off-diagonal matrix elements of the replica overlap matrix. Thus, in searching for solutions to (12), any ansatz about the structure of Q_{ab} is implicitly an ansatz about the geometry and multiplicity of free energy valleys in (2), averaged over $\mathbf{J}$.

Now the effective Hamiltonian yielding the average in (12) is symmetric with respect to permutations of the replica indices a (i.e. permuting the rows and columns of Q_{ab}). Therefore it is natural to search for a replica symmetric saddle point in which $Q_{ab} = q$ for all $a \neq b$. This is equivalent to an assumption that there is only one free energy valley, and q measures its heterogeneity. Taking the $n \to 0$ limit with this replica symmetric ansatz yields a saddle point equation for q (see (142) for the derivation):

$$q = \left\langle\!\left\langle \tanh^2(\beta\sqrt{q}z) \right\rangle\!\right\rangle_z. \tag{14}$$

At high temperature ($\beta < 1$), $q = 0$ is the only solution, representing a "paramagnetic" state (Fig. 1A) in which activity patterns fluctuate over all possible configurations, and average neural activity m_i is 0 for all i (Fig. 1A). At lower temperature ($\beta > 1$), a nonzero solution rises continuously from 0, suggesting a phase transition to a "frozen" state corresponding to a single valley (Fig. 1B) in which each neuron has a different mean activity m_i.

While this scenario seems plausible, a further analysis of this solution [6, 62] yields inconsistent physical predictions (like negative entropy for the system). Within the replica framework, this inconsistency can be detected by showing that the replica symmetric saddle point for Q_{ab} is unstable [63], and so one must search for solutions in which Q_{ab} breaks the replica symmetry. This corresponds to a physical picture in which there are many free energy minima. A great deal of work has lead to a remarkably rich ansatz which predicts a nested hierarchical, tree like organization on the space of free energy minima (see Fig. 1CD), known as an ultrametric structure [64]. It is striking

that this highly symmetric and ordered low temperature hierarchical structure emerges generically from purely random, disordered couplings $\mathbf{J}_{ij}$. Unfortunately, we will not explore this phenomenon further here, since for most of the applications of replica theory to neuronal processing and data analysis discussed below, a replica symmetric analysis turns out to be correct.

2.2. Chaos in the SK Model and the Hopfield Solution

So far, in order to introduce the replica method, we have analyzed a toy neuronal network with a random symmetric connectivity matrix $\mathbf{J}$, and found that such a network exhibits broken replica symmetry corresponding to a hierarchy of low energy states that are stable with respect to thermal or noise induced fluctuations. It is tempting to explore the possibility that this multiplicity of states may be useful for performing neural information processing tasks. However, several works have noted that while these states are stable with respect to thermal fluctuations, they are not structurally stable with respect to perturbations either to the inverse temperature β, or the connectivity matrix $\mathbf{J}$ [65, 66, 67]. Indeed very small changes to β or $\mathbf{J}$ induce macroscopic changes in the location of energy minima in the space of neuronal activity patterns. This sensitive dependence of low energy activity patterns to either β or $\mathbf{J}$ was called temperature or disorder chaos respectively in [66]. For neural information processing, it would be useful to instead have network connectivities whose noisy dynamics not only thermally stabilize a prescribed set of neuronal activity patterns, but do so in a manner that is structurally stable with respect to changes in either the connectivity or level of noise.

An early proposal to do just this was the Hopfield model [68]. Suppose one wishes to find a network connectivity $\mathbf{J}$ that stabilizes a prescribed set of P N-dimensional patterns $\boldsymbol{\xi}^\mu$, for $\mu = 1, \ldots, P$, where $\boldsymbol{\xi}^\mu_i = \pm 1$. Hopfield's proposal was to choose

$$\mathbf{J}_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi^\mu_i \xi^\mu_j. \tag{15}$$

This choice reflects the outcome of a Hebbian learning rule [69] in which each synapse from neuron j to neuron i changes its synaptic weight by an amount proportional to the correlation between the activity on its presynaptic and postsynaptic neuron. When the activity pattern $\boldsymbol{\xi}^\mu$ is imposed upon the network, this correlation is $\xi^\mu_i \xi^\mu_j$, and when all P patterns are imposed upon the network in succession, the learned synaptic weights are given by (15).

This synaptic connectivity $\mathbf{J}$ induces an equilibrium probability distribution over neuronal activity patterns $\mathbf{s}$ through (2). Ideally, this distribution should have $2P$ free energy valleys, corresponding to lumps of probability mass located near the P patterns $\boldsymbol{\xi}^\mu$ and their reflections $-\boldsymbol{\xi}^\mu$. If so, then when network activity $\mathbf{s}$ is initialized to either a corrupted or partial version of one of the learned patterns $\boldsymbol{\xi}^\mu$, the network will relax (under a dynamics whose stationary distribution is given by (2)) to the free energy valley corresponding to $\boldsymbol{\xi}^\mu$. This relaxation process is often called pattern completion. Thus Hopfield's prescription provided a unifying framework for thinking about learning

and memory: the structure of past experience (i.e. the patterns $\boldsymbol{\xi}^\mu$) are learned, or stored, in the network's synaptic weights (i.e. through (15)), and subsequent network dynamics can be viewed as motion down a free energy landscape determined by the weights. If learning is successful, the minima of this free energy landscape correspond to past experiences, and the process of recalling past experience corresponds to completing partial or corrupted initial network activity patterns induced by current stimuli.

A key issue then is storage capacity: how many patterns P can a network of N neurons store? This issue was addressed in [70, 71] via the replica method in the situation where the stored patterns $\boldsymbol{\xi}^\mu$ are random and uncorrelated (each ξ_i^μ is chosen independently to be $+1$ or -1 with equal probability). These works extensively analyzed the properties of free energy valleys in the Gibbs distribution (2) with connectivity (15), as a function of the inverse temperature β and the level of storage saturation $\alpha = \frac{P}{N}$. This problem fits the classic mold of disordered statistical physics, where the patterns $\boldsymbol{\xi}^\mu$ play the role of quenched disorder, and neuronal activity patterns play the role of thermal degrees of freedom. In particular, the structure of free energy minima can be described by a collection of self-averaging order parameters $m^\mu = \frac{1}{N}\, \boldsymbol{\xi}^\mu \cdot \mathbf{s}$, denoting the overlap of neuronal activity with pattern μ. Successful pattern completion is possible if there are $2P$ free energy valleys such that the average of m^μ in each valley is large for one pattern μ and small for all the rest. These free energy valleys can be thought of as recall states. The replica method in [70, 71] yields a set of self-consistent equations for these averages. Solutions to the replica equations, in which precisely one order parameter m^μ is large, are found at low temperature only when $\alpha < \alpha_c = 0.138$. For $\alpha > \alpha_c$, the system is in a spin glass state with many free energy minima, none of which have a macroscopic overlap with any of the patterns (in the solutions to the replica equations, no average m^μ is O(1) as $P, N \to \infty$ with $\alpha > \alpha_c$). At such high levels of storage, so many patterns "confuse" the network, so that its low energy states do not look like any one pattern $\boldsymbol{\xi}^\mu$. Indeed the free energy landscape of the Hopfield model as α becomes large behaves like the low temperature spin glass phase of the SK model discussed in the previous section.

Even for $\alpha < \alpha_c$, at low enough temperatures, spurious, metastable free energy valleys corresponding to mixtures of patterns can also arise. These mixture states are characterized by solutions to the replica equations in which the average m^μ is $O(1)$ for more than one μ. However, as the temperature is increased, such mixture states melt away. This phenomenon illustrates a beneficial role for noise in associative memory operation. However, there is a tradeoff to melting away mixture states by increasing temperature, as α_c decreases with increasing temperature. Nevertheless, in summary there is a robust region in the $\alpha - \beta$ phase plane with $\alpha = O(0.1)$ and β corresponding to low temperatures, in which the recall states dominate the free energy landscape over neural activity patterns, and the network can successfully operate as a pattern completion, or associative memory device. Many important details about the phase diagram of free energy valleys as a function of α and β and be found in [70, 71].

2.3. Cavity Method

We now return to an analysis of the SK model through an alternate method that sheds light on the physical meaning of the saddle point equation for the replica symmetric order parameter q in (14), which may seem a bit obscure. In particular, we give an alternate derivation of (14) through the cavity method [7, 72] which provides considerable physical intuition for (14) by describing it as a self-consistency condition. In general, the cavity method, while indirect, can often provide intuition for the final results derived via more direct replica methods.

The starting point involves noting that the SK Hamiltonian (1) governing the fluctuations of N neurons can be written as

$$H_N(\mathbf{s}, \mathbf{J}) = -s_1 h_1 + H_{\setminus 1}, \tag{16}$$

where

$$h_1 = \sum_{i=2}^{N} \mathbf{J}_{1i} s_i \tag{17}$$

is the local field acting on neuron 1, and

$$H_{\setminus 1} = -\frac{1}{2} \sum_{ij=2}^{N} \mathbf{J}_{ij} s_i s_j, \tag{18}$$

is the Hamiltonian of the rest of the neurons $s_2, \ldots, s_N$. Since h_1 is a sum of many terms, it is tempting to approximate its thermal fluctuations in the full system of N neurons in (16) by a Gaussian distribution. However, such a Gaussian approximation is generally invalid because the individual terms are correlated with each other. One source of correlation arises from a common coupling of all the neurons $s_2, \ldots, s_N$ to s_1. For example, because s_1 interacts with s_i through the symmetric coupling $\mathbf{J}_{1i} = \mathbf{J}_{i1}$, whenever $s_1 = +1$ (or $s_1 = -1$) this exerts a positive (or negative) effect on the combination $\mathbf{J}_{1i} s_i$. Thus all individual terms in (17) exhibit correlated fluctuations due to common coupling to the fluctuating neuron s_1.

The key idea behind the cavity method is to consider not the distribution of the local field h_1 acting on neuron 1 in the full system of N neurons in (16), but instead the distribution of h_1 in a "cavity system" of $N-1$ neurons obtained by removing s_1 from the system, thereby leaving "cavity" (see Fig. 2AB). Then h_1 is known as the cavity field, or the field exerted on neuron 1 by all the others in the *absence* of neuron 1, and its distribution is given by that of h_1 (17) in a Gibbs distribution with respect to (18):

$$P_{\setminus 1}(h_1) = \frac{1}{Z_{\setminus 1}} \sum_{s_2, \ldots, s_N} \delta(h_1 - \sum_{i=2}^{N} \mathbf{J}_{1i} s_i)\, e^{-\beta H_{\setminus 1}} \tag{19}$$

The joint distribution of s_1 and its local field h_1 in the full system of N spins can be written in terms of the cavity field distribution as follows:

$$\begin{aligned} P_N(s_1, h_1) &= \frac{1}{Z_N} \sum_{s_2, \ldots, s_N} \delta(h_1 - \sum_{i=2}^{N} \mathbf{J}_{1i} s_i)\, e^{-\beta H_N} \\ &= \frac{1}{Z} e^{-\beta V(s_1, h_1)} P_{\setminus 1}(h_1), \end{aligned} \tag{20}$$

where $V(s_1, h_1) = -s_1 h_1$.

The advantage of writing the joint distribution of s_1 and h_1 in terms of the cavity field distribution $P_{\backslash 1}(h_1)$ is that one can now plausibly make a Gaussian approximation to $P_{\backslash 1}(h_1)$, i.e. the distribution of (17) in the cavity system (18) of neurons $2, \ldots, N$ in the absence of 1. Because the cavity system does not couple to neuron 1, it does not know about the set of couplings $\mathbf{J}_{1i}$, and therefore the thermal fluctuations of cavity activity

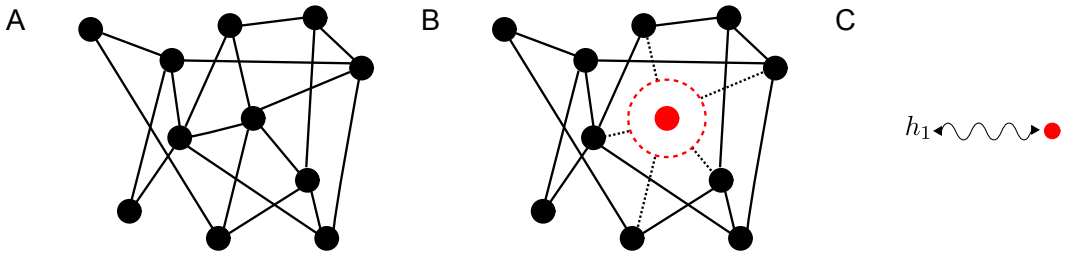

Figure 2. The cavity Method. (A) A network of neurons, or spins. (B) A cavity surrounding a single neuron, $\mathbf{s}_1$, that has been removed from the system. (C) In a replica symmetric approximation, the full distribution of the field h_1 exerted on the cavity (in the absence of $\mathbf{s}_1$) by all other neurons can be approximated by a Gaussian distribution, while the joint distribution of $\mathbf{s}_1$ and h_1 takes the form in equation (20).

patterns $\mathbf{s}_2, \ldots, s_N$, while of course correlated with each other, must be uncorrelated with the couplings $\mathbf{J}_{1i}$, unlike the case of these same fluctuations in the presence of s_1. Motivated by this lack of correlation, we can make a Gaussian approximation to the thermal fluctuations of h_1 in the cavity system $H_{\backslash 1}$. Note that this does not imply that the local field h_1 in the full system H_N is Gaussian. Indeed, if $P_{\backslash 1}(h_1)$ in (20) is Gaussian, then $P_N(h_1)$ obtained by marginalizing out s_1 in $P_N(s_1, h_1)$ cannot be Gaussian; as discussed above, this non-gaussianity arises due to positive correlations between s_1 and h_1 induced by their coupling $V(s_1, h_1)$. The simplification in replacing the network with a fluctuating field is shown in the transition from Fig.2B to 2C.

Under a Gaussian approximation, $P_{\backslash 1}(h_1)$ is characterized by its mean

$$\langle h_1 \rangle_{\backslash 1} = \sum_{i=2}^{N} \mathbf{J}_{1i} \langle s_i \rangle_{\backslash 1}, \tag{21}$$

and variance

$$\langle (\delta h_1)^2 \rangle_{\backslash 1} = \sum_{i,j=2}^{N} \mathbf{J}_{1i} \mathbf{J}_{1j} \langle \delta s_i \delta s_j \rangle_{\backslash 1} \tag{22}$$

$$= \sum_{i=2}^{N} \mathbf{J}_{1i}^2 \langle (\delta s_i)^2 \rangle_{\backslash 1} \tag{23}$$

$$= 1 - \sum_{i=2}^{N} \frac{1}{N} \langle s_i \rangle_{\backslash 1}^2$$

$$= 1 - q \tag{24}$$

where q is the order parameter

$$q = \frac{1}{N}\sum_{i=1}^{N} \langle s_i \rangle_N^2, \tag{25}$$

and $\delta s_i = s_i - \langle s_i \rangle_{\setminus 1}$. Here we have neglected various terms that vanish in the large N limit, but most importantly, in going from (22) to (23), we have made a strong assumption that the connected correlation $\langle \delta s_i \delta s_j \rangle_{\setminus 1}$ vanishes in the large N limit fast enough that we can neglect all off-diagonal terms in (22). This can be true if the cavity system (and consequently the full system) is accurately described by a single free energy valley. On the otherhand, if the system is described by multiple free energy valleys, the connected correlation will receive contributions from fluctuations across valleys, and we cannot neglect the off-diagonal terms [7]. Thus the validity of this cavity approximation is tantamount to an assumption of replica symmetry, or a single valley in the free energy landscape. As discussed above, under the assumption of a single valley, we expect q to be self-averaging: it does not depend on the detailed realization of $\mathbf{J}_{ij}$ in the large N limit. Finally, we note that the cavity method can be extended to scenarios in which replica symmetry is broken and there are multiple valleys [7].

In the replica symmetric scenario, under the Gaussian approximation to $P_{\setminus 1}(h_1)$, (20) becomes

$$P_N(s_1, h_1) = \frac{1}{Z[\langle h_1 \rangle_{\setminus 1}, 1-q]} e^{-\beta V(s_1,h_1) - \frac{1}{2(1-q)}(h_1 - \langle h_1 \rangle_{\setminus 1})^2}, \tag{26}$$

allowing us to compute the mean activity of neuron i in the full system of N neurons, in terms of the mean $\langle h_1 \rangle_{\setminus 1}$ and variance $1-q$ of its cavity field,

$$\langle s_1 \,|\, \langle h_1 \rangle_{\setminus 1}, 1-q \,\rangle_N = \sum_{s_1} s_1 P_N(s_1, h_1). \tag{27}$$

But now we must compute q, which we can do by demanding self consistency of the cavity approximation. First of all, we note that there was nothing special about neuron 1; the above procedure of forming a cavity system by removing neuron 1 could have been done with any neuron. Thus (26, 27) holds individually for all neurons i, and we can average these equations over all i to obtain an expression for q:

$$q = \frac{1}{N}\sum_{i=1}^{N} \langle s_i \,|\, \langle h_i \rangle_{\setminus i}, 1-q \,\rangle_N^2. \tag{28}$$

However, we do not yet know $\langle h_i \rangle_{\setminus i}$ for each i. For each i, $\langle h_i \rangle_{\setminus i} = \sum_{k \neq i} \mathbf{J}_{ik} \langle s_k \rangle_{\setminus i}$ is a random variable due to the randomness of the couplings $\mathbf{J}_{ik}$, which are uncorrelated with $\langle s_k \rangle_{\setminus i}$ by virtue of the fact that this thermal average occurs in a cavity system in the absence of i. Thus we expect the distribution of $\langle h_i \rangle_{\setminus i}$ over random realizations of $\mathbf{J}_{ik}$ to be gaussian, with a mean and variance that are easily computed to be 0 and q respectively. Furthermore, we expect this distribution to be self-averaging: i.e. the distribution of $\langle h_i \rangle_{\setminus i}$ for a fixed i across different realizations of $\mathbf{J}$ should be the same as the distribution of $\langle h_i \rangle_{\setminus i}$ across different neurons i for a fixed realization of $\mathbf{J}$, in the large N limit. Under this assumption, although we may not know each individual

$\langle h_i \rangle_{\setminus i}$, we can replace the average over neurons in (28) with an average over a Gaussian distribution, yielding

$$q = \left\langle \left\langle \, \langle s_i \,|\, \sqrt{q}z, 1-q \, \rangle_N^2 \, \right\rangle \right\rangle_z. \tag{29}$$

Here $\langle\langle \cdot \rangle\rangle_z$ denotes a "quenched" average with respect to a zero mean unit variance Gaussian variable z, reflecting the heterogeneity of the mean cavity field across neurons, and the thermal average $\langle \cdot \rangle_N$ is computed via (26, 27), and reflects the thermal fluctuations of a single neuron in the presence of a cavity field with mean and variance $\sqrt{q}z$ and $1-q$, respectively.

Equation (29) is a self-consistent equation for the order parameter q which is itself a measure of the heterogeneity of mean activity across neurons. So physically, (29) reflects a demand that the statistical properties of the cavity fields are consistent with the heterogeneity of mean neural activity. Now finally, we can specialize to the SK model in which $V(s,h) = -sh$ in (26), which yields $\langle s_1 \,|\, \sqrt{q}z, 1-q \, \rangle_N = \tanh \beta\sqrt{q}z$ in (27), and when this is substituted into (29), we recover the self-consistent equation for q in (14), derived via the replica method.

2.4. Message Passing

So far, we have seen two methods which allow us to calculate self-averaging quantities (for example $q = \frac{1}{N}\sum_i \langle s_i \rangle^2$) that do not depend on the detailed realization $\mathbf{J}$. However, we may wish to understand the detailed pattern of mean neural activity, i.e. $\langle s_i \rangle$ for all i, for some fixed realization of $\mathbf{J}_{ij}$. Mathematically, this corresponds to computing the marginal distribution of a single neuron in a full joint distribution given by (2). Here we introduce efficient distributed message passing algorithms from computer science [5, 11, 10] that have been developed to compute such marginals in probability distributions which obey certain factorization properties.

Consider for example a joint distribution over N variables $x_1, \cdots, x_N$ that factorizes into a set of P factors, or interactions, indexed by $a = 1, \ldots P$:

$$P(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{a=1}^{P} \psi_a(x_a). \tag{30}$$

Here x_i is any arbitrary variable that could be either continuous or discrete, and x_a denotes the collection of variables that factor a depends on. Thus we systematically abuse notation and think of each factor index a also as a subset of the N variables, with variable $i \in a$ if and only if factor ψ_a depends on x_i. The factorization properties of P can be visualized in a factor graph, which is a bipartite graph whose nodes correspond either to variables i or factors a, and there is an edge between a factor a and variable i if and only if $i \in a$, or equivalently factor ψ_a depends on x_i (see Fig. 3A). For example, the SK model, or more generally any neural system with an equilibrium distribution, corresponds to a factor graph in which the neurons s_i are the variables x_i, and the factors correspond to nonzero synaptic weights connecting pairs of neurons. Thus each a corresponds to a neuron pair $a = (ij)$, and in the SK model of equation (2), $\psi_a(s_i, s_j) = e^{\beta \mathbf{J}_{ij} s_i s_j}$.

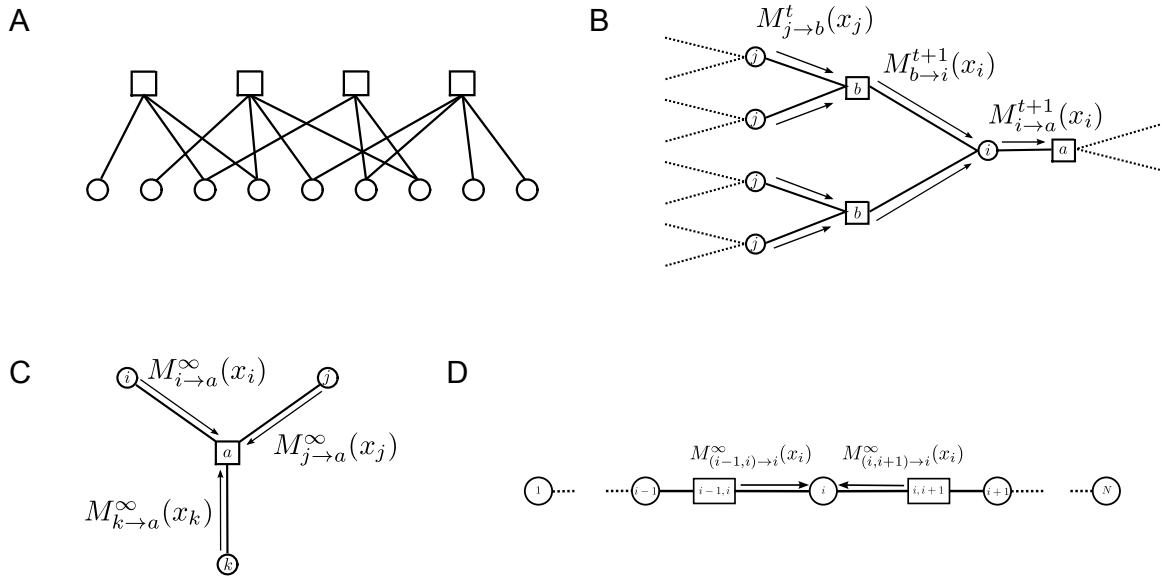

Figure 3. Message Passing. (A) A factor graph in which variable nodes are represented by circles, and factor nodes are represented by squares. (B) The flow of messages involved in the update of the message $M^{t+1}_{i\to a}(x_i)$. (C) The message passing approximation to the joint distribution of x_i, x_j, and x_k. Here the interaction a is treated exactly, while the effect of all other interactions besides a are approximated by a product of messages. (D) Exact message passing in a chain; the marginal on x_i is computed exactly as a product of two messages.

