

Accepted Manuscript

Statistical physics and representations in real and artificial neural networks

S. Cocco, R. Monasson, L. Posani, S. Rosay, J. Tubiana

PII: S0378-4371(17)31234-7
DOI: <https://doi.org/10.1016/j.physa.2017.11.153>
Reference: PHYSYA 18921

To appear in: *Physica A*



Please cite this article as: S. Cocco, R. Monasson, L. Posani, S. Rosay, J. Tubiana, Statistical physics and representations in real and artificial neural networks, *Physica A* (2017), <https://doi.org/