The utility of the factor graph representation is that an iterative algorithm to compute the marginals,

$$P(x_i) = \sum_{\{x_j\}, j\neq i} P(x_1, \ldots, x_N) \tag{31}$$

for all i can be visualized as the flow of messages along the factor graph (3B). We first define this iterative algorithm and then later give justification for it. Every message is a probability distribution over a single variable, and at any given time t there are two types of messages, one from variables to factors, and the other from factors to variables. We denote by $M^t_{j\to b}(x_j)$ the message from variable j to factor b, and by $M^t_{b\to i}(x_i)$ the message from factor b to variable i. Intuitively, we can think of $M^t_{j\to b}(x_j)$ as an approximation to distribution on x_j induced by all other interactions besides interaction b. In contrast, we can think of $M^t_{b\to i}(x_i)$ as an approximation to the distribution on x_i induced by the direct influence of interaction b alone. These messages will be used below to approximate the marginal of x_i in the full joint distribution of all interactions (see e.g. (34)).

The (unnormalized) update equation for a factor to variable message is given by

$$M^{t+1}_{b\to i}(x_i) = \sum_{x_{b\setminus i}} \psi_b(x_b) \prod_{j\in b\setminus i} M^t_{j\to b}(x_j), \tag{32}$$

where $b\setminus i$ denotes the set of all variables connected to factor node b except i (see Fig. 3b). Intuitively, the direct influence of b alone on i (the left hand side of (32)) is obtained by marginalizing out all variables other than i in the factor ψ_b, supplemented

by accounting for the effect of all of the other interactions besides b on variables $j \in b \backslash i$ by the product of messages $M^t_{j\to b}(x_j)$ (see Fig. 3b). The (unnormalized) update equation for the variable to factor messages is then given by

$$M^{t+1}_{i\to a}(x_i) = \prod_{b\in i\backslash a} M^{t+1}_{b\to i}(x_i), \tag{33}$$

Intuitively, the distribution on x_i induced by all other interactions besides interaction a (the left hand side of (33)) is simply the product of the direct influences of all interactions b that involve variable i, except for interaction a (see Fig. 3b). Message passing involves randomly initializing all the messages and then iteratively running the update equations (32,33) until convergence. One exception to the random initialization is the situation where any variable i is connected to only 1 variable node a. In this case, $M_{i\to a}(x_i)$ is initialized to be a uniform distribution over x_i, since in the absence of a, variable i feels no influence from the rest of the graph. Under the message passing dynamics, $M_{i\to a}(x_i)$ will remain a uniform distribution. Now for general factor graphs, convergence is not guaranteed, but if the algorithm does converge, then the marginal distribution of a variable x_i can be approximated via

$$P(x_i) \propto \prod_{a\in i} M^{\infty}_{a\to i}(x_i). \tag{34}$$

and indeed the the joint distribution of all variables $i \in a$ can be approximated via

$$P(x_a) \propto \psi_a(x_a) \prod_{i\in a} M^{\infty}_{i\to a}(x_i). \tag{35}$$

The update equations (32,32), while intuitive, lead to two natural questions: for which factor graphs will they converge, and if they converge, how well will the fixed point messages $M^{\infty}_{a\to i}$ and $M^{\infty}_{i\to a}$ approximate the true marginals though equations (34,35)? A key intuition arises from the structure of the approximation to the joint marginal of the variables x_a in (35) (see also Fig. 3C). This approximation treats the coupling of the variables $i \in a$ through interaction a by explicitly including the factor ψ_a. However, it approximates the effects of all other interactions b on these variables by a simple product of messages $M^{\infty}_{i\to a}(x_i)$. An exactly analogous approximation is made in the update equation (32). Such approximations might be expected to work well whenever removing the interaction a leads to a factor graph in which all the variables i that were previously connected to a are now weakly coupled (ideally independent) under all the remaining interactions $b \neq a$.

This weak coupling assumption under the removal of a single interaction holds exactly whenever the factor graph is a tree, with no loops. Indeed in such a case, removing any one interaction a removes all paths through the factor graph between variables $i \in a$. In the absence of any such paths, all pairs of variables $i \in a$ are independent, and their joint distribution factorizes, consistent with the approximations made in (32) and (35). In general, whenever the factor graph is a tree, the message passing equations converge in finite time, and the fixed point messages yield the true marginals [11]. We will not give a general proof of this fact, but we will illustrate it in the

case of a one dimensional Ising chain (see Fig. 3D). Consider the marginal distribution of a spin at position i in the chain. This spin feels an interaction to its left and right, and so (35) tells us the marginal is a product of two converged messages at time $t = \infty$:

$$P(s_i) \propto M^{\infty}_{(i-1,i)\to i}(s_i)\, M^{\infty}_{(i,i+1)\to i}(s_i), \tag{36}$$

Each of these two messages can be computed by iterating messages from either end of the chain to position i. For example, the rightward iteration for computing $M^{\infty}_{(i-1,i)\to i}(s_i)$ is

$$M^{t+1}_{k\to(k,k+1)}(s_k) = M^{t+1}_{(k-1,k)\to k}(s_k) = \sum_{s_{k-1}} e^{\beta \mathbf{J}_{k,k-1} s_k\, s_{k-1}} M^{t}_{k-1\to(k-1,k)}(s_{k-1}) \tag{37}$$

where the first equality is a special case of (33) and the second is a special case of (32). The first message in this iteration, $M^0_{1\to(1,2)}(s_1)$ is initialized to be a uniform distribution, since spin 1 is only connected to a single interaction $(1, 2)$. A similar leftward iteration leads to the calculation of $M^{\infty}_{(i,i+1)\to i}(s_i)$. Each iteration converges in an amount of time given by the path length from each corresponding end to i, and after convergence, we have

$$M^{\infty}_{(i-1,i)\to i}(s_i) = \sum_{s_1,\ldots,s_{i-1}} e^{\beta \sum_{k=1}^{i-1} \mathbf{J}_{k,k+1} s_k s_{k+1}} \tag{38}$$

$$M^{\infty}_{(i,i+1)\to i}(s_i) = \sum_{s_{i+1},\ldots,s_N} e^{\beta \sum_{k=i}^{N-1} \mathbf{J}_{k,k+1} s_k s_{k+1}}. \tag{39}$$

Inserting (38) and (39) into (36) yields the correct marginal for s_i, and the normalization factor can be fixed at the end by demanding $P(+1) + P(-1) = 1$. Note that whereas a naive sum over all spin configurations to compute the marginal over s_i would require $O(2^N)$ operations, this iterative procedure for computing the marginal requires only $O(N)$ operations. Moreover, two sweeps through the chain allow us to compute all the messages, and therefore all N marginals simultaneously, as (36) holds for all i. Overall, this method is essentially identical to the transfer matrix method for the $1D$ Ising chain, and is a generalization of the Bethe approximation [73].

Although message passing is only exact on trees, it can nevertheless be applied to graphical models with loops, and as discussed above, it should yield good approximate marginals whenever the variables adjacent to a factor node are weakly correlated upon removal of that factor node. We will see successful examples of this in the contexts of learning in section 3.6 and compressed sensing in 6.3. An early theoretical advance in partially justifying the application of message passing to graphical models with loops was a variational connection: each solution to the fixed point equations of message passing are in one to one correspondence with extrema of a certain Bethe free energy [74], an approximation to the Gibbs free energy in variational approaches to inference in graphical models that is exact on trees (see [75] for a review). However there are no known general and precise conditions under which message passing in graphical models with many loops is theoretically guaranteed to converge to messages that yield a good approximation to the marginals. Nevertheless, in practice, message passing seems

to achieve empirical success in approximating marginals when correlations between variables adjacent to a factor node are indeed weak after removal of that factor.

We conclude this section by connecting message passing back to the replica method. In general, suitable averages of the messaging passing equations reduce to both the cavity equations and the replica equations [5]. To illustrate this in the special case of the SK model, we outline the derivation of the replica saddle point equation (14) from the perspective of message passing. We first note that since every factor node in the SK model has degree 2, the update of a message from a factor (i, j) to variable j, i.e. $M_{(i,j)\to j}(s_j)$ depends only on the message $M_{i\to(i,j)}(s_i)$ through,

$$M^{t+1}_{(i,j)\to j}(s_j) = \sum_{s_i} e^{\beta \mathbf{J}_{ij} s_i s_j} M^t_{i\to(i,j)}(s_i), \tag{40}$$

which is a special case of (32). Thus we can take one set of messages, for example the node to factor messages, $M^t_{i\to(i,j)}(s_i)$, as the essential degrees of freedom upon which the message passing dynamics operates. We simplify notation a little by letting $M^t_{i\to j}(s_i) \equiv M^t_{i\to(i,j)}(s_i)$. Then the remaining message passing update (33) yields the dynamics

$$M^{t+1}_{i\to j}(s_i) = \prod_{k\in i\setminus j} \sum_{s_k} e^{\beta \mathbf{J}_{ik} s_i s_k} M^t_{k\to i}(s_k), \tag{41}$$

where $k \in i$ if and only if s_k is coupled to s_i through a nonzero $\mathbf{J}_{ik}$.

Now each message is a distribution over a binary variable, and all such distributions can be usefully parameterized by a single scalar parameter:

$$M^t_{i\to j}(s_i) \propto e^{\beta h^t_{i\to j} s_i}. \tag{42}$$

Here the scalar parameter $h^t_{i\to j}$ can be thought of as a type of cavity field; as $t \to \infty$, if message passing is successful, $h^t_{i\to j}$ converges to the field exerted on spin i by all the spins in a cavity system in which the interaction $\mathbf{J}_{ij}$ is removed. In terms of this parameterization, the message passing updates (41) yield a dynamical system on the cavity fields [76],

$$h^{t+1}_{i\to j} = \sum_{k\in i\setminus j} u(\mathbf{J}_{ki}, h^t_{k\to i}). \tag{43}$$

Here $u(J, h)$ is defined implicitly through the relation

$$e^{\beta u(J,h) s} \propto \sum_{s'} e^{\beta J s s' + \beta h s'}. \tag{44}$$

Physically $u(J, h)$ is the effective field on a binary spin s coupled with strength J to another spin s' that experiences an external field of strength h, after marginalizing out s'. Explicitly,

$$u(J, h) = \frac{1}{\beta} \text{arctanh}\left[\tanh(\beta J)\tanh(\beta h)\right]. \tag{45}$$

In the weak coupling limit of small J, $u(J, h) \approx J\tanh(\beta h)$, which reflects the simple approximation that the average magnetization of $\mathbf{s}'$, due to the external field h (which would be $\tanh(\beta h)$ if the back-reaction from s can be neglected), exerts a field $J\tanh(\beta h)$

on s. The more complex form of $u(J,h)$ in (45) reflects the back-reaction of s on s' that becomes non-negligible at larger values of the bi-directional coupling J. In (43), the updated cavity field $h_{i\to j}^{t+1}$ turns out to be a simple sum over all the spins k (besides j) of this same effective field u obtained by marginalizing out s_k in the presence of its own cavity field $h_{k\to i}^{t}$.

Using (43), we are now ready to derive (14). The key point is to consider self-consistency conditions for the distribution of cavity fields $h_{i\to j}^{t}$ as $t\to\infty$. We can think of this distribution in two ways. First, for a fixed i and j, $h_{i\to j}^{\infty}$ is a random variable due to the random choice of couplings $\mathbf{J}$. Second, for a fixed realization of $\mathbf{J}$, at a message passing fixed point, there is an empirical distribution of cavity fields $h_{i\to j}^{\infty}$ across all choices of pairs i and j. The assumption of self-averaging means that as $N\to\infty$, the latter empirical distribution converges to the distribution of the former random variable. In any case, we would like to write down a self-consistent equation for this distribution, by observing that this distribution must be self-reproducing under the update equation (43). More precisely, in (43), if the couplings J_{ik} are drawn i.i.d from a distribution $P(J)$, and the cavity fields $h_{k\to i}^{t}$ are drawn i.i.d. from a distribution $Q(h)$, then the induced distribution on $h_{i\to j}^{t+1}$ should be identical to $Q(h)$. This yields a recursive distributional equation characterizing the distribution of cavity fields $Q(h)$ at a message passing fixed point:

$$Q(h)=\int\prod_k d\mathbf{J}_k\,P(\mathbf{J}_k)\prod_k dh_k\,Q(h_k)\,\delta\left(h-\sum_k u(\mathbf{J}_k,h_k)\right). \tag{46}$$

Here we have suppressed the arbitrary indices i and j. More generally, one can track the time-dependent evolution of the distribution of cavity fields, an algorithmic analysis technique known as density evolution [5].

In general, it can be difficult to solve the distributional equation for $Q(h)$. However, in the SK model, one could make an approximation that the distribution of cavity fields is a zero mean Gaussian with variance q. Then the distributional equation for $Q(h)$ reduces to a self-consistency condition for q by taking the expectation of h^2 on each side of (46). The left hand side is by definition q. To simplify the right hand side, since the couplings J_k have variance of $\frac{1}{N}$, we can use the small coupling approximation $u(J,h)\approx J\tanh(\beta h)$. Then averaging h^2 on both sides of (46) yields

$$q=\int\prod_k d\mathbf{J}_k\,P(\mathbf{J}_k)\prod_k dh_k\,Q(h_k)\left(\sum_k \mathbf{J}_k\tanh\beta h_k\right)^2 \tag{47}$$

$$=\int\prod_k d\mathbf{J}_k\,P(\mathbf{J}_k)\prod_k dh_k\,Q(h_k)\left(\sum_k \mathbf{J}_k^2\tanh^2\beta h_k\right) \tag{48}$$

$$=\int\prod_k dh_k\,Q(h_k)\left(\frac{1}{N}\sum_k\tanh^2\beta h_k\right) \tag{49}$$

$$=\int dh\,Q(h)\,\tanh^2\beta h. \tag{50}$$

Now, since we have assumed $Q(h)$ is zero mean Gaussian with variance q, this is equivalent to the replica symmetric saddle point equation (14).

In summary, we have employed a toy model of a neural network, the SK spin glass model, to introduce the various replica, cavity and message passing approaches to analyzing disordered statistical mechanical systems. In each case we have discussed in detail the simplest possible ansatz concerning the structure of free energy landscape, namely the replica symmetric ansatz, corresponding to a single valley with weak connected correlations between degrees of freedom. While this assumption is not true for the SK model, it nevertheless provides a good example system in which to gain familiarity with the various methods. And fortunately, for many of the applications discussed below, the assumption of a single free energy valley governing fluctuations will turn out to be correct. Finally, we note that just as the replica and cavity methods can be extended [7] to scenarios in which replica symmetry is broken, corresponding to many free energy valleys and long range correlations, so too can message passing approaches. Indeed viewing optimization and inference problems through the lens of statistical physics has lead to a new algorithm, known as survey propagation [77, 78], which can find good marginals, or minimize costs, in free energy landscapes characterized by many metastable minima that can confound more traditional, local algorithms.

3. Statistical Mechanics of Learning

In the above sections, we have reviewed powerful machinery designed to understand the statistical mechanics of fluctuating neural activity patterns in the presence of disordered synaptic connectivity matrices. A key conceptual advance made by Gardner [79, 80] was that this same machinery could be applied to the analysis of learning, by performing statistical mechanics directly on the space of synaptic connectivities, with the training examples presented to the system playing the role of quenched disorder. In this section we will explore this viewpoint and its applications to diverse phenomena in neural and unsupervised learning (see [12] for an extensive review of this topic).

3.1. Perceptron Learning

The perceptron is a simple neuronal model defined by a vector of N synaptic weights $\mathbf{w}$, which linearly sums a pattern of incoming activity $\boldsymbol{\xi}$, and fires depending on whether or not the summed input is above a threshold. Mathematically, in the case of zero threshold, it computes the function $\sigma = \mathrm{sgn}(\mathbf{w}\cdot\boldsymbol{\xi})$, where $\sigma = +1$ represents the firing state and $\sigma = -1$ represents the quiescent state. Geometrically, it separates its input space into two classes, each on opposite sides of the $N-1$ dimensional hyperplane orthogonal to the weight vector $\mathbf{w}$. Since the absolute scale of the weight vector $\mathbf{w}$ is not relevant to the problem, we will normalize the weights to satisfy $\mathbf{w}\cdot\mathbf{w} = N$, so that the set of perceptrons live on an $N-1$ dimensional sphere.

Suppose we wish to train a perceptron to memorize a desired set of P input-output associations, $\boldsymbol{\xi}^\mu \to \sigma^\mu$. Doing so requires a learning rule (an algorithm for modifying the synaptic weights $\mathbf{w}$ based on the inputs and outputs) that finds a set of synaptic

weights $\mathbf{w}$ that satisfies the P inequalities

$$\mathbf{w} \cdot \sigma^\mu \boldsymbol{\xi}^\mu \geq 0 \quad \forall \quad \mu = 1, \ldots, P. \tag{51}$$

We will see below, that as long as there exists a simultaneous solution $\mathbf{w}$ to the P inequalities, then remarkably, a learning rule, known as the perceptron learning rule [13], can find the solution. The main remaining question is then, under what conditions on the training data $\{\boldsymbol{\xi}^\mu, \sigma^\mu\}$ does a solution to the inequalities exist?

A statistical mechanics based approach to answering this question involves defining an energy function on the $N-1$ dimensional sphere of perceptrons as follows,

$$E(\mathbf{w}) = \sum_{\mu=1}^{P} V(\lambda^\mu), \tag{52}$$

where $\lambda^\mu = \frac{1}{\sqrt{N}} \mathbf{w} \cdot \sigma^\mu \boldsymbol{\xi}^\mu$ is the alignment of example μ with the weight vector $\mathbf{w}$. Successfully memorizing all the patterns requires all alignments to be positive, so V should be a potential that penalizes negative alignments and favors positive ones. Indeed a wide variety of learning algorithms for the perceptron architecture can be formulated as gradient descent on $E(\mathbf{w})$ for various choices of potential functions $V(\lambda)$ in (52) [12]. However, if we are interested in probing the space of solutions to the inequalities (51), it is useful to take $V(\lambda) = \theta(-\lambda)$, where $\theta(x)$ is the Heaviside function ($\theta(x) = 1, x \geq 0$, and 0 otherwise). With this choice, the energy function in (52) simply counts the number of misclassified examples, and so the Gibbs distribution

$$P(\mathbf{w}) = \frac{1}{Z} e^{-\beta E(\mathbf{w})} \tag{53}$$

in the zero temperature ($\beta \to \infty$) limit becomes a uniform distribution on the space of perceptrons satisfying (51) (see Fig. 4). Thus the volume of the space of solutions to

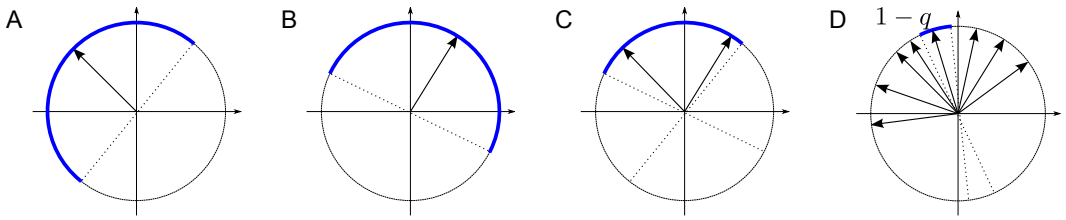

Figure 4. Perceptron Learning. (A) The total sphere of all perceptron weights (grey circle) and a single example (black arrow). The blue region is the set of perceptron weights that yield an output +1 on the example. (B) The same as (A), but for a different example. (C) The set of weights that yield +1 on both examples in (A) and (B). (D) As more examples are added, the space of correct weights shrinks, and its typical volume is given by $1-q$, where q is the replica order parameter introduced in section 3.3

(51), and in particular, whether or not it is nonzero, can be computed by analyzing the statistical mechanics of (53) in the zero temperature limit.

3.2. Unsupervised Learning

This same statistical mechanics formulation can be extended to more general unsupervised learning scenarios. In unsupervised learning, one often starts with a set of P data vectors $\boldsymbol{\xi}^\mu$, for $\mu = 1, \ldots, P$, where each vector is of dimension N. For example, each vector could be a pattern of expression of N genes across P experimental conditions, or a pattern of activity of N neurons in response to P stimuli. The overall goal of unsupervised learning is to find simple hidden structures or patterns in the data. The simplest approach is to find an interesting single dimension spanned by the vector $\mathbf{w}$, such that the projections $\lambda^\mu = \frac{1}{\sqrt{N}}\mathbf{w}\cdot\boldsymbol{\xi}^\mu$ of the data onto this single dimension yield a useful one dimensional coordinate system for the data. This interesting dimension can often be defined by minimizing the energy function (52), with the choice of $V(\lambda)$ determining the particular unsupervised learning algorithm. One choice, $V(\lambda) = -\lambda$ corresponds to Hebbian learning. Upon minimization of $E(\mathbf{w})$, this choice leads to $\mathbf{w} \propto \sum_\mu \boldsymbol{\xi}^\mu$; i.e. $\mathbf{w}$ points in the direction of the center of mass of the data. In situations in which the data has its center of mass at the origin, a useful choice is $V(\lambda) = -\lambda^2$. Under this choice, $\mathbf{w}$ points in the direction of the eigenvector of maximal eigenvalue of the data covariance matrix $\mathbf{C} = \sum_\mu \boldsymbol{\xi}^\mu \boldsymbol{\xi}^{\mu^T}$. This is the direction of maximal variance in the data, also known as the first principal component of the data; i.e. it is the direction which maximizes the variance of the distribution of λ^μ across data points μ.

Beyond finding an interesting dimension in the data, another unsupervised learning task to find clusters in the data. A popular algorithm for doing so is K-means clustering. This is an iterative algorithm in which one maintains a guess about K potential cluster centroids in the data, $\mathbf{w}_1, \ldots, \mathbf{w}_K$. At each iteration in the algorithm, each cluster i is defined to be the set of data points closer to centroid $\mathbf{w}_i$ than to any other centroid. Then all the cluster centroids $\mathbf{w}_i$ are optimized by minimizing the sum of the distances from $\mathbf{w}_i$ to those data points $\boldsymbol{\xi}^\mu$ assigned to cluster i. In the case where the distance measure is euclidean distance, this step just sets each centroid $\mathbf{w}_i$ to be the center of mass of the data points assigned to cluster i. The cluster assignments of the data are then recomputed with the new centroids, and the whole process repeats. The idea is that if there are K well separated clusters in the data, this iterative procedure should converge so that each $\mathbf{w}_i$ is the center of mass of cluster i.

For general K, this iterative procedure can be viewed as an alternating minimization of a joint energy function over cluster centroids and cluster membership assignments. For the special case of $K = 2$, and when both the data and cluster centroids are normalized to have norm N, this energy function can be written as

$$E(\mathbf{w}_1, \mathbf{w}_2) = \sum_{\mu=1}^{P} V(\lambda_1^\mu, \lambda_2^\mu), \tag{54}$$

where

$$\lambda_i^\mu = \frac{1}{\sqrt{N}}\mathbf{w}_i\cdot\boldsymbol{\xi}^\mu, \tag{55}$$

and,

$$V(\lambda_1, \lambda_2) = -\lambda_1\theta(\lambda_1 - \lambda_2) - \lambda_2\theta(\lambda_2 - \lambda_1) \tag{56}$$

$$= \frac{1}{2}(\lambda_1 + \lambda_2) - \frac{1}{2}|\lambda_1 - \lambda_2|. \tag{57}$$

Gradient descent on this energy function forces each centroid $\mathbf{w}_i$ to perform Hebbian learning only on the data points that are currently closest to it.

3.3. Replica Analysis of Learning

Both perceptron learning and unsupervised learning, when formulated as statistical mechanics problems as above, can be analyzed through the replica method. A natural question for perceptron learning is how many associations P can a perceptron with N synapses memorize? One benchmark is the case of random associations where $\boldsymbol{\xi}^\mu$ is a random vector drawn from a uniform distribution on a sphere of radius N and $\sigma^u = \pm 1$ each with probability half. Similarly, a natural question for unsupervised learning, is how do we assess the statistical significance of any structure or pattern we find in a high dimensional dataset consisting of P points in N dimensions? To address this question it is often useful to analyze what structure we may find in a null data distribution that itself has no structure, for example when the data points $\boldsymbol{\xi}^\mu$ are drawn from a uniform distribution on the $N-1$ sphere (or equivalently, in the large N limit, from a Gaussian distribution with identity covariance matrix).

In both cases, the analysis simplifies in the "thermodynamic" limit $P, N \to \infty$ with the ratio $\alpha = P/N$ held constant. Fortunately, this is the limit of relevance to neural models with many synaptic weights, and to high dimensional data. The starting point of the analysis involves understanding the low energy configurations of the Gibbs distribution in (53). In the thermodynamic limit, important observables, like the volume of low energy configurations or the distribution of data along the optimal direction(s) become self-averaging; they do not depend on the detailed realization of $\boldsymbol{\xi}^\mu$ or σ^μ. Therefore we can compute these observables by averaging $\log Z$ over these realizations. This can be done by first averaging the replicated partition function

$$\langle\langle\, Z^n \,\rangle\rangle = \left\langle\left\langle \int \prod_{a=1}^n d\mathbf{w}^a e^{-\sum_{a=1}^n \sum_{\mu=1}^P V(\lambda_a^\mu)} \right\rangle\right\rangle, \tag{58}$$

where $\lambda_a^\mu = \frac{1}{\sqrt{N}}\mathbf{w}_a \cdot \boldsymbol{\xi}^\mu$. (For the case of perceptron learning, we can make the redefinition $\sigma^\mu\ \boldsymbol{\xi}^\mu \to \boldsymbol{\xi}^\mu$, since both have the same distribution; in essence we absorb the sign of the desired output into the input yielding only positive examples.) Averaging over $\boldsymbol{\xi}^\mu$ then reduces to averaging over λ_a^μ. These variables are jointly Gaussian distributed with zero mean and covariance matrix $\langle\langle\, \lambda_a^\mu \lambda_b^\nu \,\rangle\rangle = Q_{ab}\delta_{\mu\nu}$ where $Q_{ab} = \frac{1}{N}\mathbf{w}^a \cdot \mathbf{w}^b$ is the replica overlap matrix. Thus after averaging over λ_a^μ, the integrand depends on the configuration of replicated weights only through their overlap. Therefore it is useful to separate the integral over $\mathbf{w}^a$ into an integral over all possible overlaps Q_{ab}, and an integral over all configurations with the same overlap. Following section 8, this yields

$$\langle\langle\, Z^n \,\rangle\rangle = \int \prod_{ab} dQ_{ab}\, e^{-N[E(Q) - S(Q))]}, \tag{59}$$

where

$$E(Q) = -\alpha \ln \int \prod_{a=1}^{n} \frac{d\lambda_a}{\sqrt{2\pi}} \frac{1}{\sqrt{\det Q}} e^{-\frac{1}{2}\lambda_a Q^{-1}_{ab} \lambda_b - \sum_a \beta V(\lambda_a)} \tag{60}$$

and

$$S(Q) = \frac{1}{2} \operatorname{Tr} \log Q \tag{61}$$

is the entropy of the volume of weight vectors with overlap matrix Q. For example, for perceptron learning when $V(\lambda) = \theta(-\lambda)$, in the zero temperature limit $\beta \to \infty$, $E(Q)$ is an energetic term that promotes the alignment of the replicated weights so that they all yield the correct answer on any given set of examples (i.e. $\lambda^a > 0$ for all a), while $S(Q)$ is an entropic term that promotes replicated weight configurations with small overlaps, since they have larger volume.

At large N the integral over Q_{ab} can be done via the saddle point method, and the competition between entropy and energy selects a saddle point overlap matrix. We make the ansatz that the saddle point has a replica symmetric form $Q_{ab} = (1-q)\delta_{ab} + q$. Given the connection (explained in section 8.2) between replica overlap matrix elements, and the distribution of overlaps of pairs of random weights $\mathbf{w}$ drawn independently from (53), this choice suggests the existence of a single free energy valley. This is reasonable to expect as most of the energy functions we will be analyzing for unsupervised learning are convex. Also, in the zero temperature limit, this ansatz suggests that the space of ground state energy configurations, if degenerate, should form a convex, connected set. This is indeed true for perceptron learning, since the space of ground states is the intersection of a set of P half-spheres (see Fig. 4). Thus unlike the SK model, we expect a replica symmetric assumption to be a good approximation.

Taking the $n \to 0$ limit yields a saddle point equation for q which, as explained in section 8.3.2 can be derived from extremizing a free energy

$$F(q) = \alpha \langle\langle \ln \zeta \rangle\rangle_z + \frac{1}{2}\Big[\frac{q}{1-q} + \ln(1-q)\Big], \tag{62}$$

where

$$\zeta = \int \frac{d\lambda}{\sqrt{2\pi(1-q)}} e^{-\frac{1}{2}\frac{(\lambda - \sqrt{q}z)^2}{1-q} - \beta V(\lambda)} \tag{63}$$

is the partition function of the distribution appearing inside the average over z in (155). Now in the case of perceptron learning, $1-q$ reflects the typical volume of the solution space to (51) (see Fig. 4D), in that q in the $\beta \to \infty$ limit is the typical overlap between two zero energy synaptic weight configurations (see section 8.2). q arises from a minimization of the sum of two terms in $F(q)$ in (62). The first term is an energetic term which is a decreasing function of q, reflecting a pressure for synaptic weights to agree on all examples (promoting larger q). The second term is an entropic term that is an increasing function of q, which thus promotes smaller values of q which reflect larger volumes in weight space. As α increases, placing greater weight on the first term in $F(q)$, q increases as energy becomes more important than entropy. As shown in [80],

$q \to 1$ as $\alpha \to \alpha_c = 2$. Thus a perceptron with N weights can store at most $2N$ random associations.

3.4. Perceptrons and Purkinje Cells in the Cerebellum

Interestingly, in [81] the authors developed a replica based analysis of perceptron learning and applied it to make predictions about the distribution of synaptic weights of Purkinje cells in the cerebellum. Indeed an analogy between the Purkinje cell and the perceptron was first posited over 40 years ago [82, 83]. The Purkinje cell has one of the largest and most intricate dendritic arbors of all neuronal cell types; this arbor is capable of receiving excitatory synaptic inputs from about 100,000 granule cells which, in areas of the cerebellum devoted to motor control, convey a sparse representation of ongoing internal motor states, sensory feedback, and contextual states. The Purkinje cell output in turn can exert an influence on outgoing motor control signals. In addition to the granule cell input, each Purkinje cell receives input from on average one cell in the inferior olive through a climbing fiber input, whose firing induces large complex spikes in the Purkinje cell as well as plasticity in the granule cell to Purkinje cell synapses. Since inferior olive firing is often correlated with errors in motor tasks, climbing fiber input is thought to convey an error signal that can guide plasticity. Thus at a qualitative level, the Purkinje cell can be thought of as performing supervised learning in order to map ongoing task related inputs to desired motor outputs, where the desired mapping is learned over time using error corrective signals transmitted through the climbing fibers.

Now the actual distribution of synaptic weights between granule cells and Purkinje cells has been measured [84], and a prominent feature of this distribution is that it has a delta function at 0, while the rest of the nonzero weights follow a truncated Gaussian distribution. In particular about 80 percent of the synaptic weights are exactly 0. If the Purkinje cell is implementing an important sensorimotor mapping, why then are a majority of the synapses silent? In general, the distribution of synaptic weights in a network should reflect the properties of the learning rule as well as the statistics of inputs and outputs. Thus one might be able to quantitatively derive the distribution of weights by positing a particular learning rule and input-output statistics and then derive the weight distribution. However, the authors of [81] took an even more elegant approach that did not depend on even positing any particular learning rule. They simply modeled the Purkinje cell architecture as a perceptron, assumed that it operated optimally at capacity, and derived the distribution of synaptic weights of perceptrons operating at capacity via a replica based Gardner type analysis. Remarkably, for a wide range of input-output statistics, whenever the perceptron implemented the maximal number of input-output associations at a given level of reliability (its capacity), its distribution of synaptic weights consisted of a delta-function at 0 plus a truncated Gaussian for the nonzero weights. Indeed, like the data, a majority of the synapses were silent. This prediction only relies on the perceptron operating at (or near) capacity, and does not depend on the learning rule; any learning rule that can achieve capacity would

necessarily yield such a weight distribution.

The key intuition for why a majority of the synapses are silent comes from the constraint that all the granule cell to Purkinje cell synapses are excitatory. Thus the either the Purkinje cell or the perceptron faces a difficult computational task: it must find a nonnegative synaptic weight vector that linearly combines nonnegative granule cell activity patterns and fires for some fraction of granule cell patterns while not firing for the rest. It turns out that false positive errors dominate the weight structure of the optimal perceptron operating at or near capacity: there are many granule cell activation patterns for which the perceptron must remain below threshold and not fire, and the only way to achieve this requirement with nonnegative weights is to set many synapses exactly to zero. Indeed by quantitatively matching the parameters of the replica based perceptron learning theory to physiological data, the capacity of the generic Purkinje cell was estimated to be about 40,000 input-output associations, corresponding to 5 kilobytes of information stored in the weights of a single cell [81].

3.5. Illusions of Structure in High Dimensional Noise

In contrast to perceptron learning, in the applications of the statistical mechanics formulation in (52) and (53) to unsupervised learning discussed here, $V(\lambda)$ has a unique minimum leading to a non-degenerate ground state. Thus in the zero temperature $\beta \to \infty$ limit we expect thermal fluctuations in the synaptic weights, reflected by $1-q$, to vanish. Indeed we can find self consistent solutions to the extremization of $F(q)$ in (62) by assuming $1-q = \frac{\Delta}{\beta}$ as $\beta \to \infty$ with Δ remaining $O(1)$. In this limit, (62) and (63) become

$$\frac{1}{\beta}F(\Delta) = -\alpha \Big\langle \Big\langle \min_\lambda \Big[\tfrac{(\lambda - z)^2}{2\Delta} + V(\lambda) \Big] \Big\rangle \Big\rangle_z + \frac{1}{2\Delta}. \tag{64}$$

Extremization of (64) over Δ determines the value of Δ as a function of α.

Furthermore, the interesting observable for unsupervised learning is the distribution of alignments across examples with the optimal weight vector,

$$P(\lambda) = \frac{1}{P}\sum_{\mu=1}^{P} \delta(\lambda - \lambda^\mu), \tag{65}$$

where $\lambda^\mu = \frac{1}{\sqrt{N}} \mathbf{w}\cdot\boldsymbol{\xi}^\mu$, and $\mathbf{w}$ minimizes $E(\mathbf{w})$ in (52). This is essentially the distribution of the data $\boldsymbol{\xi}^\mu$ along the dimension discovered by unsupervised learning. This distribution is derived via the replica method in section 8.4 at finite temperature, and is given by (155). Its zero temperature limit yields

$$P(\lambda) = \langle\langle\, \delta(\lambda - \lambda^*(z,\Delta)\,)\rangle\rangle_z, \tag{66}$$

where

$$\lambda^*(z,\Delta) = \mathrm{argmin}_\lambda \Big[\frac{(\lambda - z)^2}{2\Delta} + V(\lambda) \Big] \tag{67}$$

Equations (64), (66) and (67) have a simple interpretation within the zero temperature cavity method applied to unsupervised learning [85, 86]. Consider a cavity

system in which one of the examples, say example 1 in (52) is removed, and let $\mathbf{w}^{\setminus 1}$ be the"cavity" weight vector which optimizes $E(\mathbf{w})$ in the presence of all other examples $\boldsymbol{\xi}^\mu$ for $\mu = 2, \ldots, P$. Since $\mathbf{w}^{\setminus 1}$ does not know about about the random example $\boldsymbol{\xi}^1$, its overlap with this example, $z = \frac{1}{\sqrt{N}} \mathbf{w}^{\setminus 1} \cdot \boldsymbol{\xi}^1$ is a zero mean unit variance random gaussian variable. Now suppose example 1 is then included in the unsupervised learning problem. Then upon re-minimization of the total energy $E(\mathbf{w})$ in the presence of $\boldsymbol{\xi}^1$, the weight vector $\mathbf{w}^{\setminus 1}$ will change to a new weight vector, and consequently its alignment with $\boldsymbol{\xi}^1$ will also change from z to an optimal alignment λ^*. It can be shown [86] that for large N and P, this new optimal alignment arises through the minimization in (67). This minimization reflects a competition between two effects; the second term in (67) favors optimizing the alignment with respect to the new example, but the first term tries to prevent changes from the original alignment z. This term arises because $\mathbf{w}$ was already optimal with respect to all the other examples, so any changes in $\mathbf{w}$ incur an energy penalty with respect to the old examples. The parameter Δ plays the role of an inverse

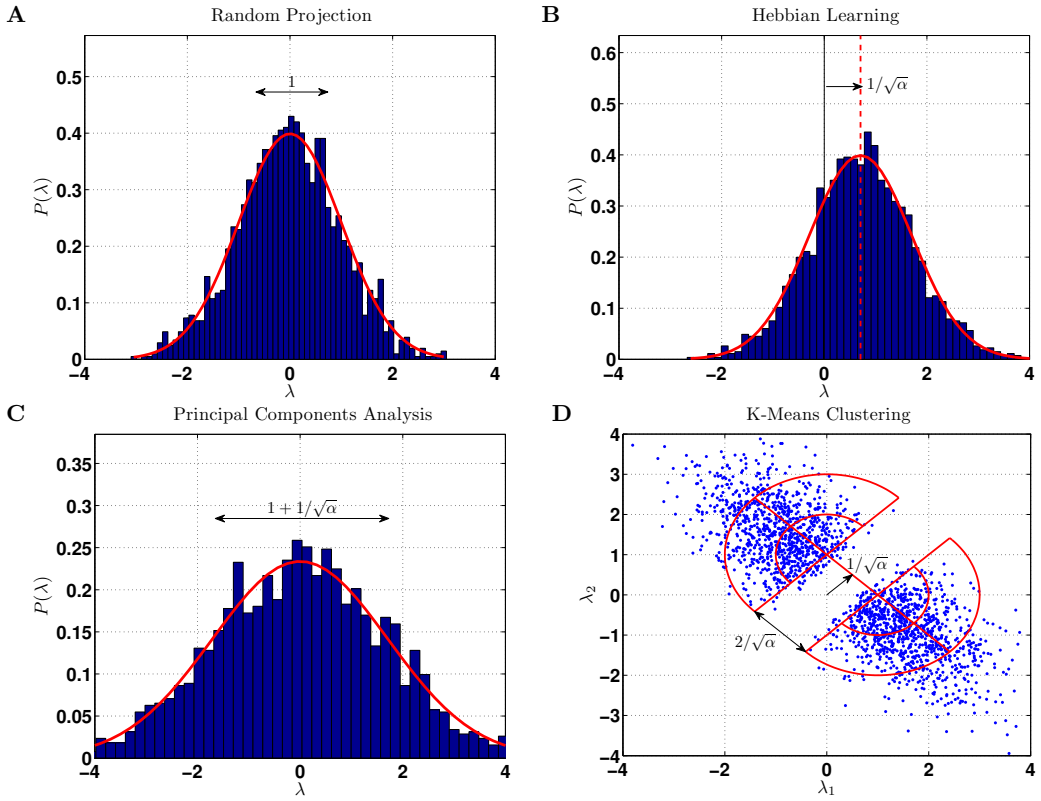

Figure 5. Illusions of structure. $P = 2000$ random points in $N = 1000$ dimensional space (so $\alpha = 2$) are drawn from a structureless zero mean, identity covariance Gaussian distribution. These points are projected onto different directions. (A) A histogram of the projection of these points onto a random direction; in the large N, P limit this histogram is Gaussian with 0 mean and unit variance. (B) A histogram of the same point cloud projected onto the Hebbian weight vector. (C) A projection onto the principal component vector. (D) The same point cloud projected onto two clusters directions found by K-means clustering with $K = 2$.

stiffness constant which determines the scale of a possible realignment of a weight vector with respect to a new example, and a self-consistency condition for Δ can be derived within the cavity approximation and is identical to the extremization of (64). This extremization makes Δ implicitly a function of α, and it is usually a decreasing function of α. Thus unsupervised learning becomes stiffer as α, or the number of examples, increases and the weight vector responds less to the presentation of any new example. Finally, example 1 is not special in any way. Thus repeating this analysis for each example, and averaging over the gaussian distribution of alignments z before learning any example, yields the distribution of alignments across examples after learning in (66).

We can now apply these results to an analysis of illusions of structure in high dimensional data. Consider an unstructured dataset, i.e. a random gaussian point cloud consisting of P points, $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^P$ in N dimensional space, where each point $\boldsymbol{\xi}^\mu$ is drawn i.i.d from a zero mean multivariate gaussian distribution whose covariance matrix is the identity matrix. Thus if we project this data onto a random direction $\mathbf{w}$, the distribution of this projection $\lambda^\mu = \frac{1}{\sqrt{N}} \mathbf{w} \cdot \boldsymbol{\xi}^\mu$ across examples μ will be a zero mean unit variance gaussian (see Fig. 5A). However, suppose we performed Hebbian learning to find the center of mass of the data. This corresponds to the choice $V(\lambda) = -\lambda$, and leads to $\lambda^*(z, \Delta) = z + \Delta$ from (67) with $\Delta = \frac{1}{\sqrt{\alpha}}$ from (64). Thus Hebbian learning yields an additive shift in the alignment to a new example whose magnitude decreases with the number of previous examples as $\frac{1}{\sqrt{\alpha}}$. After learning, we find that the distribution of alignments in (66) is a unit variance gaussian with a *nonzero* mean given by $\frac{1}{\sqrt{\alpha}}$ (see Fig. 5B). Thus a high dimensional random gaussian point cloud typically has a nonzero center of mass when projected onto the optimal Hebbian weight vector.

Similarly, we could perform principal components analysis to find the direction of maximal variance in the data. This corresponds to the choice $V(\lambda) = -\lambda^2$ and leads through (67) and (64) to $\lambda^*(z, \Delta(\alpha)) = (1 + \frac{1}{\sqrt{\alpha}})z$. Thus PCA scales up the alignment to a new example, and (66) leads to a gaussian distribution of alignments along the principal component with zero mean, but a standard deviation equal to $1 + \frac{1}{\sqrt{\alpha}}$ (see Fig 5C). This extra width is larger than any unity eigenvalue of the covariance matrix and leads to an illusion that the high dimensional gaussian point cloud has a large width along the principal component direction.

Finally, K-means clustering for $K = 2$, defined by the energy function in (56), involves a projection of the data onto two dimensions, determined by the two cluster centroids. However, the form of this energy function in (57) reveals a lack of interaction between the projected coordinates $\lambda_+ = \lambda_1 + \lambda_2$ and $\lambda_- = \lambda_1 + \lambda_2$. Along the direction λ_+, the algorithm behaves like Hebbian learning, so we should expected a gaussian distribution of the data along $\lambda_1 + \lambda_2$ with a mean of $\frac{1}{\sqrt{\alpha}}$. However, along $\lambda_1 - \lambda_2$ the algorithm is maximizing the absolute value of the projection, so that $V(\lambda) = -|\lambda|$. With this choice, (67) yields $\lambda^*(z, \Delta) = z + \mathrm{sgn}(z)\Delta$ with $\Delta = \frac{1}{\sqrt{\alpha}}$ determined by (64). Note that this implies that the distribution of alignments in (66) has a gap of zero density in the region $-\frac{1}{\sqrt{\alpha}} \leq \lambda \leq \frac{1}{\sqrt{\alpha}}$ and outside this region, the distribution is a split gaussian. The joint distribution of high dimensional data in K-means clustering

factorizes along $\lambda_1 + \lambda_2$ and $\lambda_1 - \lambda_2$ and does indeed have a gap of width $\frac{2}{\sqrt{\alpha}}$ along the $\lambda_1 - \lambda_2$ direction (see Fig. 5D). So quite remarkably, K-means clustering (with $K = 2$) of a random high dimensional gaussian point cloud reveals the illusion that there are 2 well separated clusters in the cloud. There is not a perfect match between the replica symmetric theory and numerical experiments because the discontinuity in the derivative of the energy in (57) actually leads to replica symmetry breaking [87]. However the corrections to the replica symmetric result are relatively small, and replica symmetry is a good approximation in this case; in contrast it is exact for Hebbian learning and PCA (see e.g. Fig 5B,C).

In summary, Fig. 5 reveals different types of illusions in high dimensional data whose effects diminish rather slowly as $O(\frac{1}{\sqrt{\alpha}})$ as the amount of data α increases. Indeed it should be noted that the very ability of the perceptron to store random patterns also depends on a certain illusion of structure: P random points in an N dimensional space will typically lie on one side of some hyperplane as long as $\alpha = P/N \leq 2$.

3.6. From Message Passing to Synaptic Learning

We have seen in section 3.1, that a perceptron with N synapses has the capacity to learn P random associations as long as $\alpha = P/N < \alpha_c = 2$. But what learning algorithm can allow a perceptron to learn these associations, up to the critical capacity? In the case of analog valued synaptic weights that we have been discussing, a simple algorithm, known as the perceptron learning algorithm [13, 14] can be proven to learn any set of associations that can be implemented (i.e. those associations $\{\boldsymbol{\xi}^\mu, \sigma^\mu\}$ for which a solution weight vector to (51) exists). The perceptron learning algorithm iteratively updates a set of randomly initialized weights as follows (for simplicity, we assume, without loss of generality, that $\sigma^\mu = 1$, for all patterns).

- When presented pattern μ, compute the current input $I = \mathbf{w} \cdot \boldsymbol{\xi}^\mu$.
- Rule 1: If $I \geq 0$, do nothing.
- Rule 2: If $I < 0$, update all weights: $\mathbf{w}_i \to \mathbf{w}_i + \boldsymbol{\xi}_i^\mu$.
- Iterate to the next pattern, until all patterns are learned correctly.

Such an algorithm will find realizable solutions to (51) in finite time for analog synaptic weights. However, what if synaptic weights cannot take arbitrary analog values? Indeed evidence suggests that biological synapses behave like noisy binary switches [88, 89], and thus can reliably code only two levels of synaptic weights, rather than a continuum. The general problem of learning in networks with binary weights (or weights with a finite discrete set of values) is much more difficult than the analog case; it is in fact an NP-complete problem [15, 16]. An exact enumeration and theoretical studies have revealed that when weights are binary (say $\mathbf{w}_i = \pm 1$), the perceptron capacity is is reduced to $\alpha_c = 0.83$, i.e. the space of binary weight vector solutions to (51) is nonempty only when $\alpha = P/N < \alpha_c = 0.83$ [90, 91]. Of course, below capacity, one can always find a solution through a brute-force search, but such a search

will require a time that is exponential in N. It is unlikely to expect to find a learning algorithm that provably finds solutions in a time that is polynomial in N, as this would imply $P = NP$. However, is it possible to find a learning algorithm that can typically (but not provably) find solutions in polynomial time at large $\alpha < 0.83$, and moreover, can this algorithm be biologically plausible?

The work of [17, 18] provided the first such algorithm. Their approach was to consider message passing on the joint probability distribution over all synaptic weights consistent with the desired associations (again we assume $\sigma^\mu = 1$):

$$P(\mathbf{w}) = \frac{1}{Z} \prod_{\mu=1}^{P} \theta(\mathbf{w}\cdot\boldsymbol{\xi}^\mu). \tag{68}$$

Here the factors are indexed by examples. The messages from examples to synapses and synapses to examples are all distributions on binary variables, and therefore can be parameterized by real numbers, as $u_{\mu\to i} = M_{\mu\to i}(+1) - M_{\mu\to i}(-1)$, and $M_{i\to\mu}(w_i) \propto e^{h_{i\to\mu} w_i}$. The message passing equations (32)-(33) then yield a dynamical system on the variables $u_{\mu\to i}$ and $h_{i\to\mu}$. This system drives the messages to approximate the marginal distribution of a synapse across all synaptic weight configurations which correctly learn all P associations. However, we would like to find a single synaptic weight configuration, not a distribution. To do this, in [18] the message passing equations are supplemented by a positive feedback term on the updates for $h_{i\to\mu}$. This positive feedback amplifies the messages over time and forces the weights to polarize to a single configuration, so that the approximation to the marginals through (34) becomes a delta function on ± 1 for all synapses i. Furthermore, in the large N limit, one can approximate the dynamical system on the $2PN$ variables $u_{\mu\to i}$ and $h_{i\to\mu}$ via an approximate message passing dynamical system on the time dependent variable $h_i^t = \sum_{t'<t} \sum_{\mu=1}^{P} u_{\mu\to i}^{t'}$ [17, 18]. Thus one obtains a learning rule in which each synapse maintains a single analog hidden variable h_i.

This rule was further simplified by allowing h_i to only take a finite number of discrete values, and the actual value of the synaptic weight was related to the hidden variable via $w_i = \mathrm{sgn}(h_i)$ [18]. After this simplification, the (amplified) message passing equations can be written in an online form in terms of the following algorithm (for convenience, h_i is allowed to take only odd integer values to avoid the ambiguous state $h_i = 0$):

- For pattern μ, compute the current input $I = \mathbf{w}\cdot\boldsymbol{\xi}^\mu$, where $w_i = \mathrm{sgn}(h_i)$.
- Rule 1: If $I \geq 1$, do nothing.
- Rule 2: If $I < 0$, update all internal states : $h_i \to h_i + 2\boldsymbol{\xi}_i^\mu$.
- Rule 3: If $I = 1$, then update each internal state $h_i \to h_i + 2\boldsymbol{\xi}_i^\mu$, but only if $h_i \boldsymbol{\xi}_i^\mu \geq 1$.
- Iterate to the next pattern, until all patterns are learned correctly.

The resulting rule is quite similar to the perceptron learning rule above, except for the modification of rule 1 and the addition of rule 3. Rule 3 concerns situations in which pattern μ is barely correct, i.e. a change in a single synaptic weight, or a single pattern

component would cause I to be below threshold (which is 0), resulting in an error. For barely learned patterns, rule 3 reinforces those internal variables that are already pointing in the right direction, i.e. contributing positively to the input current I on pattern μ, by making them larger in absolute value. Note that rule 3 cannot change any synaptic weight w_i; it is thus a *metaplastic* rule [92], or a rule that changes an internal state of the synapse without changing its synaptic efficacy.

Remarkably, the addition of rule 3, while seeming to be an innocuous modification of the perceptron learning rule, turns out to have a large impact on the learning capabilities of the discrete perceptron. For example, for a neuron with $N = O(10^5)$ synapses, when $\alpha \in [0.3 \dots 0.6]$, the message passing derived algorithm finds a solution with a few tens of presentations per pattern, whereas a similar clipped perceptron algorithm obtained by removing rule 3 is unable to find such a solution in $O(10^4)$ presentations per pattern [18]. Given the remarkable performance of message passing, it is intriguing to speculate whether some signature of message passing may exist within synapses. The key prediction is that in neurons that learn via error signals, metaplastic changes should occur whenever an error signal is absent, but the neuron is close to threshold.

4. Random Matrix Theory

The eigenvalue distributions of large random matrices play a central role in a variety of fields [19, 20]. For example, within the context of neuroscience, these distributions determine the stability of linear neural networks, the transition to chaos in nonlinear networks [21], and they are relevant to the statistical analysis of high dimensional data. Replica theory provides a powerful method to compute the eigenvalue spectrum of many different classical random matrix ensembles, including random symmetric [93] and asymmetric [94] matrices. More recently, it has been applied to matrices whose connectivity obeys Dale's law, which stipulates that all the outgoing synaptic weights of any neuron have the same sign [95]. Here we will introduce the replica formalism for symmetric matrices, focusing on the Wishart matrix ensemble [96, 97] because of its applications to high dimensional statistics discussed in section 5.3.

4.1. Replica Formalism for Random Matrices

Suppose $\mathbf{W}$ is an N by N random matrix whose elements are drawn from some probability distribution. For any specific realization of $\mathbf{W}$, its eigenvalue distribution is

$$\rho_{\mathbf{W}}(z) = \frac{1}{N} \sum_{i=1}^{N} \delta(z - z_i), \tag{69}$$

where z_i are the eigenvalues of $\mathbf{W}$. Now for large N, and for many distributions on the matrix elements of $\mathbf{W}$, this eigenvalue distribution is self-averaging; for any realization of $\mathbf{W}$, it converges as $N \to \infty$, to its average over $\mathbf{W}$, which we denote by $\langle\langle\, \rho_{\mathbf{W}}(\lambda)\, \rangle\rangle_{\mathbf{W}}$. We would like to theoretically compute this average, but it is difficult to average (69)

directly, since the eigenvalues z_i are complicated functions of the matrix elements of $\mathbf{W}$ (i.e. the roots of the characteristic polynomial $\det(z - \mathbf{W})$).

To perform this average, it is useful to physically think of the eigenvalues z_i as a collection of coulomb charges in the complex plane. In two dimensions, such charges repel each other with a force that decays inversely with distance. Then the resolvent,

$$R_{\mathbf{W}}(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i - z} = \int dz' \, \frac{\rho_{\mathbf{W}}(z')}{z' - z}, \tag{70}$$

can be thought of as (the negative of) the electric force field on a test charge placed at a point z in the complex plane, due to the presence of all the other charges $z_1, \ldots, z_N$. In mathematics, the transformation from $\rho_{\mathbf{W}}(z)$ to $R_{\mathbf{W}}(z)$ in (70) is known as the Stieltjes transform. For the case of symmetric $\mathbf{W}$, the charge density is confined to the real axis, and one can recover the charge density from its force field via the relation

$$\rho_{\mathbf{W}}(z) = \lim_{\epsilon \to 0^+} \frac{1}{\pi} \, Im \, R_{\mathbf{W}}(z + i\epsilon). \tag{71}$$

See section 8.5 for a derivation of this relation. Now the force on a test charge at a point z is the derivative of the electrostatic potential $\Phi_{\mathbf{W}}(z)$, and it turns out this potential, as opposed to either the charge density $\rho_{\mathbf{W}}(z)$ or the electric force $R_{\mathbf{W}}(z)$, will be easy to average over $\mathbf{W}$. We can derive a simple expression for the potential via the following sequence:

$$\begin{aligned} R_{\mathbf{W}}(z) &= \frac{1}{N} \operatorname{Tr} \frac{1}{\mathbf{W} - z} \\ &= -\frac{\partial}{\partial z} \frac{1}{N} \operatorname{Tr} \log (\mathbf{W} - z) \\ &= \frac{\partial}{\partial z} \frac{2}{N} \log \Big[\det (\mathbf{W} - z) \Big]^{-\frac{1}{2}} \\ &= \frac{\partial}{\partial z} \, \Phi_{\mathbf{W}}(z), \end{aligned} \tag{72}$$

where,

$$\Phi_{\mathbf{W}}(z) = \frac{2}{N} \log Z_{\mathbf{W}}(z) \tag{73}$$

and

$$Z_{\mathbf{W}}(z) = \int d\mathbf{u} \, e^{-\frac{i}{2}\mathbf{u}^T(\mathbf{W} - z)\mathbf{u}}. \tag{74}$$

Here we have used a Gaussian integral representation of $[\det(z - \mathbf{W})]^{-\frac{1}{2}}$ in (74) and neglected factors which do not survive differentiation by z in (72).

Now the electrostatic potential $\Phi_{\mathbf{W}}(z)$ is expressed in (73) as the free energy of a partition function $Z_{\mathbf{W}}(z)$ given by (74). We can use this representation to average the potential over $\mathbf{W}$, via the replica method to appropriately take care of the logarithm:

$$\begin{aligned} \langle\langle \, \Phi_{\mathbf{W}}(z) \, \rangle\rangle_{\mathbf{W}} &= \frac{2}{N} \, \langle\langle \log Z_{\mathbf{W}}(z) \, \rangle\rangle_{\mathbf{W}}, \\ &= \frac{2}{N} \lim_{n \to 0} \frac{\partial}{\partial n} \, \langle\langle \, Z^n_{\mathbf{W}}(z) \, \rangle\rangle_{\mathbf{W}}. \end{aligned} \tag{75}$$

This yields a general procedure for computing the average eigenvalue spectrum (i.e. charge density) of random Hermitian matrices. We first average a replicated version of the partition function in (74) (see (76) below). This allows us to recover the average electrostatic potential through (75), which then leads to the the average electric field through (72), which in turn leads to the average charge density through (71). We note that although we have focused on the case of hermitian matrices, this analogy between eigenvalues and coulomb charges extends to non-hermitian matrices in which the eigenvalue density is not confined to the real axis.

4.2. The Wishart Ensemble and the Marcenko-Pastur Distribution

As seen in the previous section, the first step in computing the eigenvalue spectrum of a Hermitian random matrix involves computing the average

$$\left\langle\left\langle\, Z^n_{\mathbf{W}}(z) \,\right\rangle\right\rangle_{\mathbf{W}} = \left\langle\left\langle\, \int \prod_{a=1}^n d\mathbf{u}^a \, e^{-\frac{i}{2}\sum_a \mathbf{u}^{a\,T}(\mathbf{W}-z)\mathbf{u}^a} \,\right\rangle\right\rangle_{\mathbf{W}}. \tag{76}$$

At this point we must choose a probability distribution over $\mathbf{W}$. When the matrix elements of $\mathbf{W}$ are chosen i.i.d. from a gaussian distribution, one obtains Wigner's semicircular law [98] for the eigenvalue distribution, which was derived via the replica method in [99].

Here we will focus on the Wishart random matrix ensemble in which

$$\mathbf{W} = \frac{1}{P}\mathbf{A}^T\mathbf{A}, \tag{77}$$

where $\mathbf{A}$ is a P by N matrix whose elements are chosen i.i.d. from a zero mean, unit variance Gaussian. This matrix has a simple interpretation in terms of high dimensional data analysis. We can think of each row of the matrix $\mathbf{A}$ as a data vector in an N dimensional feature space. Each data vector, or row of $\mathbf{A}$ is then a single draw from a multivariate Gaussian distribution in N dimensions, whose covariance matrix is the identity matrix. $\mathbf{W}$ is then the empirical covariance matrix of the P samples in the data set $\mathbf{A}$. In the low dimensional limit where the amount of data $P \to \infty$ and N remains $O(1)$, the empirical covariance matrix $\mathbf{W}$ will converge to the identity, and its spectrum will be a delta-function at 1. However, in the high dimensional limit in which $P, N \to \infty$ and $\alpha = P/N = O(1)$, then even though on average $\mathbf{W}$ will be the identity matrix, fluctuations in its elements are strong enough that its eigenvalue spectrum for typical realizations of the data $\mathbf{A}$ will not converge to that of the identity matrix. Even when $\alpha > 1$, the case of interest here, there will be some spread in the density around 1, and this spread can be thought of as another illusion of structure in high dimensional data, which we now compute via the replica method.

Inserting (77) into (76) we obtain

$$\left\langle\left\langle\, Z^n_{\mathbf{W}}(z) \,\right\rangle\right\rangle_{\mathbf{W}} = \int \prod_{a=1}^n d\mathbf{u}^a \left[\left\langle\left\langle\, e^{-\frac{i}{2}\sum_a \frac{1}{P}\mathbf{u}^{a\,T}(\mathbf{A}^T\mathbf{A})\mathbf{u}^a} \,\right\rangle\right\rangle_{\mathbf{A}} e^{\frac{iz}{2}\sum_a \mathbf{u}^{a\,T}\mathbf{u}^a}\right]. \tag{78}$$

Now the integrand depends on the quenched disorder $\mathbf{A}$ only through the variables $\lambda^a_\mu = \frac{1}{\sqrt{N}}\mathbf{a}_\mu \cdot \mathbf{u}^a$, where $\mathbf{a}_\mu$ is row μ of the matrix $\mathbf{A}$. These variables are jointly gaussian

distributed with zero mean and covariance $\left\langle\!\left\langle\, \lambda_\mu^a \lambda_\nu^b \,\right\rangle\!\right\rangle = Q_{ab}\delta_{\mu\nu}$ where $Q_{ab} = \frac{1}{N}\mathbf{u}^a \cdot \mathbf{u}^b$. Thus the average over $\mathbf{A}$ can be done by a gaussian integral over the variables λ_μ^a:

$$\left\langle\!\left\langle\, e^{-\frac{i}{2}\sum_a \frac{1}{P}\mathbf{u}^{a\,T}(\mathbf{A}^T\mathbf{A})\mathbf{u}^a} \,\right\rangle\!\right\rangle_{\mathbf{A}} = \left\langle\!\left\langle\, e^{-\frac{1}{2}\frac{i}{\alpha}\sum_a\sum_\mu (\lambda_a^\mu)^2} \,\right\rangle\!\right\rangle_{\{\lambda_\mu^a\}} \tag{79}$$

$$= \left[\left\langle\!\left\langle\, e^{-\frac{1}{2}\frac{i}{\alpha}\sum_a (\lambda_a)^2} \,\right\rangle\!\right\rangle_{\{\lambda^a\}}\right]^P \tag{80}$$

$$= \left[\det\left(I + \frac{i}{\alpha}Q\right)^{-\frac{1}{2}}\right]^P \tag{81}$$

$$= e^{-N\frac{\alpha}{2}\operatorname{Tr}\log(I+\frac{i}{\alpha}Q)} \tag{82}$$

Here in going from (79) to (80), we have exploited the fact that the variables λ_μ^a are uncorrelated for different μ, yielding a single average over variables λ_a with covariance $\langle\!\langle\, \lambda_a\lambda_b \,\rangle\!\rangle = Q_{ab}$, raised to the power P. In going from (80) to (82) we performed the gaussian integral over λ_a.

Thus consistent with the general framework in section 8.1, averaging over the disorder introduces interactions between the replicated degrees of freedom $\mathbf{u}^a$ which depend only on the overlap matrix Q_{ab}. Therefore we can compute the remaining integral over $\mathbf{u}^a$ in (78) by integrating over all overlaps Q_{ab}, and integrating over all configurations of $\mathbf{u}^a$ with a given overlap Q. This latter integral yields an entropic factor that depends on the overlap. In the end (78) becomes

$$\left\langle\!\left\langle\, Z_{\mathbf{W}}^n(z) \,\right\rangle\!\right\rangle_{\mathbf{W}} = \int \prod_{ab} dQ_{ab}\, e^{-N(E(Q)-S(Q))}, \tag{83}$$

where

$$E(Q) = \frac{\alpha}{2}\operatorname{Tr}\log(I + \frac{i}{\alpha}Q) - \frac{iz}{2}\operatorname{Tr} Q, \tag{84}$$

and

$$S(Q) = \frac{1}{2}\operatorname{Tr}\log Q, \tag{85}$$

is the usual entropic factor. The first term in (84) comes from (82) while the second term in (84) comes from the part outside the average over $\mathbf{A}$ in (78).

Now the final integral over Q_{ab} can be done via the saddle point method, and the integral can be approximated by the value of the integrand at the saddle point matrix Q which extremizes $F(Q) = E(Q) - S(Q)$. We can make a decoupled replica symmetric ansatz for this saddle point, $Q_{ab} = q\delta_{ab}$. With this choice, (75) leads to the electrostatic potential

$$\langle\!\langle\, \Phi_{\mathbf{W}}(z) \,\rangle\!\rangle_{\mathbf{W}} = -\alpha\log(1 + \frac{i}{\alpha}q) + izq + \log q \tag{86}$$

and (72) leads to the electric field

$$\langle\!\langle\, R_{\mathbf{W}}(z) \,\rangle\!\rangle_{\mathbf{W}} = iq. \tag{87}$$

Here q satisfies the saddle point equation obtained by extremizing $F(q)$, or equivalently the right hand side of (86),

$$-\frac{\alpha}{\alpha + iq} + z + \frac{1}{iq} = 0. \tag{88}$$

This is a z dependent quadratic equation for iq, and due to the relation between the electric field and charge density in (71), we are interested in those real values of z for which the solution iq has a nonzero imaginary part. It is in these regions of z that charges (eigenvalues) will accumulate, and their density will be proportional to this imaginary part. In the regime in which $\alpha > 1$ (so we have more data points than dimensions), a little algebra shows that iq has an imaginary part only when $z_- < z < z_+$ where $z_\pm = (1 \pm \frac{1}{\sqrt{\alpha}})^2$. In this region the charge density is

$$\langle\langle\, \rho_{\mathbf{W}}(z)\,\rangle\rangle_{\mathbf{W}} = \frac{\alpha\sqrt{(z-z_-)(z_+-z)}}{2\pi z}, \tag{89}$$

which is the Marcenko-Pastur (MP) distribution (see Fig. 6A below). Thus due to the high dimensionality of the data, the eigenvalues of the sample covariance matrix spread out around 1 over a range of $O(\pm\frac{1}{\alpha})$. This illusory spread becomes smaller as we obtain more data (increased α).

4.3. Coulomb Gas Formalism

In the previous section we found the marginal density of eigenvalues for a Wishart random matrix, but what about the entire joint distribution of all N eigenvalues? This distribution has a physically appealing interpretation that provides intuition for applications in high dimensional statistics discussed below. Consider the distribution of $\mathbf{W} = \mathbf{A}^T\mathbf{A}$, i.e. the matrix in (77) without the $\frac{1}{P}$ scaling. Because the P by N matrix $\mathbf{A}$ has i.i.d zero mean unit variance Gaussian elements, the distribution of $\mathbf{A}$ is given by

$$P(\mathbf{A}) \propto e^{-\frac{1}{2}\,\mathrm{Tr}\,\mathbf{A}^T\mathbf{A}}, \tag{90}$$

Now each matrix $\mathbf{A}$ has a unique singular value decomposition (SVD), $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Sigma}$ is a P by N matrix whose only N nonzero elements are on the diagonal: $\mathbf{\Sigma}_{ii} = \sigma_i$. The σ_i's are the singular values of $\mathbf{A}$, and the eigenvalues λ_i of $\mathbf{W}$ are simply the square of these singular values. Thus to obtain the joint distribution for λ_i, we first perform the change of variables $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ in the measure (90).

Fortunately, $P(\mathbf{A})$ is independent of $\mathbf{U}$ and $\mathbf{V}$, and depends only on $\mathbf{\Sigma}$. However, we need to transform the full measure $P(\mathbf{A})\,d\mathbf{A}$, and therefore we must include the Jacobian of the change of variables, given by (see e.g. [100])

$$d\mathbf{A} = \prod_{i<j}\left(\sigma_i^2 - \sigma_j^2\right)\prod_{i=1}^{N}\sigma_i^{P-N}(\mathbf{U}^T d\mathbf{U})(d\mathbf{\Sigma})(\mathbf{V}^T d\mathbf{V}). \tag{91}$$

Now the angular variables $\mathbf{U}$ and $\mathbf{V}$ decouple from the singular values σ_i, so we can integrate them out, yielding a constant. Furthermore, we can perform the change of variables $\lambda_i = \sigma_i^2$ to obtain

$$P(\lambda_1, \ldots, \lambda_N) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}\lambda_i}\prod_{i=1}^{N}\lambda_i^{\frac{1}{2}(P-N-1)}\prod_{j<k}|\lambda_j - \lambda_k|. \tag{92}$$

Here the first factor in the product arises from $P(\mathbf{A})$ in (90) while the second two factors arise from the Jacobian incurred by the change of measure in (91). This joint distribution can be written as a Gibbs distribution at unit temperature, $P(\{\lambda_i\}) \propto e^{-E(\{\lambda_i\})}$, where the energy is

$$E = \frac{1}{2}\sum_{i=1}^{N}(\lambda_i - (P-N-1)\ln\lambda_i) - \sum_{j\neq k}\ln|\lambda_j - \lambda_k|. \tag{93}$$

This energy function has a simple interpretation in which each eigenvalue is a Coulomb charge on confined to the real axis on the 2D complex plane. Each charge moves in a linear plus logarithmic potential which confines the charges, and there is a pairwise repulsion between all charges governed by a logarithmic potential (the Coulomb interaction in two dimensions). The Coulomb repulsion balances the confinement due to the external potential when the charges, or eigenvalues spread out over a typical range of $O(N)$. More precisely, this range is given by $(1 \pm \sqrt{\alpha})^2 N$ where $\alpha = P/N$ (note this is consistent with $z_\pm$ defined above (89) after rescaling by $\frac{1}{P}$), and within this range, the charge density in the $N \to \infty$ limit is given by the MP distribution.

4.4. Tracy-Widom Fluctuations

In the previous sections we have seen that full joint distribution of eigenvalues behaves like a Coulomb gas and its typical density at large N is given by the MP distribution. However, what do typical as well as large fluctuations of the maximal eigenvalue, or right most charge behave like? The distribution of the maximal eigenvalue forms a null distribution to test for the statistical significance of outcomes in PCA, and also plays role in random dimensionality reduction, so its fluctuations are of great interest. The mean of the maximal eigenvalue λ_{MAX} of course lies at the end of the MP charge density and is given, to leading order in N, by $\langle\lambda_{MAX}\rangle = (1+\sqrt{\alpha})^2 N$. Typical fluctuations about this mean have been found to scale as $O(N^{\frac{1}{3}})$ [24]. More precisely, for large N they have the limiting form

$$\lambda_{MAX} = (1+\sqrt{\alpha})^2 N + \alpha^{-\frac{1}{6}}(1+\sqrt{\alpha})^{\frac{4}{3}} N^{\frac{1}{3}}\chi, \tag{94}$$

where χ is the Tracy-Widom distribution that has a range of $O(1)$ [24].

The computation of these typical fluctuations is involved, but often we are interested in the probability of large deviations in which $|\lambda_{MAX} - \langle\lambda_{MAX}\rangle| = O(N)$. These large deviations were computed in [26, 27] in a very simple way using the Coulomb gas picture. Suppose for example the largest eigenvalue occurs at distance that is $O(N)$ to the right of the typical edge of the MP density, $(1+\sqrt{\alpha})^2 N$. The most likely way this could happen (i.e. the saddle point configuration of charges in (92) [27]), is that a *single* eigenvalue pops out of the MP density, while the remaining eigenvalues are unperturbed, and preserve the shape of the MP density. The energy paid by a single eigenvalue popping out of the MP density is dominated by the linear confining term in (93), and is therefore proportional to the distance it pops out. Since the probability of

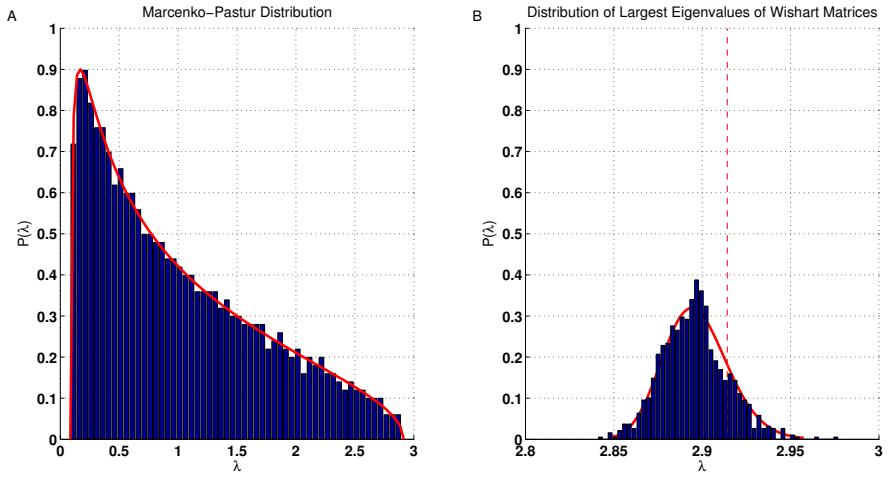

Figure 6. Spectral distributions of empirical noise covariance matrices. (A) $P = 2000$ random points in $N = 1000$ dimensional space ($\alpha = 2$) are drawn from a 0 mean identity covariance normal distribution. The blue histogram is the distribution of eigenvalues of the empirical covariance matrix $\mathbf{W} = \frac{1}{P}\mathbf{A}^T\mathbf{A}$, where $\mathbf{A}$ is a P by N data matrix whose rows correspond to the points (see Eq. (77)). The red curve is the Marcenko-Pastur distribution (see Eq. (89)) for $\alpha = 2$. (B) A histogram, in blue, of the maximal eigenvalues of 1000 random covariances matrices $\mathbf{W}$, each constructed exactly as in (A). The red curve is the Tracy-Widom distribution in (94) for $\alpha = 2$, rescaled by $\frac{1}{P}$. The dashed red line marks the edge of the Marcenko-Pastur distribution in (A). The discrepancy between this edge and the mean of the maximal eigenvalue distribution is a finite size effect; this discrepancy, like the fluctuations in the maximal eigenvalue, vanishes as $O(N^{-2/3})$.

a fluctuation is exponentially suppressed by its energy cost (here entropy plays no role because the MP density is unperturbed), we obtain

$$\text{Prob}\left(\lambda_{MAX} = \langle\lambda_{MAX}\rangle + cN\right) \propto e^{-N\Phi_+(c)} \quad \text{for} \quad cN \gg \langle\lambda_{MAX}\rangle. \tag{95}$$

Thus large deviations of $O(N)$ are exponentially suppressed by N, and the $O(1)$ large deviation constant $\Phi_+(c)$ can be computed explicitly by quantitatively working out this Coulomb gas argument [26, 27].

On the other hand, suppose that the maximal eigenvalue λ_{MAX} occurs at a distance cN to the left of right edge of the MP density. In order for this fluctuation to occur, the entire MP density must become compressed, incurring a much larger energy cost compared to a positive or rightward fluctuation of λ_{MAX}. Indeed because of the Coulomb repulsion between all pairs of charges in (93), the energy cost of compression is $O(N^2)$ leading to the stronger suppression,

$$\text{Prob}\left(\lambda_{MAX} = \langle\lambda_{MAX}\rangle - cN\right) \propto e^{-N^2\Phi_-(c)} \quad \text{for} \quad cN \gg \langle\lambda_{MAX}\rangle, \tag{96}$$

where $\Phi_-(c)$ is an $O(1)$ large deviation function computed in [26, 27]. Thus the physics of Coulomb gases gives a nice explanation for the asymmetry in the large deviations of the Tracy-Widom distribution.

For the reader's convenience, we summarize the implications of the Coulomb gas formalism for high dimensional statistics by reintroducing the $\frac{1}{P}$ scaling in the definition (77) of the empirical covariance matrix. The above results then tell us that the maximal eigenvalue of the empirical covariance matrix of P random Gaussian points in N dimensional space, in the limit $N, P \to \infty$ with $\alpha = P/N$ remaining $O(1)$ (but strictly greater than 1), has a mean $\langle \lambda_{MAX} \rangle = (1 + \frac{1}{\sqrt{\alpha}})^2$ with typical fluctuations about this mean that are $O(N^{-\frac{2}{3}})$ (see Fig. 6B). Moreover, the probability of large $O(1)$ positive deviations of λ_{MAX} are $O(e^{-N})$, while the probability of large $O(1)$ negative deviations of λ_{MAX} are $O(e^{-N^2})$.

5. Random Dimensionality Reduction

We have seen in section 3.5 that we may need to be careful when we perform dimensionality reduction of high dimensional data by looking for optimal directions along which to project the data, as this process can potentially lead to illusions of structure. An alternate approach might be to skip the optimization step responsible for the illusion, and simply project our data onto randomly chosen directions. However, it is not at all obvious that such random dimensionality reduction would preserve the true or interesting structure that is present in the data. Remarkably, a collection of theoretical results reveal that random projections (RP's) preserve much more structure than one might expect.

5.1. Point Clouds

A very generic situation is that data often lies along a low dimensional manifold embedded in a high dimensional space. An extremely simple manifold is a point cloud consisting of a finite set of points, as in Fig. 7A. Suppose this cloud consists of P points $\mathbf{s}^\alpha$, for $\alpha = 1, \ldots, P$, embedded in an N dimensional space, and we project them down to the points $\mathbf{x}^\alpha = \mathbf{A}\mathbf{s}^\alpha$ in a low M dimensional space through an appropriately normalized random $M \times N$ random projection matrix $\mathbf{A}$. The squared euclidean distances between pairs of points in the high dimensional space are given by $||\mathbf{s}^\alpha - \mathbf{s}^\beta||^2$ and in the low dimensional space by $\|\mathbf{x}^\alpha - \mathbf{x}^\beta\|^2$. The fractional distortion in the squared distance incurred by the projection is given by

$$D_{\alpha\beta} = \frac{\|\mathbf{x}^\alpha - \mathbf{x}^\beta\|^2 - \|\mathbf{s}^\alpha - \mathbf{s}^\beta\|^2}{\|\mathbf{s}^\alpha - \mathbf{s}^\beta\|^2}. \tag{97}$$

How small can we make M before the point cloud becomes distorted in the low dimensional space, so that the low and high dimensional distances are no longer similar?

The celebrated Johnson-Lindenstrauss (JL) lemma [101] (see [102, 103] for more recent and simpler proofs) and provides a striking answer. It states that for any distortion level $0 < \delta < 1$, as long as $M > O(\frac{\ln P}{\delta^2})$, with high probability, one can find a projection such that

$$-\delta \leq D_{\alpha\beta} \leq \delta \tag{98}$$

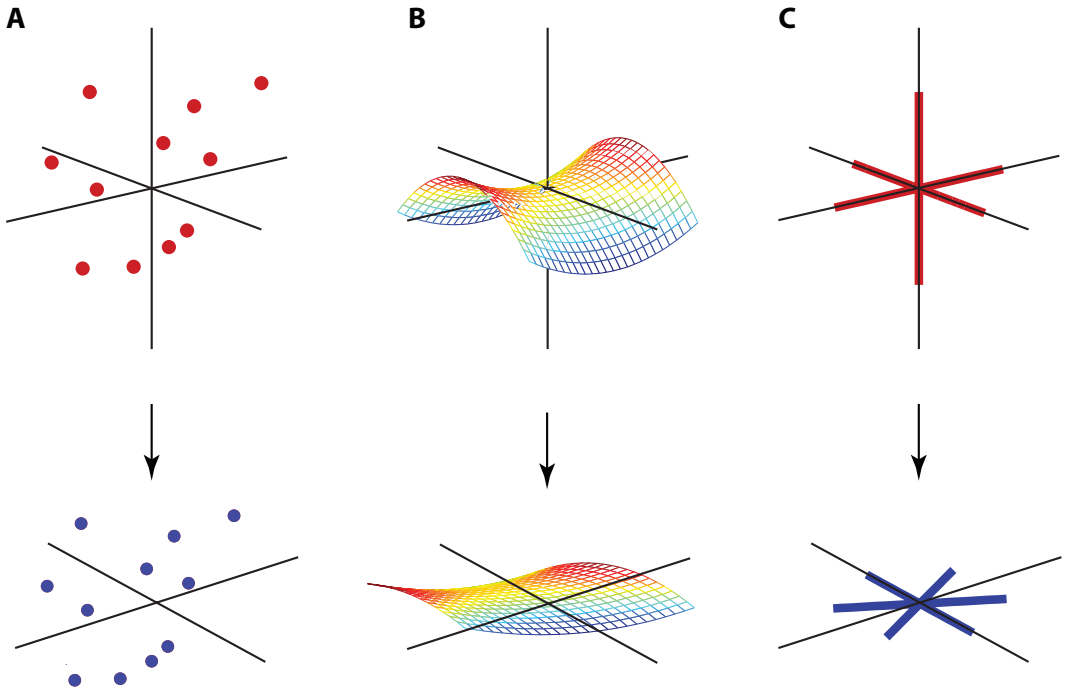

Figure 7. Random Projections. (A,B) Projection of a point cloud, and a nonlinear manifold respectively. (C) A manifold of K-sparse signals (red) in N dimensional space is randomly projected down to an M dimensional space (here $K = 1$, $N = 3$, $M = 2$).

for all pairs of points α and β. Thus the distortion between any pair of points rarely exceeds δ. This is striking because the number of projected dimensions M need only be *logarithmic* in the number of points P *independent* of the embedding dimension of the source data, N. Of course, with so few projections, one cannot reconstruct the original data from its projections. Nevertheless, surprisingly, with so few random projections the geometry of the entire point cloud is preserved. We will discuss a statistical mechanics based approach for understanding the JL lemma in section 5.3.

5.2. Manifold Reduction

Consider data distributed along a nonlinear K dimensional manifold embedded in N dimensional space, as in Fig 7B. An example might be a set of images of a single object observed under different lighting conditions, perspectives, rotations and scales. Another example would be the set of neural firing rate vectors in a brain region in response to a continuous family of stimuli. [104, 105, 106] show that $M > O(\frac{K}{\delta^2} \log NC)$ random projections preserve the geometry of the manifold up to distortion δ. Here C is a number related to the curvature of the manifold, so that highly curved manifolds require more projections. Overall, the interesting result is that the required number of projections depends linearly on the intrinsic dimensionality of the manifold, and only logarithmically on its ambient embedding dimension.

The simplest finite dimensional manifold is a K dimensional linear subspace in an

N dimensional space. It can be shown [107] that $M > O(\frac{K}{\delta^2})$ RP's are sufficient to preserve all pairwise distances between data points within a distortion level δ. We will give an alternate proof of this result in section 5.3 below using the results of sections 4.2 and 4.3. Of course, for such a simple manifold, there exists a nonrandom, optimal geometry preserving projection, namely the PCA basis consisting of $M = K$ orthogonal vectors spanning the manifold. Thus we pay a price in the number of projections for choosing random projections rather than the optimal ones. Of course, for data that is not distributed along a hyperplane, a PCA based projection will no longer be optimal, and will not generically preserve geometry.

Sparsity is another example of an interesting low dimensional structure. Consider for example, the (nonsmooth) manifold of N dimensional signals with only K nonzero components. This is a manifold of $\binom{N}{K}$ coordinate hyperplanes in N dimensional space, as in Fig. 7C. The geometry of this manifold can also be preserved by random projections. In particular [107] shows that random projections down to an $M > O(\frac{K}{\delta^2} \log \frac{N}{K})$ dimensional space preserve the distance between any pair of K-sparse signals, with distortion less than δ.

Beyond the issue of preserving geometry under an RP, one might be interested in situations in which one can invert the random projection; i.e. given the projection of a data vector, or signal in the low M dimensional space, how might one recover the original signal in the high N dimensional space? For the case of point clouds (Fig. 7A) and general nonlinear manifolds (Fig. 7B), there are no general computationally tractable algorithms capable of achieving this high dimensional signal recovery. However, for the case of K-sparse signals (Fig. 7C), there exists a simple, computationally tractable algorithm, known as L_1 minimization, reviewed below, that can provably, under various assumptions on the RP matrix $\mathbf{A}$, recover the high dimensional signal from its projection. It turns out that geometry preservation is a sufficient condition for signal recovery; in particular, [108] shows that any projection which preserves the geometry of all K-sparse vectors, allows one to reconstruct these vectors from the low dimensional projection, efficiently and robustly using L_1 minimization. This is one of the observations underlying the field of compressed sensing, reviewed below.

However, even in situations where one cannot accurately reconstruct the high dimensional signal from its projection, RP's can still be very useful by allowing for compressed computation directly in the low dimensional projection space, without the need for signal recovery. This can be done because many interesting machine learning and signal processing algorithms depend only on pairwise distances between data points. For example, regression [109], signal detection [110], classification [111, 112, 113, 114], manifold learning [115], and nearest neighbor finding [102] can all be accomplished in a low dimensional space given a relatively small number of RP's. Moreover, task performance is often comparable to what can be obtained by performing the task directly in the original high dimensional space. The reason for this remarkable performance is that these computations rely only on the distances between data points, which are preserved by RP's.

5.3. Correlated Extreme Value Theory and Dimensionality Reduction

The proofs [101, 102, 103, 107, 104] behind the remarkable sufficient conditions for low distortion of submanifolds under RP's rely on sequences of potentially loose inequalities. Thus they leave open the question of whether these sufficient conditions are actually necessary, and how the typical, as opposed to worst case, distortion behaves. Is it possible to use a direct approach, more in the spirit of statistical mechanics, to simply compute the probability distribution of the typical distortion of random manifolds under random projections? Here the random choice of manifold plays the role of quenched disorder, the random choice of projection plays the role of thermal degrees of freedom, and the observable of interest is the distribution, across choices of RP's, of the maximal distortion over all pairs of points in the fixed manifold. The hope is that this distribution is self-averaging, in that it does not depend on the choice of a particular manifold from a suitable ensemble of manifolds. In general, this approach is challenging, but here we show that it can be carried out for two very simple classes of manifolds: random point clouds and hyperplanes. Our main goal in this section is simply to obtain intuition for the scaling behavior of some of the inequalities discussed in the previous section.

Consider a fixed realization of a Gaussian point cloud consisting of P points $\mathbf{s}^\alpha$, $\alpha = 1, \ldots, P$ in N dimensional space. Let $\mathbf{A}$ be an M by N random projection operator whose matrix elements are i.i.d Gaussian with zero mean and variance $\frac{1}{M}$, and let $\mathbf{x}^\alpha = \mathbf{A}\mathbf{s}^\alpha$ be the low dimensional image of the cloud. With this choice of scaling for the variance of the projection operator, it is straightforward to show that any one distortion $D_{\alpha\beta}$ in (97) is in the large N limit, for a fixed (i.e. quenched) point cloud, a Gaussian random variable with zero mean and variance $O(\frac{1}{M})$, due to the random choice of $\mathbf{A}$. Now there are $\binom{P}{2} = O(P^2)$ pairs of points, or possible distortions, and we are interested in the maximum distortion, whose behavior could in principle depend on the correlations between pairs of distortions. For random Gaussian point clouds, the correlation coefficient between two pairs of distortions are weak, in fact $O(\frac{1}{N})$, and can be neglected. In this manner, the ambient dimensionality of the point cloud disappears from the problem. Thus the maximum distortion over all pairs of points can be well approximated by the maximum of $O(P^2)$ independent Gaussian variables each with variance $O(\frac{1}{M})$. In general the maximum of R independent Gaussian variables with variance σ^2 approaches a Gumbel distribution in the large R limit, and takes typical values of $O(\sigma\sqrt{\ln R})$. Indeed the Gumbel distribution is a universal distribution governing extreme values of any random variables whose tails vanish faster than exponentially [116]; this strong suppression of extreme values in any single variable directly leads to an extremely slow $\sqrt{\ln R}$ growth of the maximum over R realizations of such variables. Applying this general result with $\sigma^2 = O(\frac{1}{M})$ and $R = O(P^2)$ yields the conclusion that the maximal distortion over all pairs of points obeys a Gumbel distribution, and its typical values scale with P and M as $O(\sqrt{\frac{\ln P}{M}})$. Thus the origin of the extremely slow $\sqrt{\ln P}$ growth of the maximal distortion with the number of points P is due to the strong, Gaussian suppression of any individual distortion. This effect

is directly responsible for the remarkable JL lemma. For example, if we desire our maximal distortion to be less than δ, we must have

$$O\left(\sqrt{\frac{\ln P}{M}}\right) < \delta, \tag{99}$$

or equivalently $M > O\left(\frac{\ln P}{\delta^2}\right)$, which, up to constants, is the JL result. Thus extreme value theory for uncorrelated Gaussian variables provides a natural intuition for why the number of random projections M need only be logarithmic in the number of points P, and independent of the ambient dimension N, in order to achieve an $O(1)$ distortion δ.

For random Gaussian point clouds, we were able to neglect correlations in the distortion between different pairs of points. For more general manifold ensembles, we will no longer be able to do this. However, for the ensemble of random hyperplanes, an exact analysis is still possible despite the presence of these correlations. Let $\mathbf{U}$ be an N by K random matrix whose K orthonormal columns form a basis for a random K dimensional subspace of N dimensional space (drawn uniformly from the space of such subspaces). What is the distribution of the maximal distortion in (97) where $\mathbf{s}^\alpha$ and $\mathbf{s}^\beta$ range over all pairs of points in this subspace? First, by exploiting rotational invariance of the ensemble of $\mathbf{A}$ and $\mathbf{U}$, we can always perform a change of basis in N dimensional space so that the columns of $\mathbf{U}$ are mapped to the first K coordinate axes. Thus points in the hyperplane can be parameterized by N dimensional vectors whose only nonzero components are the first K coordinates, and the statistics of their projection to M dimensional space can be determined simply by the M by K submatrix of $\mathbf{A}$ consisting of its first K columns. In this manner, the dimensionality N of the ambient subspace again disappears from the problem. Second, by exploiting the linearity of the projection operator, to compute the maximal distortion over all pairs of points in the plane, it suffices to compute the maximal distortion over all points on the unit sphere in K dimensional space. Thus if we let $\bar{\mathbf{A}}$ denote the M by K submatrix of $\mathbf{A}$, and let $\mathbf{s}$ denote a K-dimensional coordinate vector for the hyperplane, then we have

$$\max_{\alpha\beta} D_{\alpha\beta} = \max_{\{\mathbf{s}\,,\,||\mathbf{s}||^2=1\}} \sqrt{\mathbf{s}^T \bar{\mathbf{A}}^T \bar{\mathbf{A}} \mathbf{s}} - 1. \tag{100}$$

Here, to obtain a slightly cleaner final result, we are now measuring the distortion $D_{\alpha\beta}$ in terms of fractional change in Euclidean distance as opposed to the squared Euclidean distance used in (97), hence the square root in (100). The constrained maximum over $\mathbf{s}$ of $\mathbf{s}^T \bar{\mathbf{A}}^T \bar{\mathbf{A}} \mathbf{s}$ in (100) is simply the maximum eigenvalue of the matrix $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$, and its distribution over the random choice of $\mathbf{A}$ has been characterized in sections 4.2 and 4.3. In fact the results in these sections carry over with the replacements $P \to M$ and $N \to K$. The maximal eigenvalue is with high probability equal to $(1 + \frac{1}{\sqrt{\alpha}})^2$, with $\alpha = \frac{M}{K}$. Its typical fluctuations are $O(M^{-\frac{2}{3}})$, while its large positive deviations of $O(1)$ are exponentially suppressed in M, i.e. are $O(e^{-M})$. So the maximal distortion in (100) is close to $\frac{1}{\sqrt{\alpha}}$. As long as $M > K$, a similar argument holds for the minimum distortion, which will be close to $-\frac{1}{\sqrt{\alpha}}$. Indeed, if $M < K$, then $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$ will have zero eigenvalues,

which correspond geometrically to vectors in the random hyperplane $\mathbf{U}$ that lie in the kernel of the random projection $\mathbf{A}$. So as long as $\alpha > 1$, the distribution of distance distortions $D_{\alpha\beta}$ will with high probability lie in the range $-\frac{1}{\sqrt{\alpha}}$ to $+\frac{1}{\sqrt{\alpha}}$. This means of course, that if one wants all distortions $D_{\alpha\beta}$ to obey $-\delta < D_{\alpha\beta} < +\delta$, then one can achieve this with high probability as long as $\delta < \frac{1}{\sqrt{\alpha}}$, or equivalently, the number of random projections obeys $M > \frac{K}{\delta^2}$, which proves the claim about RP's of hyperplanes made in section 5.2. Overall, this argument shows how the extremal fluctuations of correlated random variables (i.e. the charges of a Coulomb gas described in section 4.3) can be used to understand geometric distortions induced by RP's of simple manifolds, namely hyperplanes.

6. Compressed Sensing

We have seen in section 5 that random projections can preserve the geometric structure of low dimensional signal manifolds. Furthermore, in the case in which the manifold is the space of K-sparse signals (Fig. 7C), as discussed above, one can actually recover the high dimensional signal from its projection using a computationally tractable algorithm, known as L_1 minimization. Here we review this algorithm and its analysis based on statistical mechanics and message passing. As discussed in the introduction, many applications of the ideas in this section and the previous one are described in [30].

6.1. L_1 Minimization

Suppose $\mathbf{s}^0$ is an unknown sparse N dimensional vector which has only a fraction $f = K/N$ of its elements nonzero. Thus $\mathbf{s}^0$ is a point in the top manifold of Fig. 7C. Suppose we are given a vector $\mathbf{x}$ of $M < N$ measurements, which is linearly related to $\mathbf{s}^0$ by an M by N measurement matrix $\mathbf{A}$, i.e. $\mathbf{x} = \mathbf{A}\mathbf{s}^0$. $\mathbf{x}$ is then the projection of $\mathbf{s}^0$ in the bottom manifold of Fig. 7C. Each measurement x_μ, for $\mu = 1, \ldots, M$ is a linear function $\mathbf{a}_\mu \cdot \mathbf{s}^0$ of the unknown signal $\mathbf{s}^0$, where $\mathbf{a}_\mu$ is the μ'th row of $\mathbf{A}$. In the context of signal processing, $\mathbf{s}^0$ could be a temporal signal, and the $\mathbf{a}_\mu$ could be a set of N temporal filters. In the context of network reconstruction, $\mathbf{s}^0$ could be a vector of presynaptic weights governing the linear response of a single postsynaptic neuron x_u to a pattern of presynaptic stimulation $\mathbf{a}_\mu$ on a trial μ [35].

Now how might one recover $\mathbf{s}^0$ from $\mathbf{x}$? In general, this is an underdetermined problem; there is an $N - M$ dimensional space of candidate signals $\mathbf{s}$ that satisfy the measurement constraint $\mathbf{x} = \mathbf{A}\mathbf{s}$. The true signal $\mathbf{s}^0$ is but one point in this large space. However, we can try to exploit our prior knowledge that $\mathbf{s}^0$ is sparse by searching for sparse solutions to the measurement constraints. For example, one could solve the optimization problem

$$\hat{\mathbf{s}} = \text{argmin}_{\mathbf{s}} \sum_{i=1}^{N} V(s_i) \quad \text{subject to } \mathbf{x} = \mathbf{A}\mathbf{s}, \tag{101}$$

to obtain an estimate $\hat{\mathbf{s}}$ of $\mathbf{s}^0$. Here $V(x)$ is any sparsity promoting function. A natural choice is $V(x) = 1$ if $x = 0$ and $V(x) = 0$ otherwise, so that (101) yields the the signal consistent with the measurements $\mathbf{x}$ with the minimum number of nonzero elements. However, this is in general a hard combinatorial optimization problem. One could relax this optimization problem by choosing $V(s) = |s|^p$, so that (101) finds a solution to the measurement constraints with minimal L_p norm. However, this optimization problem is nonconvex for $p < 1$. Thus a natural choice is $p = 1$, the lowest value of p for which the recovery algorithm (101) becomes a convex optimization problem, known as L_1 minimization.

6.2. Replica Analysis

Much of the seminal theoretical work in CS [117, 118, 108] has focused on sufficient conditions on $\mathbf{A}$ to guarantee perfect signal recovery, so that $\hat{\mathbf{s}} = \mathbf{s}^0$ in (101), in the case of L_1 minimization. But often, large *random* measurement matrices $\mathbf{A}$ which violate these sufficient conditions nevertheless typically yield good signal reconstruction performance. Thus these sufficient conditions are not necessary. Here we review a statistical mechanics approach to CS based on the replica method [119, 120, 121], which allows one to directly compute the typical performance of L_1 minimization. Some of these results have also been derived using message passing [55] and polyhedral geometry [122].

To understand the properties of the solution $\hat{\mathbf{s}}$ to the optimization problem in (101), we define an energy function on the residual $\mathbf{u} = \mathbf{s} - \mathbf{s}^0$ given by

$$E(\mathbf{u}) = \frac{\lambda}{2N}\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u} + \sum_{i=1}^{N}|u_i + s_i^o|, \tag{102}$$

and analyze the statistical mechanics of the Gibbs distibution

$$P_G(\mathbf{u}) = \frac{1}{Z}e^{-\beta E(\mathbf{u})}. \tag{103}$$

By taking the limit $\lambda \to \infty$ we enforce the constraint $\mathbf{x} = \mathbf{A}\mathbf{s}$. Then taking the low temperature $\beta \to \infty$ limit condenses the Gibbs distribution onto the vicinity of the global minimum of (101). Then we can compute the average error

$$Q_0 = \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{u}_i \rangle_{P_G}^2, \tag{104}$$

and, if needed, the thermal fluctuations

$$\Delta Q = \frac{1}{N}\sum_{i=1}^{N}\langle (\delta \mathbf{u}_i)^2 \rangle_{P_G}^2 \tag{105}$$

Now, P_G, and therefore its free energy $-\beta F = \log Z$, average error Q_0, and fluctuations ΔQ all depend on the measurement matrix $\mathbf{A}$ and on the signal $\mathbf{s}^0$. We take these to be random variables; the matrix elements $A_{\mu i}$ are drawn independently from a standard normal distribution, while $\mathbf{s}^0$ has fN randomly chosen nonzero elements each drawn independently from a distribution $P(s^0)$. Thus $\mathbf{A}$ and $\mathbf{s}^0$ play the role

of quenched disorder in the thermal distribution P_G. In the limit $M, N \to \infty$ with $\alpha = M/N$ held fixed, we expect interesting observables, including the free energy, Q_0 and ΔQ to be self-averaging; i.e. the thermal average of these observables over P_G for any typical realization of $\mathbf{A}$ and $\mathbf{s}^0$ coincides with their thermal averages over P_G, further averaged over $\mathbf{A}$ and $\mathbf{s}^0$. Thus the typical error Q_0 does not depend on the detailed realization of $\mathbf{A}$ and $\mathbf{s}^0$. We can therefore compute Q_0 by computing the average free energy $-\beta\bar{F} \equiv \langle\langle \log Z \rangle\rangle_{\mathbf{A},\mathbf{s}^0}$ using the replica method.

Details of the replica calculation can be found in [121]. Basically, averaging over $\mathbf{A}$ in the replicated Gibbs distribution corresponding to the energy (102) reduces to averaging over the variables $b_\mu^a = \frac{1}{\sqrt{N}}\mathbf{a}_\mu \cdot \mathbf{u}^a$, where $\mathbf{u}^a$, $a = 1, \ldots, n$ are the replicated residuals. These variables are are jointly gaussian distributed with zero mean and covariance $\left\langle\!\left\langle \delta b_\mu^a \, \delta b_\nu^b \right\rangle\!\right\rangle = Q_{ab}\delta_{\mu\nu}$, where $Q_{ab} \equiv \frac{1}{T}\sum_{i=1}^T u_i^a u_i^b$. The replica method yields a set of saddle point equations for the overlap matrix Q. Given the convexity of the energy function (102), it is reasonable to choose a replica symmetric ansatz for the saddle point, $Q_{ab} = \Delta Q \delta_{ab} + Q_0$. Under this replica symmetric ansatz, further averaging over $\mathbf{s}^0$, and taking the $\lambda \to \infty$ limit, yields a set of self-consistent equations

$$Q_0 = \langle\langle \langle u \rangle_{H^{MF}}^2 \rangle\rangle_{z,s^0} \tag{106}$$

$$\Delta Q = \langle\langle \langle \delta u^2 \rangle_{H^{MF}} \rangle\rangle_{z,s^0}. \tag{107}$$

Here the thermal average $\langle\cdot\rangle_{H^{MF}}$ is performed with respect to a Gibbs distribution

$$P^{MF}(s \,|\, s^0) = \frac{1}{Z} e^{-H^{MF}}, \tag{108}$$

with an effective mean-field Hamiltonian

$$H^{MF} = \frac{\alpha}{2\Delta Q}\left(s - s^0 - z\sqrt{Q_0/\alpha}\right)^2 + \beta|s|, \tag{109}$$

where we make the substitution $s - s^0 = u$. Furthermore the quenched average $\langle\langle \cdot \rangle\rangle_{z,s^0}$ denotes an average over a standard normal variable z and the full distribution of the signal component s^0 (given by $f\delta(s^0) + (1-f)P(s^0)$).

The relationship between the mean field theory P^{MF} in (108) and the original Gibbs distribution P_G in (103) is as follows. The replica parameters Q_0 and ΔQ in (106) and (107) are identified with the order parameters (104) and (105). Thus solving (106) and (107) in the zero temperature $\beta \to \infty$ limit allows us to compute the typical error Q_0 of CS as a function of α and f. Furthermore, consider the marginal distribution of a single signal component $\mathbf{s}_k$ in $P_G(\mathbf{u}) = P_G(\mathbf{s} - \mathbf{s}^0)$, given the true signal component is $\mathbf{s}_k^0$. According to replica theory, the mean field theory prediction for the distribution of this marginal is given by

$$P_G(\mathbf{s}_k = s \,|\, \mathbf{s}_k^0 = s^0) = \left\langle\!\left\langle P^{MF}(s \,|\, s^0) \right\rangle\!\right\rangle_z, \tag{110}$$

where $P^{MF}(s \,|\, s^0)$ is defined by (108) and (109), and Q_0 and ΔQ are the solutions to (106) and (107).

Now in solving (106) and (107) in the $\beta \to \infty$ limit, one finds two distinct classes of solutions [121] depending on the values of α and f. For $\alpha > \alpha_c(f)$ one finds solutions

in which both ΔQ and Q_0 vanish as $O(\frac{1}{\beta^2})$. It is expected that thermal fluctuations captured by ΔQ should always vanish in the low temperature limit, but the fact that Q_0, which captures the typical error of CS, also vanishes, suggests that for $\alpha > \alpha_c(f)$, L_1 minimization should exactly recover the true signal, so that $\hat{\mathbf{s}} = \mathbf{s}^0$ in (101). On the otherhand, for $\alpha < \alpha_c(f)$, this class of solutions no longer exists, and instead a new class of solutions occurs in which ΔQ is $O(\frac{1}{\beta})$ but Q_0 remains $O(1)$ as $\beta \to \infty$. This class of solutions predicts an error regime in which $\hat{\mathbf{s}} \neq \mathbf{s}^0$ due to too few measurements. Thus replica theory predicts phase a transition between a perfect and an imperfect reconstruction regime in the α-f plane, as verified in Fig. 8A. This phase boundary was first derived in [123, 122] using very different methods of convex geometry.

The phase boundary simplifies in the $f \to 0$ limit of high sparsity. In this limit, $\alpha_c(f) = f \log 1/f$. This result can be understood from an information theoretic perspective. First, the entropy of a sparse signal of dimension N is $O(Nf \log 1/f)$. Second, assuming that each of our measurements carries $O(1)$ bits of entropy, and that they are not redundant or highly correlated, then the entropy of our measurements is $O(M)$. It will not be possible to perfectly reconstruct the signal using *any* reconstruction algorithm whatsoever, if the entropy of our measurements is less than the entropy of our signal. The requirement that the measurement entropy exceed the signal entropy then yields the inequality $\alpha = M/N > O(f \log 1/f)$. Thus from the perspective of information theory, it is not surprising that we can reconstruct the signal when $\alpha > \alpha_c(f)$. What is surprising is that a very simple, polynomial time algorithm, L_1 minimization, is capable of performing the reconstruction, down to a number of measurements that approaches the information theoretic limit at small f, up to constant factors.

What is the nature of this phase transition? For example if we decrease α from above $\alpha_c(f)$ to below, do we see a catastrophic rise in the error, or does performance gracefully degrade? In the language of statistical physics, does $Q_0(\alpha, f)$ undergo a first or second order phase transition in α? Fortunately, it is a second order phase transition, so that Q_0 rises continuously from 0. The exponent governing the rise depends on the distribution of non-zeros $P(s^0)$; namely the more confined this distribution is to the origin, the shallower the rise (see Fig. 8B). Note that the phase boundary $\alpha_c(f)$ in contrast is universal, in that it does not depend on the distribution of non-zeros in the signal.

Finally, we can understand the nature of the errors made by CS by looking at the distribution the signal reconstruction components conditioned on the true signal component. This is of course interesting only in the error regime. To take the zero temperature limit we can make the change of variables $\Delta Q = \frac{\alpha}{\beta}\Delta q$, and $Q_0 = \alpha q_0$ where Δq and q_0 remain $O(1)$ as $\beta \to \infty$. Then the mean field Hamiltonian in (109) becomes

$$H^{MF} = \beta\Big[\frac{1}{2\Delta q}\left(s - s^0 - z\sqrt{q_0}\right)^2 + |s|\Big]. \tag{111}$$

Since the entire Hamiltonian is proportional to β, in the large β limit, the statistics of

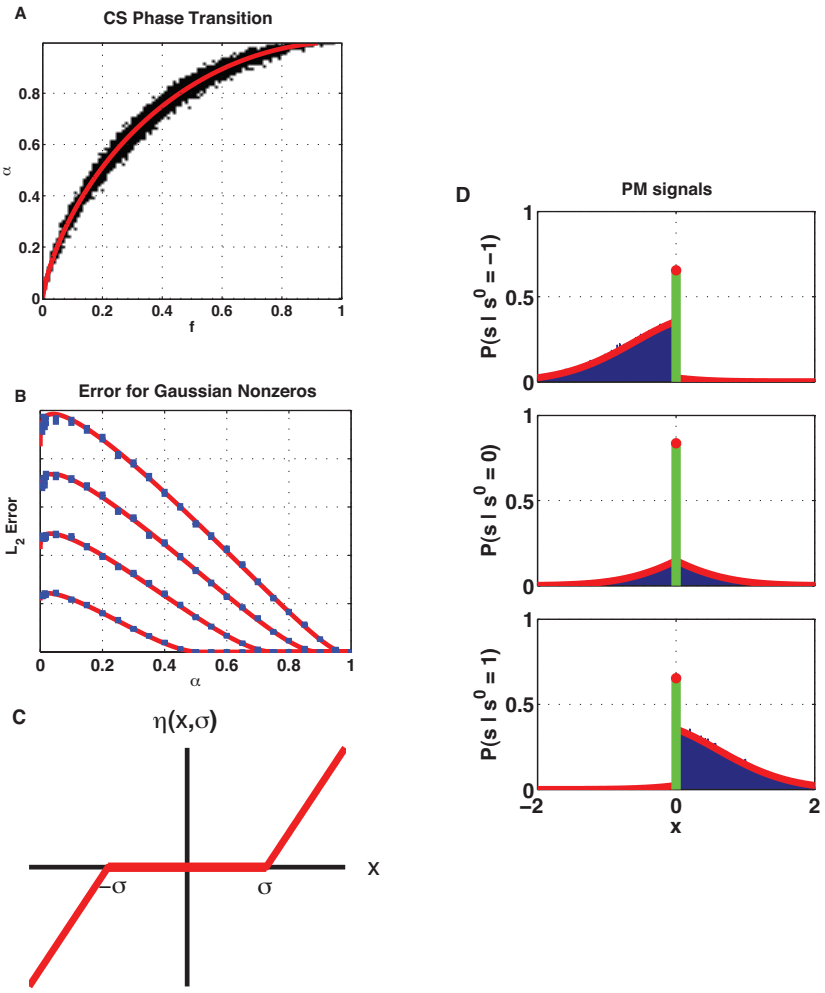

Figure 8. Compressed sensing analysis. (A) The red curve is the theoretical phase boundary $\alpha_c(f)$ obtained by solving (106) and (107). We also use linear programming to solve (101) 50 times for each value of α and f in increments of 0.01, with $N = 500$. The black transition region shows when the fraction of times perfect recovery occurs is neither 0 nor 1. For all other $\alpha > \alpha_c(f)$, we obtained perfect recovery all 50 times, and for all other $\alpha < \alpha_c(f)$ we never once obtained perfect recovery. The width of this transition region narrows as N is increased (not shown), yielding a sharp transition in the $N \to \infty$ limit at the phase boundary $\alpha_c(f)$. (B) Blue points are the average L_2 reconstruction error obtained by solving (101) 100 times for each for 4 values of $f = 0.2, 0.4, 0.6, 0.8$, and various α, with $N = 500$. Error bars reflect the standard error. Red curves are plots of Q_0 obtained by solving (106) and (107) in the error phase. (C) The soft thresholding function defined in (113). (D) The blue histograms are the conditional distribution of nonzero signal reconstruction components $\hat{\mathbf{s}}_k$ obtained from solving (101) 2000 times, while the height of the green bar represents the average fraction of components $\hat{\mathbf{s}}_k$ that are zero, all conditioned on the value of the true signal component s^0_k. Here $N = 500$, $\alpha = f = 0.2$, and $P(s^0) = \frac{1}{2}\delta(s^0 - 1) + \frac{1}{2}\delta(s^0 + 1)$. For these values of α and f, the order parameters were numerically found to take the values $q_0 = 1.06$ and $\Delta q = 1.43$. The red curves are the theoretically predicted distribution of nonzero reconstruction components in (114), while the red dot is the theoretically predicted height of the delta function at $s = 0$ predicted in (114), all conditioned on the three possible values of the true signal s^0, -1 (top), 0 (middle) and $+1$, bottom. Each distribution can be thought of as arising from a Gaussian distribution with mean s^0 and variance q_0, fed through the soft threshold function in (C), with a noise threshold $\sigma = \Delta q$.

s are dominated by the global minimum of (111). In particular, we have

$$\langle s \rangle_{HMF} = \eta(\, s^0 + z\sqrt{q_0}\,,\, \Delta q), \tag{112}$$

where

$$\eta(x\,, \sigma) = \text{argmin}_s \Big(\frac{1}{2} \frac{(s-x)^2}{\sigma} + |s| \Big) = \text{sgn}(x)(|x| - \sigma)_+, \tag{113}$$

is a soft thresholding function (see Fig. 8D) which also arises in message passing approaches [55] to solving the CS problem in (101), and $(y)_+ = y$ if $y > 0$ and is otherwise 0.

The optimization in (113) can be understood intuitively as follows: suppose one measures a scalar value x which is a true signal s^0 corrupted by additive gaussian noise with variance σ. Under a Laplace prior $e^{-|s^0|}$ on the true signal, $\eta(x, \sigma)$ is simply the MAP estimate of s^0 given the data x, which basically chooses the estimate $s = 0$ unless the data exceeds the noise level σ. Thus we see that in (111) and (112), $s^0 + z\sqrt{q_0}$ plays the role of the observed, corrupted data x and Δq plays the role of an effective noise level σ.

This optimization has an interpretation within the cavity method [119]. It is the optimization that a new signal component s added to a cavity system in the absence of that component must perform to minimize its total energy in (103). This minimization reflects a compromise between minimizing its own absolute value, and satisfying all the measurement constraints, whose sum total effect is encapsulated by the quadratic term in the mean field theory of (111). The cavity field encapsulating the effect of all other measurements will vary from component to component, and the average over components can be approximated by the average over the Gaussian variable z, in analogy to the SK model in going from (28) to (29).

The distribution of the signal reconstruction components, conditioned on the true signal component s^0 in (110) reduces to

$$P_G(\mathbf{s}_k = s \,|\, \mathbf{s}_k^0 = s^0) = \Big\langle \Big\langle\, \eta(s^0 + z\sqrt{q_0}, \Delta q) \,\Big\rangle \Big\rangle_z, \tag{114}$$

and it reflects the Gaussian distribution of the zero-temperature cavity fields across components, fed through the soft-thresholding function which arises from the scalar L_1 minimization problem in (113). Here q_0 and Δq are determined through (106) and (107), which can be thought of as a self-consistency condition within the cavity approximation demanding that the distribution across components of the cavity field is consistent with the distribution across components of the signal reconstruction, in analogy to the corresponding self-consistency condition (29) in the SK model. An example of the match between replica theory and simulations for the signal reconstruction distribution is shown in Fig. 8D.

6.3. From Message Passing to Network Dynamics

The L_1 minimization problem in (101) can also be formulated as a message passing problem [55, 124], and approximate formulations of the message passing dynamical

system yield neural network-like dynamics which provide a fast iterative way to solve the L_1 problem. The graphical model consists of N variable nodes, one for each component of the unknown signal $\mathbf{s}_i$, and M degree N factor nodes, one for each measurement, plus N more degree 1 factor nodes to implement the L_1 norm. For example, the Gibbs distribution defined by (102) and (103) decomposes as

$$P(\mathbf{s}) = \prod_{\mu=1}^{M} \psi_\mu(\mathbf{s}) \prod_{\imath=1}^{N} e^{-\beta|\mathbf{S}_i|}, \tag{115}$$

where the factor $\psi_\mu(\mathbf{s})$ is given by

$$\psi_\mu(\mathbf{s}) = e^{-\frac{1}{2N}\beta\lambda(x_\mu - \mathbf{a}_\mu \cdot \mathbf{s})^2}, \tag{116}$$

and $\mathbf{a}_\mu$ is row μ of the measurement matrix $\mathbf{A}$. This decomposition is in a form suitable for the application of the message passing equations (32) and (33). However, a straightforward application of these equations is computationally complex. First, since each unknown component $\mathbf{s}_i$ is a real number, every message becomes a distribution over the real numbers, and one must keep track of MN such messages (the degree 1 factors do not require associated messages and can be incorporated into the updates of the other messages).

The first approximation made in [55] is to restrict the messages to be Gaussian, which is reasonable because the density of the random measurement matrix $\mathbf{A}$ implies that in each update, each message receives contributions from a large number of messages, allowing one to invoke the central limit theorem. Thus one need keep track of only 2 numbers for each message, leading to a dynamical system on $2MN$ variables. This system can be further simplified by noting [55, 124] that messages from the same variable i to different factors μ are all quite similar to each other; they differ only in excluding the effects of one factor μ out of M possible factors. Thus one might assume that $M_{i\to\mu} = M_i + O(\frac{1}{\sqrt{M}})$. A similar argument holds for the factor to variable messages, suggesting $M_{\mu\to i} = M_\mu + O(\frac{1}{\sqrt{N}})$. By doing a careful expansion in $\frac{1}{N}$, one can then reduce the message passing equations to a dynamical system on $M + N$ variables.

A readable account of this reduction can be found in [124]. Here we simply quote the main result. The dynamical system on $M + N$ variables can be interpreted as an iterative update on two pairs of variables, a current estimate $\mathbf{s}^t$ for the unknown N dimensional signal, and the resulting residual $\mathbf{r}^t$ in M dimensional measurement space. The update equations are [124],

$$\mathbf{s}^{t+1} = \eta(\mathbf{s}^t + \mathbf{A}^T\mathbf{r}^t, \theta) \tag{117}$$

$$\mathbf{r}^t \quad = \mathbf{x} - \mathbf{A}\mathbf{s}^t + b\mathbf{r}^{t-1}, \tag{118}$$

where η is the soft-thresholding function defined in (113), and here is applied component wise to its vector inputs. In [124], it was shown that if these equations converge, then the resulting $\mathbf{s}^\infty$ is a global minimum of the L_1 energy function in (102) with $\frac{1}{\lambda} = \theta(1-b)$.

The crucial term is the message passing derived term involving b in (118) that endows the temporal evolution of the residual with a history dependence. Without this term, reconstruction performance is severely impaired. This dynamics can loosely be

interpreted as that of a two layer interacting neuronal network, where the residual $\mathbf{r}^t$ is stored in the activity of M neurons in the first layer, while the current estimate of the sparse representation $\mathbf{s}^t$ is stored in N neurons in a second layer. The first layer receives feedforward external input $\mathbf{x}$ and a top down inhibitory prediction of that input through synaptic connectivity $\mathbf{A}$. The second layer neurons receive a feedforward drive from the residuals through a synaptic connectivity $\mathbf{A}^T$ and have a nonlinear transfer function η. Interestingly, this network dynamics is different from other proposals for the implementation of L_1 minimization and sparse coding [48, 47]. Given the potential role of L_1 minimization as a computational description of early visual [56] and olfactory [57] processing, it would be interesting to explore more fully the space of neuronal architectures and dynamics capable of implementing L_1 minimization type computations.

7. Discussion

We have reviewed the applications of replicas, cavities, and message passing to basic models of network dynamics, memory storage, machine learning algorithms, and statistical models of data. While the ideas and models reviewed yield a rich and sometimes surprisingly striking picture, it is natural to ask how the above results are modified when assumptions in the models are relaxed, and what theoretical progress can be made through a statistical mechanics based analysis of these more complex scenarios. Here we will briefly discuss a few more prominent examples. However, we warn the reader that in this discussion we will barely be scratching the surface of a deep literature lying at the intersection of statistical mechanics, computer science and neuroscience.

7.1. Network Dynamics

First, in section 2, when we introduced the SK model and the Hopfield model, we were considering models of neuronal networks that had many simplifying assumptions, including: symmetric connectivity, binary neurons, and lack of external inputs. What happens for example, when the connectivity becomes asymmetric? Then there is no simple form for the stationary distribution of neuronal activity analogous to the equilibrium Gibbs distribution in (2). One must then posit a dynamical model for the network dynamics, and in many situations, one can use dynamic mean field theory methods [125, 126] to understand time averaged asymptotic statistical properties of neuronal activity. This was done in for example in [127, 128, 129] for asymmetric networks. Interestingly, while fully asymmetric networks become ergodic, partially asymmetric networks retain fixed points, but the time it takes for transients to reach these fixed points can diverge exponentially with network size.

Moreover dynamical versions of the SK and Hopfield models have a Lyapunov function that is bounded from below, implying that the long time asymptotic behavior

at zero temperature consists of fixed points only. In asymmetric networks, two more dynamical possibilities arise: oscillations and chaos. Seminal work showed analytically (via dynamic mean field theory) and through simulations that neuronal networks can exhibit deterministic high dimensional chaos [21, 130, 131]. Even when driven by a constant current input, such networks exhibit chaos by dynamically achieving a balance between excitatory and inhibitory inputs to individual neurons. This balance leads to spontaneous irregular neural activity characteristic of cortical states, in which neurons spike due to fluctuations in their input, as opposed to a mean superthreshold input current. Interestingly, when the inputs have more nontrivial temporal structure, such networks exhibit a sharp phase transition from a chaotic state to an ordered state, that is entrained by the input, as the input strength increases. This happens for example when the external input is either noisy [132] or oscillatory [133]. In the case of oscillatory input there is an interesting non-monotonic dependence in the input strength at which this phase transition occurs, as a function of oscillation frequency [133].

In section 2, we discussed binary models of neurons. However, biological neurons are characterized by analog internal states describing membrane voltage and ion channel conductance states, and their dynamics exhibits large spiking events in their membrane voltage. Dynamic mean field theory methods can be extended to spiking networks and were used to characterize the phase diagram of networks of excitatory and inhibitory leaky integrate and fire neurons [134]. For such neurons, whose internal state is characterized solely by a membrane voltage, one can derive an appropriate mean field theory by maintaining a distribution of membrane voltages across neurons in the network, and self-consistently solving for this distribution using Fokker-Planck methods (see [135, 136] for reviews). This work [134] lead to four possible macroscopic phases of network dynamics, characterized by two possibilities for the temporal statistics of single neurons (regular periodic spike trains, or irregular aperiodic spike trains) times two possibilities for the population average firing rates (synchronous or temporally structured rates, or asynchronous or constant rates). Varying strengths of excitation, inhibition, and single neuron properties allow all four combinations to occur.

More recently, in [137, 138] the authors went beyond mean field theory to track entire microstate trajectories in spiking neural networks consisting of neurons in which it is possible to analytically compute the time of the first neuron to spike next, given the internal state of all neurons in the network. This allowed the authors to perform numerically exact computations of the entire spectrum of Lyapunov exponents by computing products of Jacobians associated with every future spike starting from an initial condition. They found classes of networks that exhibit extensive chaos, in which a finite fraction of all Lyapunov exponents were positive. Moreover, they showed that the Lyapunov spectrum is highly sensitive to the details of the action potential shape, as positive feedback effects associated with the rise of the action potential contribute most heavily to the divergence of microstate trajectories. Even more interestingly, the authors found "flux" tubes of stability surrounding trajectories: small perturbations to the network state decayed quickly, whereas larger perturbations lead to an exponential

divergence between trajectories. Thus each trajectory is surrounded by a stability tube. However, the radius of this tube shrinks with the number of neurons, N. This reveals that the calculation of Lyapunov exponents in spiking networks in the thermodynamic ($N \to \infty$) limit is extremely subtle, due to the non-commutation of limits. Computing Lyapunov exponents requires taking a small perturbation limit, which if taken before the thermodynamic limit would yield negative exponents, but if taken after the thermodynamic limit, would yield positive exponents. In any case, injecting extra spikes into the network constitutes a large perturbation even at finite N, which leads to a divergence in trajectories. This picture is consistent with recent experimental results suggesting that the injection of extra spikes into a cortical network leads to a completely different spiking trajectory, without changing the overall population statistics of neural activity [139]. More generally, for reviews on network dynamics in neuroscience, see [140, 141].

7.2. Learning and Generalization

In the beginning of section 3, we considered the capacity of simple network architectures to store, or memorize a set of input-output mappings. While memory is certainly important, the goal of most organisms is not simply to memorize past responses to past inputs, but rather to generalize from past experience in order to learn rules that can yield appropriate responses to novel inputs the organism has never seen before. This idea has been formalized in a statistical mechanics framework for the perceptron in [142, 143]. Here the P training inputs and outputs are no longer random, but are generated from a "teacher" perceptron. The observable of interest then becomes the generalization error $\epsilon_g(\alpha)$, which is by definition the probability that the trained perceptron disagrees with the teacher perceptron's correct answer on a novel input, not present in the training set. Here $\alpha = \frac{P}{N}$ is the ratio of the number of training examples to the number of synapses N. For a wide variety of training procedures, or learning algorithms, statistical mechanics approaches found that $\epsilon_g(\alpha)$ decays as $O(\frac{1}{\alpha})$ for large α, indicating that the number of examples should be proportional to the number synapses in order for good generalization to occur.

The perceptron, while acting in some sense as the *Drosophila* of statistical learning theory, is a very limited architecture in that it can only learn linearly separable classifications in which the two classes fall on opposite sides of a hyperplane. Statistical mechanics approaches have been used to analyze memory [144, 145, 146] and generalization [147, 148, 149] in more sophisticated multilayered networks. In a multilayered network, only the input and output layers are constrained to implement a desired mapping, while the internal, hidden layer activities remain unspecified. This feature generically leads to replica symmetry breaking, where the space of solutions to a desired input-output mapping breaks into multiple disconnected components, where each component corresponds to a different internal representation of hidden layer activities capable of implementing the desired mapping. Statistical mechanics has also

had success in the analysis of learning in other architectures and machine learning algorithms, including support vector machines [150, 151, 152], and Gaussian processes [153].

Another generalization of the perceptron is the tempotron, an architecture and learning rule capable of learning to classify spatiotemporal patterns of incoming spikes [154]. The tempotron can be trained to fire a spike for one class of input spike time patterns, and no spikes for another class, while the precise timing of the output spike can be left unspecified. A statistical mechanics analysis of a simplified binary tempotron was carried out in [155]. Interestingly, the space of solutions in synaptic weight space to any given spike time classification problem can be well described by a one-step replica symmetry broken phase shown schematically in Fig. 1C. Each component corresponds to a different output spike time for the positive classifications, in direct analogy to the replica symmetry broken phase of multilayered networks in which each component corresponds to a different internal representation. The various solution components are both small (implying that very similar weights can yield very different classifications) and far apart (implying that very different weights can yield an identical classification). The authors verified that these properties persist even in a more biologically realistic Hodgkin-Huxley model of a single neuron [155]. Overall this reveals a striking double dissociation between structure (synaptic connectivity) and function (implemented classification) even at the level of single neurons. This double dissociation has important implications for the interpretation of incoming connectomics data [156]. More generally, for reviews on applications of statistical mechanics to to memory, learning and generalization, see [12, 157].

7.3. Machine Learning and Data Analysis

Starting in the latter part of section 3, we turned our attention to the statistical mechanics based analysis of machine learning algorithms designed to extract structured patterns from data, focusing on illusions of structure returned by such algorithms when applied to high dimensional noise. In real data analysis problems, we have to protect ourselves from such illusions, and so understanding these illusions present in pure noise is an important first step. However, we would ideally like to analyze the performance of machine learning algorithms when data contain both structured patterns as well as random noise. A key question in the design and analysis of experiments is then how much data do we need to reliably uncover structure buried within noise? Since, in the statistical mechanics based analysis of learning algorithms, the data plays the role of quenched disorder, we must analyze statistical mechanics problems in which the quenched disorder is no longer simply random, but itself has structure.

This has been done for example in [158, 96] for PCA applied to signals confined to a low dimensional linear space, but corrupted by high dimensional noise. The data is then drawn from a covariance matrix consisting of the identity plus a low rank part. A replica based computation of the typical eigenvalue spectrum of empirical covariance

matrices for data of this type revealed the presence of a series of phase transitions as the ratio between the amount of data and its ambient dimensionality increases. As this ratio increases, signal eigenvalues associated with the low rank part pop out of a Marcenko-Pasteur sea (i.e. Fig.6A) associated with the high dimensional noise. Thus this work reveals sharp thresholds in the amount of data required to resolve signal from noise. Also, interesting work has been done on the statistical mechanics based analysis of typical learning outcomes for other structured data settings, including finding a direction separating two Gaussian clouds [159, 160], supervised learning from clustered input examples [161], phase transitions in clustering as a function of cluster scale [162], and learning Gaussian mixture models [163, 164]. Moreover, statistical mechanics approaches to clustering have yielded interesting new algorithms and practical results, including superparamagnetic clustering [165, 166], based on an isomorphism between cluster assignments and Potts model ground states, and a method for computing p-values for cluster significance [167], using extreme value theory to compute a null distribution for the maximal number of data points over all regions in feature space of a given size.

In section 5.3 we initiated a statistical mechanics based analysis of random dimensionality reduction by connecting the maximally incurred geometric distortion to correlated extreme value theory. For the simple case of point clouds, correlations could be neglected, while for hyperplanes, the correlations arose from fluctuations in the Coulomb gas interactions of eigenvalues of random matrices, and could be treated exactly. It would be interesting to study more complex manifolds. For example, rigorous upper bounds on the maximal distortion were proven in [104] by surrounding arbitrary manifolds and their tangent planes by a scaffold of points, and then showing that if the geometry of this scaffold remains undistorted under any projection, then so does the geometry of the manifold. An application of the JL lemma to the scaffold then suffices to obtain an upperbound on the distortion incurred by the manifold under a random projection. To understand how tight or loose this upperbound is, it would useful to compute the typical distortion incurred by more complex manifold ensembles. For example, for manifolds consisting of unions of planes, one would be interested in the fluctuations of the maximal eigenvalue of multiple correlated matrices, corresponding to the restriction of the same random projection to each plane. Thus results from the eigenvalue spectra of random correlated matrices [168, 169, 170] could become relevant.

Finally, we note that throughout most of this review we have focused on situations in which replica symmetry holds, though we have noted that several neuronal and machine learning problems, including multilayer networks, tempotrons, and clustering, are described by replica symmetry broken phases in which the solution space breaks up into many clusters, as well as suboptimal, higher energy metastable states. As noted at the end of section 2.4, statistical mechanics based approaches have inspired a new algorithm, known as survey propagation [77, 78] that can find good solutions despite the proliferation of metastable states whose presence can confound simpler optimization and inference algorithms. Despite the power of survey propagation, its applications to neuroscience remain relatively unexplored.

In summary, decades of interactions between statistical physics, computer science, and neuroscience, have lead to beautiful insights, both into how neuronal dynamics leads to computation, as well as how our brains might create machine learning algorithms to analyze themselves. We suspect that further interactions between these fields are likely to provide exciting and insightful intellectual adventures for many years to come.

7.4. Acknowledgements

We thank DARPA, the Swartz foundation, Burroughs-Wellcome Foundation, Genentech Foundation, and Stanford Bio-X Neuroventures for support. S.G. thanks Haim Sompolinsky for many interesting discussions about replicas, cavities, and messages.

8. Appendix: Replica Theory

8.1. Overall Framework

Suppose we wish to do statistical mechanics on a set of N thermal degrees of freedom encoded in the N dimensional vector $\mathbf{x}$, where the components are coupled to each other through some quenched disorder $\mathbf{D}$, in a Hamiltonian $H(\mathbf{x}, \mathbf{D})$. In the above applications, $\mathbf{x}$ could be the spins $\mathbf{s}$ in a spin glass (then $\mathbf{D}$ is the connectivity matrix $\mathbf{J}$), the synaptic weights $\mathbf{w}$ of a perceptron (then $\mathbf{D}$ is the set of examples to be stored), the variables $\mathbf{u}$ in the Stieltjes transform of an eigenvalue spectrum (then $\mathbf{D}$ is a random matrix), or the residuals $\mathbf{u}$ in a compressed sensing problem (then $\mathbf{D}$ is the measurement matrix). As discussed above, to properly average over the quenched disorder $\mathbf{D}$, we must average the replicated partition function

$$\langle\langle\, Z^n \,\rangle\rangle_{\mathbf{D}} = \left\langle\left\langle \int \prod_{a=1}^{n} d\mathbf{x}^a e^{-\sum_{a=1}^{n} H(\mathbf{x}^a, \mathbf{D})} \right\rangle\right\rangle_{\mathbf{D}}. \tag{119}$$

Conditioned on any particular realization of the quenched disorder $\mathbf{D}$, the different replicated degrees of freedom $\mathbf{x}^a$ are independent. However, integrating out the quenched disorder introduces interactions among the replicated variables. In all of the above applications, the resulting interactions depend only on the overlap matrix between replicas, defined as $Q_{ab} = \frac{1}{N}\mathbf{x}^a \cdot \mathbf{x}^b$. More precisely, the following identity holds,

$$\left\langle\left\langle e^{-\sum_{a=1}^{n} H(\mathbf{x}^a, \mathbf{D})} \right\rangle\right\rangle_{\mathbf{D}} = e^{-NE(Q)}, \tag{120}$$

for some function E over the overlap matrix Q. Therefore it is useful to separate the remaining integral over $\mathbf{x}^a$ in (119) into an integral over all possible overlaps Q_{ab}, and then all possible $\mathbf{x}^a$ configurations with a prescribed set of overlaps by introducing a δ-function:

$$\langle\langle\, Z^n \,\rangle\rangle_{\mathbf{D}} = \int \prod_{ab} dQ_{ab}\, e^{-NE(Q)} \int \prod_{a=1}^{n} d\mathbf{x}^a \prod_{ab} \delta[\mathbf{x}^a \cdot \mathbf{x}^b - NQ_{ab}]. \tag{121}$$

The integral over $\mathbf{x}^a$ with a fixed set of overlaps Q_{ab} can be carried out by introducing the exponential representation of the δ function,

$$\delta[\mathbf{x}^a \cdot \mathbf{x}^b - NQ_{ab}] = \int d\hat{Q}_{ab}\, e^{-\hat{Q}_{ab}(\mathbf{x}^a \cdot \mathbf{x}^b - NQ_{ab})}, \tag{122}$$

where the integral over $\hat{Q}_{ab}$ is understood to be along the imaginary axis. Inserting (122) into (121) decouples the components of the vectors $\mathbf{x}^a$, yielding an integral over n scalar variables x^a raised to the N'th power. This final result can be written as

$$\langle\langle\, Z^n \,\rangle\rangle_{\mathbf{D}} = \int \prod_{ab} dQ_{ab}\; d\hat{Q}_{ab}\, e^{-N[E(Q) - \sum_{ab} Q_{ab}\hat{Q}_{ab} + G(\hat{Q}_{ab})]}, \tag{123}$$

where

$$G(\hat{Q}_{ab}) = -\ln \int \prod_{a} dx^a e^{-H_{\text{eff}}(x^1,\ldots,x^n)}, \tag{124}$$

is the partition function of an effective Hamiltonian

$$H_{\text{eff}} = \sum_{ab} x^a\, \hat{Q}_{ab}\, x^b. \tag{125}$$

Now in the large N limit, the final integrals over Q_{ab} and $\hat{Q}_{ab}$ can be done via the saddle point method, yielding a set of self-consistent equations for the saddle point by extremizing the exponent in (123):

$$\hat{Q}_{ab} = \frac{\partial E}{\partial Q_{ab}} \tag{126}$$

$$Q_{ab} = \langle x^a x^b \rangle_n, \tag{127}$$

where $\langle\cdot\rangle_n$ denotes an average with respect to a Gibbs distribution with effective Hamiltonian H_{eff} in (125).

In general both these equations must be solved in the $n \to 0$ limit. Now in the case where $\mathbf{x}_i^a$ are real valued-variables (as opposed to binary variables in the SK model), these equations can be further simplified because the integral over x^a in (124) can be done exactly, since it is Gaussian, and furthermore, the extremum over $\hat{Q}_{ab}$ in (124) can be performed. Together, this yields an entropic factor (up to a multiplicative constant involving n and N),

$$\int \prod_{a=1}^{n} d\mathbf{x}^a \prod_{ab} \delta[\mathbf{x}^a \cdot \mathbf{x}^b - N Q_{ab}] = \mathrm{e}^{NS(Q)}, \tag{128}$$

where

$$S(Q) = \frac{1}{2} \operatorname{Tr} \log Q \tag{129}$$

represents (up to an additive constant) the entropy of replicated configurations $\mathbf{x}^a$ with a prescribed overlap matrix Q. (123) reduces to

$$\langle\langle\, Z^n \,\rangle\rangle_{\mathbf{D}} = \int \prod_{ab} dQ_{ab}\, e^{-N[E(Q) - S(Q))]}, \tag{130}$$

and the saddle point overlap configuration Q represents a compromise between energy and entropy extremization in the exponent of (130).

8.2. Physical meaning of overlaps

Here we make the connection between the replica overlap matrix Q_{ab} and the disorder averaged distribution of overlaps, $P(q)$, of two states $\mathbf{x}^1$ and $\mathbf{x}^2$ both drawn from a Gibbs distribution with Hamiltonian $H(\mathbf{x}, \mathbf{D})$. For a given realization of the disorder, the overlap distribution is

$$P_{\mathbf{D}}(q) = \frac{1}{Z(\mathbf{D})^2} \int d\mathbf{x}^1 \, d\mathbf{x}^2 \, \delta\left(q - \frac{1}{N}\mathbf{x}^1 \cdot \mathbf{x}^2\right) e^{-H(\mathbf{x}^1,\mathbf{D}) - H(\mathbf{x}^2,\mathbf{D})} \tag{131}$$

where $Z(\mathbf{D}) = \int d\mathbf{x}\, e^{-H(\mathbf{x},\mathbf{D})}$. Averaging $P_{\mathbf{D}}(q)$ over the disorder is difficult because $\mathbf{D}$ appears both in the numerator and denominator of (131). To circumvent this, one can introduce replicas via the simple identity $Z^{-2} = \lim_{n\to 0} Z^{n-2}$. Using this, one can perform the easier average at integer $n > 2$, and then take the limit $n \to 0$ at the end. Thus

$$P(q) = \langle\langle\, P_{\mathbf{D}}(q)\,\rangle\rangle_{\mathbf{D}} \tag{132}$$

$$= \lim_{n\to 0} \left\langle \left\langle \int \prod_{a=1}^{n} d\mathbf{x}^a e^{-\sum_{a=1}^{n} H(\mathbf{x}^a,\mathbf{D})}\, \delta(q - \tfrac{1}{N}\mathbf{x}^1 \cdot \mathbf{x}^2) \right\rangle \right\rangle_{\mathbf{D}} \tag{133}$$

Here $\mathbf{x}^1$ and $\mathbf{x}^2$ are the original degrees of freedom with $n-2$ additional replicas added to yield Z^{n-2}. One can then average the right hand side of (133) over $\mathbf{D}$ using a sequence of steps very similar to section 8.1. The final answer yields

$$P(q) = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{a\neq b} \delta(q - Q_{ab}) \tag{134}$$

where Q_{ab} is the saddle point replica overlap matrix. In situations where replica symmetry is broken, there will be multiple equivalent saddle points related to each other by the action of the permutation group on the replica indices a, b. The sum over these saddle points yields the sum in (134). In summary, the probability that two states have overlap q is, according to replica theory, equal to the fraction of off-diagonal matrix elements Q_{ab} that take the value q.

8.3. Replica symmetric equations

Here we show how to take the $n \to 0$ limit for various problems, in the replica symmetric approximation. We use Einstein summation convention in which repeated indices are meant to be summed over.

8.3.1. SK Model We will now apply (124)-(127) to the SK model from section 2. As we saw in (10), we have

$$E(Q) = -\left(\frac{\beta}{2}\right)^2 Q_{ab}^2. \tag{135}$$

Then, (126) gives

$$\hat{Q}_{ab} = -\frac{\beta^2}{2} Q_{ab}, \tag{136}$$

$$E(Q) - Q_{ab}\hat{Q}_{ab} = \left(\frac{\beta}{2}\right)^2 Q_{ab}^2. \tag{137}$$

We make the following replica-symmetric ansatz for the saddle point:

$$Q_{ab} = q + (1-q)\delta_{ab}, \tag{138}$$

where we used the fact that (11) guarantees that $Q_{aa} = 1$. We will determine q by minimizing the free energy. This leads to

$$E(Q) - Q_{ab}\hat{Q}_{ab} = \left(\frac{\beta}{2}\right)^2 \left[(1-q^2)n + q^2 n^2\right] \tag{139}$$

$$H_{\text{eff}} = -\frac{\beta^2}{2}\left[(1-q)n + q\left(\sum_a s^a\right)^2\right]. \tag{140}$$

We can now evaluate (124) using the identity (9) with $\sigma = 1$:

$$\begin{aligned} G(Q) &= -\ln\left[\sum_{\{s^a\}} e^{\frac{\beta^2}{2}\left[(1-q)n + q\left(\sum_a s^a\right)^2\right]}\right] \\ &= -\frac{\beta^2}{2}(1-q)n - \ln\left[\sum_{\{s^a\}} \left\langle\!\left\langle\, e^{\beta\sqrt{q}z\sum_a s^a} \,\right\rangle\!\right\rangle_z\right] \\ &= -\frac{\beta^2}{2}(1-q)n - \ln\left[\left\langle\!\left\langle\, \left[2\cosh(\beta\sqrt{q}z)\right]^n \,\right\rangle\!\right\rangle_z\right]. \end{aligned}$$

This gives us the free energy density:

$$\begin{aligned} \left\langle\!\left\langle\, \tfrac{\beta F}{N} \,\right\rangle\!\right\rangle_{\mathbf{J}} &= -\frac{1}{N}\frac{\partial}{\partial n}\langle\langle\, Z^n \,\rangle\rangle_{\mathbf{J}}\Big|_{n=0} \\ &= -\left(\frac{\beta J}{2}\right)^2 (1-q)^2 - \left\langle\!\left\langle\, \ln[2\cosh(\beta\sqrt{q}z)] \,\right\rangle\!\right\rangle_z. \end{aligned} \tag{141}$$

As mentioned above, we determine q by minimizing this. We will need the identity

$$\langle\langle\, z\, f(z) \,\rangle\rangle_z = \langle\langle\, f'(z) \,\rangle\rangle_z,$$

which can be derived with integration by parts. We find

$$\begin{aligned} \frac{\partial}{\partial q}\left\langle\!\left\langle\, \tfrac{\beta F}{N} \,\right\rangle\!\right\rangle_{\mathbf{J}} &= \frac{\beta^2}{2}(1-q) - \frac{\beta}{2\sqrt{q}}\left\langle\!\left\langle\, z\tanh(\beta\sqrt{q}z) \,\right\rangle\!\right\rangle_z \\ &= \frac{\beta^2}{2}\left(1 - q - \left\langle\!\left\langle\, \operatorname{sech}^2(\beta\sqrt{q}z) \,\right\rangle\!\right\rangle_z\right) \\ &= \frac{\beta^2}{2}\left(\left\langle\!\left\langle\, \tanh^2(\beta\sqrt{q}z) \,\right\rangle\!\right\rangle_z - q\right), \end{aligned} \tag{142}$$

therefore, the minimum satisfies (14).

8.3.2. Perceptron and Unsupervised Learning The starting point for the learning applications discussed here is the energy (60) and entropy (129) in the replicated partition function (130). These can be derived by following sections 3.3 and 8.1. Here we take the $n \to 0$ limit. For the energy, we obtain

$$\lim_{n\to 0} E(Q) = -\alpha \ln \int \prod_{a=1}^{n} \frac{d\lambda_a}{\sqrt{2\pi}} \frac{1}{\sqrt{\det Q}} e^{-\frac{1}{2}\lambda_a Q^{-1}_{ab}\lambda_b - \sum_a \beta V(\lambda_a)} \tag{143}$$

$$= -\alpha \ln \int \prod_{a=1}^{n} \frac{d\lambda_a \, d\hat{\lambda}_a}{2\pi} \, e^{i\lambda_a \hat{\lambda}_a - \frac{1}{2}\hat{\lambda}_a Q_{ab} \hat{\lambda}_b - \sum_a \beta V(\lambda_a)}$$

$$= -\alpha \ln \int \prod_{a=1}^{n} \frac{d\lambda_a \, d\hat{\lambda}_a}{2\pi} \, e^{i\lambda_a \hat{\lambda}_a - \frac{1}{2}(1-q)\sum_a (\hat{\lambda}_a)^2 - \frac{1}{2}(\sqrt{q}\sum_a \hat{\lambda}_a)^2 - \sum_a \beta V(\lambda_a)} \tag{144}$$

$$= -\alpha \ln \langle\langle \, \zeta^n \, \rangle\rangle_z \tag{145}$$

$$= -n\alpha \langle\langle \, \ln \zeta \, \rangle\rangle_z, \tag{146}$$

where

$$\zeta = \int \frac{d\lambda \, d\hat{\lambda}}{2\pi} \, e^{i\hat{\lambda}(\lambda - \sqrt{q}z) - \frac{1}{2}(1-q)\hat{\lambda}^2 - \beta V(\lambda)}$$

$$= \int \frac{d\lambda}{\sqrt{2\pi(1-q)}} \, e^{-\frac{1}{2}\frac{(\lambda - \sqrt{q}z)^2}{1-q} - \beta V(\lambda)} \tag{147}$$

is the partition function of a distribution whose interpretation will be given in section 8.4.

In going from (143) to (144) we used the identity

$$\int \frac{d\lambda_a}{\sqrt{2\pi}} \frac{1}{\sqrt{\det Q}} \, e^{-\frac{1}{2}\lambda_a Q^{-1}_{ab} \lambda_b} = \int \prod_{a=1}^{n} \frac{d\lambda_a \, d\hat{\lambda}_a}{2\pi} \, e^{i\lambda_a \hat{\lambda}_a - \frac{1}{2}\hat{\lambda}_a Q_{ab} \hat{\lambda}_b}, \tag{148}$$

and inserted the replica symmetric ansatz $Q_{ab} = (1-q)\delta_{ab} + q$. Then the only coupling between the various λ_a's in (144) occurs through the term $\frac{1}{2}(\sqrt{q}\sum_a \hat{\lambda}_a)^2$. We can thus decouple the λ_a variables at the expense of introducing a Gaussian integral via the identity $e^{-\frac{1}{2}b^2} = \left\langle\left\langle \, e^{ibz} \, \right\rangle\right\rangle_z$, where z is a zero mean, unit variance Gaussian variable and $\langle\langle \cdot \rangle\rangle_z$ denotes an average with respect to z. This transformation yields (145), and, as $n \to 0$, (146).

Now for the entropy, we obtain

$$\lim_{n\to 0} S(Q) = \lim_{n\to 0} \frac{1}{2} \operatorname{Tr} \log Q \tag{149}$$

$$= \frac{n}{2}\Big[\frac{q}{1-q} + \ln(1-q)\Big], \tag{150}$$

Here we have used the fact that the replica symmetric Q has 1 eigenvalue equal to $1 + (n-1)q$ and $n-1$ eigenvalues equal to $1-q$. Finally, inserting (146) and (150) into (130) and performing the integration over q via a saddle point yields a saddle point equation for q corresponding to extremizing $F(q)$ in (62).

8.4. Distribution of Alignments

Suppose we wish to compute the probability distribution across examples μ of the alignment of each example $\boldsymbol{\xi}^\mu$ with an optimal weight vector $\mathbf{w}$ derived from an unsupervised learning problem. Alternatively, one can think of this as the distribution of the data projected onto the optimal dimension. This distribution is

$$P(\lambda) = \frac{1}{P} \sum_{\mu=1}^{P} \delta(\lambda - \lambda^\mu), \tag{151}$$

where $\lambda^\mu = \frac{1}{\sqrt{N}}\mathbf{w}\cdot\boldsymbol{\xi}^\mu$, and $\mathbf{w}$ is drawn from the distribution (53). For large N and P we expect this distribution to be self-averaging, so for any fixed realization of the examples, it will be close to

$$P(\lambda) = \left\langle\!\!\left\langle \tfrac{1}{Z}\int d\mathbf{w}\,\delta(\lambda-\lambda^1)\,e^{-\beta\sum_\mu V(\lambda^\mu)} \right\rangle\!\!\right\rangle, \tag{152}$$

where

$$Z = \int d\mathbf{w}\,e^{-\beta\sum_\mu V(\lambda^\mu)}, \tag{153}$$

and $\langle\langle\,\cdot\,\rangle\rangle$ denotes an average over the examples $\boldsymbol{\xi}^\mu$. This average is hard to perform because the examples occur both in the numerator and denominator. This difficulty can be circumvented by introducing replicas via the simple identity $\frac{1}{Z} = \lim_{n\to 0} Z^{n-1}$. Thus

$$P(\lambda) = \lim_{n\to 0}\left\langle\!\!\left\langle \int \prod_{a=1}^n d\mathbf{w}_a\,\delta(\lambda-\lambda_1^1)\,e^{-\beta\sum_{a=1}^n\sum_\mu V(\lambda_a^\mu)} \right\rangle\!\!\right\rangle, \tag{154}$$

where $\lambda_a^\mu = \frac{1}{\sqrt{N}}\mathbf{w}_a\cdot\boldsymbol{\xi}^\mu$. Here the first replica plays the role of the numerator in (152) and replicas $2,\ldots,n$ play the role of $\frac{1}{Z}$ in the $n\to 0$ limit. Now we can introduce an integral representation of $\delta(\lambda-\lambda_1^1)$, perform the Gaussian average over λ_a^μ and take the $n\to 0$ limit using a sequence of steps very similar to those in sections 8.1 and 8.3.2. This yields

$$P(\lambda) = \left\langle\!\!\left\langle \tfrac{1}{\zeta}\tfrac{1}{\sqrt{2\pi(1-q)}}\,e^{-\frac{1}{2}\frac{(\lambda-\sqrt{q}z)^2}{1-q}-\beta V(\lambda)} \right\rangle\!\!\right\rangle_z, \tag{155}$$

where ζ is the partition function given by (63) and q extremizes the free energy (62).

8.5. Inverting the Stieltjies Transform

It is helpful to think of (70) as a complex contour integral, with the contour running along the real axis. We can't simply set $\epsilon = 0$ in (71), as the pole at $z' = z + i\epsilon$ would hit the contour. However, Cauchy's theorem tells us that we can deform the contour without changing the integral, provided that it doesn't cross any singularities. We will use the following contour, which takes a semicircle detour below the singularity:

$$C(\delta): \quad \begin{aligned} z' &= x, && x \in (-\infty, -\delta], \\ z' &= z + \delta\,e^{i\theta}, && \theta \in [-\pi, 0], \\ z' &= x, && x \in [\delta, \infty). \end{aligned}$$

It will help to take the limit $\delta \to 0$.

We can write

$$\begin{aligned} \lim_{\epsilon\to 0^+}\frac{1}{\pi}\operatorname{Im} R_{\mathbf{w}}(z+i\epsilon) &= \lim_{\delta\to 0^+}\frac{1}{\pi}\operatorname{Im}\int_{C(\delta)}\mathrm{d}z'\,\frac{\rho_{\mathbf{w}}(z')}{z'-z} \\ &= \lim_{\delta\to 0^+}\frac{1}{\pi}\operatorname{Im}\left[\int_{-\infty}^{-\delta}\mathrm{d}x\,\frac{\rho_{\mathbf{w}}(x)}{x-z}\right. \\ &\qquad \left. + \int_{-\pi}^{0}\mathrm{d}\theta\,\left(i\delta\,e^{i\theta}\right)\frac{\rho_{\mathbf{w}}\left(z+\delta\,e^{i\theta}\right)}{\delta\,e^{i\theta}} + \int_{\delta}^{\infty}\mathrm{d}x\,\frac{\rho_{\mathbf{w}}(x)}{x-z}\right]. \end{aligned}$$

The first and third terms diverge as $\delta \to 0$. However, their sum is finite. It is referred to as the Cauchy principal value of the integral. It also happens to be real, and we are only interested in the imaginary part. This leaves the second term:

$$\lim_{\epsilon \to 0^+} \frac{1}{\pi} \operatorname{Im} R_{\mathbf{W}}(z + i\epsilon) = \lim_{\delta \to 0^+} \frac{1}{\pi} \operatorname{Im} \int_{-\pi}^{0} \mathrm{d}\theta \, i\rho_{\mathbf{W}} \left(z + \delta \, e^{i\theta} \right) = \rho_{\mathbf{W}}(z).$$

References

[1] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principles of neural science*. Appleton & Lange, 1991.

[2] P. Dayan and LF Abbott. *Theoretical neuroscience. Computational and mathematical modelling of neural systems*. MIT Press, Cambridge, MA, 2001.

[3] F. Reike, D. Warland, R. van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1996.

[4] K.Y. Lau, S. Ganguli, and C. Tang. Function constrains network architecture and dynamics: a case study on the yeast cell cycle boolean network. *Phys. Rev. E*, 75(5 Pt 1):051907–051907, May 2007.

[5] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, USA, 2009.

[6] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792–1796, 1975.

[7] M. Mezard, G. Parisi, and M.A. Virasoro. *Spin glass theory and beyond*. World scientific Singapore, 1987.

[8] K.H. Fischer and J.A. Hertz. *Spin glasses*, volume 1. Cambridge University Press, 1993.

[9] H. Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Oxford University Press, USA, 2001.

[10] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[11] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[12] A. Engel and C. V. den Broeck. *Statistical Mechanics of Learning*. Cambridge Univ. Press, 2001.

[13] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[14] HD Block. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1):123, 1962.

[15] A.L. Blum and R.L. Rivest. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117–127, 1992.

[16] E. Amaldi. On the complexity of training perceptrons. *Artificial Neural Networks*, 1:55–60, 1991.

[17] A. Braunstein and R. Zecchina. Learning by message passing in networks of discrete synapses. *Physical review letters*, 96(3):30201, 2006.

[18] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina. Efficient supervised learning in networks with binary synapses. *BMC Neuroscience*, 8(Suppl 2):S13, 2007.

[19] M.L. Mehta. *Random matrices*, volume 142. Academic press, 2004.

[20] G. Akemann, J. Baik, and P. Di Francesco. *The Oxford Handbook of Random Matrix Theory*. Oxford Handbooks in Mathematics. OUP Oxford, 2011.

[21] H. Sompolinsky, A. Crisanti, and HJ Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262, 1988.

[22] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20(1/2):32–52, 1928.

[23] V.A. Marchenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[24] C.A. Tracy and H. Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994.

[25] C.A. Tracy and H. Widom. Distribution functions for largest eigenvalues and their applications. *arXiv preprint math-ph/0210034*, 2002.

[26] P. Vivo, S.N. Majumdar, and O. Bohigas. Large deviations of the maximum eigenvalue in wishart random matrices. *Journal of Physics A: Mathematical and Theoretical*, 40(16):4317, 2007.

[27] S.N. Majumdar and M. Vergassola. Large deviations of the maximum eigenvalue for wishart and gaussian random matrices. *Physical review letters*, 102(6):60601, 2009.

[28] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *Siam Review*, 51(1):34–81, 2009.

[29] E. Candes and M. Wakin. An introduction to compressive sampling. *IEEE Sig. Proc. Mag.*, 25(2):21–30, 2008.

[30] S. Ganguli and H. Sompolinsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.*, 35:485–508, 2012.

[31] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing mri. *Signal Processing Magazine, IEEE*, 25(2):72–82, 2008.

[32] M. Lustig, D. Donoho, and J.M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[33] T. Parrish and X. Hu. Continuous update with random encoding (cure): a new strategy for dynamic imaging. *Magnetic resonance in medicine*, 33(3):326–336, 1995.

[34] W. Dai, M.A. Sheikh, O. Milenkovic, and R.G. Baraniuk. Compressive sensing dna microarrays. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009.

[35] T. Hu and D.B. Chklovskii. Reconstruction of sparse circuits using multi-neuronal excitation (rescume). *Advances in Neural Information Processing Systems 22*, 2009.

[36] Y. Mishchenko. Reconstruction of complete connectivity matrix for connectomics by sampling neural connectivity with fluorescent synaptic markers. *Journal of neuroscience methods*, 196(2):289–302, 2011.

[37] B.A. Wilt, L.D. Burns, E.T.W. Ho, K.K. Ghosh, E.A. Mukamel, and M.J. Schnitzer. Advances in light microscopy for neuroscience. *Annual review of neuroscience*, 32:435, 2009.

[38] J.W. Taraska and W.N. Zagotta. Fluorescence applications in molecular neurobiology. *Neuron*, 66(2):170–189, 2010.

[39] A.F. Coskun, I. Sencan, T.W. Su, and A. Ozcan. Lensless wide-field fluorescent imaging on a chip using compressive decoding of sparse objects. *Optics express*, 18(10):10510, 2010.

[40] D. Takhar, J.N. Laska, M. Wakin, M.F. Duarte, D. Baron, S. Sarvotham, K.F. Kelly, and R.G. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. *IS&T/SPIE Computational Imaging IV*, 6065, 2006.

[41] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, T. Sun, K.F. Kelly, and R.G. Baraniuk. Single-pixel imaging via compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):83–91, 2008.

[42] T.T. Rogers and J.L. McClelland. *Semantic cognition: A parallel distributed processing approach*. The MIT Press, 2004.

[43] R. Kiani, H. Esteky, K. Mirpour, and K. Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296, 2007.

[44] N. Kriegeskorte, M. Mur, D.A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P.A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

[45] S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proc. Natl. Acad. Sci.*, 105(48):18970, 2008.

[46] S. Ganguli and H. Sompolinsky. Short-term memory in neuronal networks through dynamical compressed sensing. In *Neural Information Processing Systems (NIPS)*, 2010.

[47] C.J. Rozell, D.H. Johnson, R.G. Baraniuk, and B.A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20(10):2526–2563, 2008.

[48] B.A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[49] L.U. Perrinet. Role of homeostasis in learning sparse representations. *Neural computation*, 22(7):1812–1836, 2010.

[50] A. Lage-Castellanos, A. Pagnani, and M. Weigt. Statistical mechanics of sparse generalization and graphical model selection. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P10009, 2009.

[51] W.K. Coulter, C.J. Hillar, G. Isley, and F.T. Sommer. Adaptive compressed sensing—a new class of self-organizing coding models for neuroscience. In *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 5494–5497. IEEE, 2010.

[52] Guy Isely, Christopher J. Hillar, and Friedrich T. Sommer. Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication. *Advances in Neural Information Processing Systems*, 2010.

[53] Christopher J. Hillar and Friedrich T. Sommer. Ramsey theory reveals the conditions when sparse coding on subsampled data is unique. *Arxiv preprint arxiv.org/abs/1106.3616*, 2011.

[54] S.M. Kim, S. Ganguli, and L.M. Frank. Spatial information outflow from the hippocampal circuit: Distributed spatial coding and phase precession in the subiculum. *The Journal of Neuroscience*, 32(34):11539–11558, 2012.

[55] D.L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914, 2009.

[56] T. Hu, Druckmann S., and D.B. Chklovskii. Early sensory processing as predictive coding: subtracting sparse approximations by circuit dynamics. *Front. Neurosci. Conf. Abs: COSYNE*, 2011.

[57] A.A. Koulakov and D. Rinberg. Sparse incomplete representations: A novel role for olfactory granule cells. *Neuron*, 72(1):124–136, 2011.

[58] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[59] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[60] J. Shlens, G.D. Field, J.L. Gauthier, M.I. Grivich, D. Petrusca, A. Sher, A.M. Litke, and EJ Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *The Journal of neuroscience*, 26(32):8254–8266, 2006.

[61] F. Krzakała, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.

[62] S. Kirkpatrick and D. Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384, 1978.

[63] JRL De Almeida and DJ Thouless. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 2001.

[64] R. Rammal, G. Toulouse, and M.A. Virasoro. Ultrametricity for physicists. *Reviews of Modern Physics*, 58(3):765, 1986.

[65] D.A. Huse and D.S. Fisher. Pure states in spin glasses. *Journal of Physics A: Mathematical and General*, 20(15):L997, 1999.

[66] AJ Bray and MA Moore. Chaotic nature of the spin-glass phase. *Physical review letters*, 58(1):57–60, 1987.

[67] S. Chatterjee. Disorder chaos and multiple valleys in spin glasses. *arXiv preprint arXiv:0907.3381*, 2009.

[68] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *PNAS*, 79(8):2554, 1982.

[69] D.O. Hebb. *The organization of behavior*. Wiley, New York, 1949.

[70] DJ Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin glass model of neural networks. *Phys. Rev. Lett*, 55:1530, 1985.

[71] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, 1987.

[72] M. Shamir and H. Sompolinsky. Thouless-anderson-palmer equations for neural networks. *Physical Review E*, 61(2):1839, 2000.

[73] H.A. Bethe. *Proc. R. Soc. London, Ser. A*, 151:552, 1935.

[74] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[75] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[76] M. Mézard and G. Parisi. The bethe lattice spin glass revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20(2):217–233, 2001.

[77] M. Mézard, G. Parisi, and R. Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[78] A. Braunstein, M. Mézard, and R. Zecchina. Survey propagation: An algorithm for satisfiability. *Random Structures & Algorithms*, 27(2):201–226, 2005.

[79] E. Gardner. The space of interactions in neural network models. *J. of Physics A*, 21:257–270, 1988.

[80] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, 1999.

[81] N. Brunel, V. Hakim, P. Isope, J.P. Nadal, and B. Barbour. Optimal information storage and the distribution of synaptic weights:: Perceptron versus purkinje cell. *Neuron*, 43(5):745–757, 2004.

[82] D. Marr. A theory of cerebellar cortex. *The journal of physiology*, 202(2):437, 1969.

[83] J.S. Albus. A theory of cerebellar function. *Math. Biosci.*, 10:26–51, 1971.

[84] P. Isope and B. Barbour. Properties of unitary granule cell purkinje cell synapses in adult rat cerebellar slices. *The Journal of neuroscience*, 22(22):9668–9678, 2002.

[85] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181, 1999.

[86] M. Griniasty. Cavity-approach analysis of the neural-network learning problem. *Physical Review E*, 47(6):4496, 1993.

[87] E. Lootens and C. van den Broeck. Analysing cluster formation by replica method. *EPL (Europhysics Letters)*, 30(7):381, 2007.

[88] C.C.H. Petersen, R.C. Malenka, R.A. Nicoll, and J.J. Hopfield. All-or-none potentiation at ca3-ca1 synapses. *PNAS*, 95(8):4732, 1998.

[89] D.H. O'Connor, G.M. Wittenberg, and S.S.H. Wang. Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *PNAS*, 102(27):9679, 2005.

[90] W. Krauth and M. Opper. Critical storage capacity of the j=±1 neural network. *Journal of Physics A: Mathematical and General*, 22:L519, 1989.

[91] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50(20):3057–3066, 1989.

[92] J.M. Montgomery and D.V. Madison. Discrete synaptic states define a major mechanism of synapse plasticity. *Trends in neurosciences*, 27(12):744–750, 2004.

[93] G.S. Dhesi and RC Jones. Asymptotic corrections to the wigner semicircular eigenvalue spectrum of a large real symmetric random matrix using the replica method. *Journal of Physics A: Mathematical and General*, 23(23):5577, 1999.

[94] HJ Sommers, A. Crisanti, H. Sompolinsky, and Y. Stein. Spectrum of large random asymmetric matrices. *Physical review letters*, 60(19):1895–1898, 1988.

[95] K. Rajan and LF Abbott. Eigenvalue spectra of random matrices for neural networks. *Physical review letters*, 97(18):188104, 2006.

[96] AM Sengupta and P.P. Mitra. Distributions of singular values for some random matrices. *Physical Review E*, 60(3):3389, 1999.

[97] D.C. Hoyle and M. Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69(2):026124, 2004.

[98] E.P. Wigner. On the distribution of the roots of certain symmetric matrices. *The Annals of Mathematics*, 67(2):325–327, 1958.

[99] G.S. Dhesi and RC Jones. Asymptotic corrections to the wigner semicircular eigenvalue spectrum of a large real symmetric random matrix using the replica method. *Journal of Physics A: Mathematical and General*, 23:5577, 1990.

[100] A. Edelman and N.R. Rao. Random matrix theory. *Acta Numerica*, 14(1):233–297, 2005.

[101] W.B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1–1, 1984.

[102] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[103] S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[104] R.G. Baraniuk and M.B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

[105] R.G. Baraniuk, V. Cevher, and M.B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, 2010.

[106] H.L. Yap, M.B. Wakin, and C.J. Rozell. Stable manifold embeddings with operators satisfying the restricted isometry property. In *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pages 1–6. IEEE, 2011.

[107] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[108] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51:4203–4215, 2005.

[109] S. Zhou, J. Lafferty, and L. Wasserman. Compressed and privacy-sensitive sparse regression. *Information Theory, IEEE Transactions on*, 55(2):846–866, 2009.

[110] M.F. Duarte, M.A. Davenport, M.B. Wakin, and R.G. Baraniuk. Sparse signal detection from incoherent projections. In *Acoustics, Speech and Signal Processing, ICASSP Proceedings.*, volume 3, pages III–III. IEEE, 2006.

[111] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk. The smashed filter for compressive classification and target recognition. *Proc. Computational Imaging V at SPIE Electronic Imaging*, 2007.

[112] M.F. Duarte, M.A. Davenport, M.B. Wakin, J.N. Laska, D. Takhar, K.F. Kelly, and R.G. Baraniuk. Multiscale random projections for compressive classification. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 6, pages VI–161. IEEE, 2007.

[113] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*, pages 1430–1434. IEEE, 2006.

[114] A. Blum. Random projection, margins, kernels, and feature-selection. *Subspace, Latent Structure and Feature Selection*, pages 52–68, 2006.

[115] C. Hegde, M.B. Wakin, and R.G. Baraniuk. Random projections for manifold learning. In *Neural Information Processing Systems*. Citeseer, 2007.

[116] S. Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

[117] D.L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via l1 minimization. *PNAS*, 100:2197–2202, 2003.

[118] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[119] Y. Kabashima, T. Wadayama, and T. Tanaka. A typical reconstruction limit for compressed sensing based on l p-norm minimization. *J. Stat. Mech.*, L09003, 2009.

[120] S. Rangan, A.K. Fletcher, and Goyal V.K. Asymptotic analysis of map estimation via the replica method and applications to compressed sensing. *CoRR*, abs/0906.3234, 2009.

[121] S. Ganguli and H. Sompolinsky. Statistical mechanics of compressed sensing. *Phys. Rev. Lett.*, 104(18):188701, 2010.

[122] D.L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *PNAS*, 102:9452–7, 2005.

[123] D.L. Donoho and J. Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *PNAS*, 102:9446–51, 2005.

[124] A. Montanari. Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*, pages 394–438, 2010.

[125] P.C. Martin, ED Siggia, and HA Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973.

[126] C. De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

[127] A. Crisanti and H. Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Physical Review A*, 36(10):4922, 1987.

[128] A. Crisanti and H. Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Ising spins and glauber dynamics. *Physical Review A*, 37(12):4865, 1988.

[129] B. Derrida, E. Gardner, and A. Zippelius. An exactly solvable asymmetric neural network model. *EPL (Europhysics Letters)*, 4(2):167, 2007.

[130] C. van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274:1724–1726, 1996.

[131] C. Vreeswijk and H. Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural Computation*, 10(6):1321–1371, 1998.

[132] L. Molgedey, J. Schuchhardt, and HG Schuster. Suppressing chaos in neural networks by noise. *Physical review letters*, 69(26):3717–3719, 1992.

[133] K. Rajan, LF Abbott, and H. Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 82(1):011903, 2010.

[134] N. Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of computational neuroscience*, 8(3):183–208, 2000.

[135] A. Renart, N. Brunel, and X.J. Wang. *Mean-field theory of irregularly spiking neuronal populations and working memory in recurrent cortical networks*. Boca Raton, CRC Press, 2004.

[136] J. Hertz, A. Lerchner, and M. Ahmadi. Mean field methods for cortical network dynamics. *Computational neuroscience: Cortical dynamics*, pages 71–89, 2004.

[137] M. Monteforte and F. Wolf. Dynamical entropy production in spiking neuron networks in the balanced state. *Physical review letters*, 105(26):268104, 2010.

[138] M. Monteforte and F. Wolf. Dynamic flux tubes form reservoirs of stability in neuronal circuits. *Physical Review X*, 2(4):041007, 2012.

[139] M. London, A. Roth, L. Beeren, M. Häusser, and P.E. Latham. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature*, 466(7302):123–127, 2010.

[140] T.P. Vogels, K. Rajan, and LF Abbott. Neural network dynamics. *Annu. Rev. Neurosci.*, 28:357–376, 2005.

[141] M.I. Rabinovich, P. Varona, A.I. Selverston, and H.D.I. Abarbanel. Dynamical principles in

neuroscience. *Reviews of modern physics*, 78(4):1213, 2006.

[142] HS Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056, 1992.

[143] H. Sompolinsky, N. Tishby, and HS Seung. Learning from examples in large neural networks. *Physical Review Letters*, 65(13):1683–1686, 1990.

[144] E. Barkai, D. Hansel, and I. Kanter. Statistical mechanics of a multilayered neural network. *Physical review letters*, 65(18):2312–2315, 1990.

[145] E. Barkai, D. Hansel, and H. Sompolinsky. Broken symmetries in multilayered perceptrons. *Physical Review A*, 45(6):4146, 1992.

[146] A. Engel, HM Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius. Storage capacity and learning algorithms for two-layer neural networks. *Physical Review A*, 45(10):7590, 1992.

[147] M. Opper. Learning and generalization in a two-layer neural network: The role of the vapnik-chervonvenkis dimension. *Physical review letters*, 72(13):2113–2116, 1994.

[148] R. Monasson and R. Zecchina. Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks. *Physical review letters*, 75(12):2432–2435, 1995.

[149] H. Schwarze. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21):5781, 1999.

[150] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14):2975–2978, 1999.

[151] M. Opper and R. Urbanczik. Universal learning curves of support vector machines. *Physical Review Letters*, 86(19):4410–4413, 2001.

[152] D. Malzahn and M. Opper. A statistical physics approach for the analysis of machine learning algorithms on real data. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11001, 2005.

[153] M. Urry and P. Sollich. Replica theory for learning curves for gaussian processes on random graphs. *arXiv preprint arXiv:1202.5918*, 2012.

[154] R. Gütig and H. Sompolinsky. The tempotron: a neuron that learns spike timing–based decisions. *Nature neuroscience*, 9(3):420–428, 2006.

[155] R. Rubin, R. Monasson, and H. Sompolinsky. Theory of spike timing-based neural classifiers. *Physical review letters*, 105(21):218102, 2010.

[156] J.W. Lichtman and J.R. Sanes. Ome sweet ome: what can the genome tell us about the connectome? *Current opinion in neurobiology*, 18(3):346–353, 2008.

[157] T.L.H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics;(United States)*, 65(2), 1993.

[158] D.C. Hoyle. Statistical mechanics of learning orthogonal signals for general covariance models. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(04):P04009, 2010.

[159] M. Biehl and A. Mietzner. Statistical mechanics of unsupervised structure recognition. *Journal of Physics A: Mathematical and General*, 27(6):1885, 1999.

[160] M. Biehl. An exactly solvable model of unsupervised learning. *EPL (Europhysics Letters)*, 25(5):391, 1994.

[161] C. Marangi, M. Biehl, and S.A. Solla. Supervised learning from clustered input examples. *EPL (Europhysics Letters)*, 30(2):117, 2007.

[162] K. Rose, E. Gurewitz, and G.C. Fox. Statistical mechanics and phase transitions in clustering. *Physical review letters*, 65(8):945–948, 1990.

[163] N. Barkai, HS Seung, and H. Sompolinsky. Scaling laws in learning of classification tasks. *Physical review letters*, 70(20):3167–3170, 1993.

[164] N. Barkai and H. Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Physical Review E*, 50(3):1766, 1994.

[165] M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Physical review letters*, 76(18):3251–3254, 1996.

[166] S. Wiseman, M. Blatt, and E. Domany. Superparamagnetic clustering of data. *Physical Review E*, 57(4):3767, 1998.

[167] M. Łuksza, M. Lässig, and J. Berg. Significance analysis and statistical mechanics: an application to clustering. *Physical review letters*, 105(22):220601, 2010.

[168] F.J. Dyson. A brownian-motion model for the eigenvalues of a random matrix. *Journal of Mathematical Physics*, 3(6), 1962.

[169] G. Schehr, S.N. Majumdar, A. Comtet, and J. Randon-Furling. Exact distribution of the maximal height of p vicious walkers. *Physical review letters*, 101(15):150601, 2008.

[170] C.A. Tracy and H. Widom. Nonintersecting brownian excursions. *The Annals of Applied Probability*, 17(3):953–979, 2007.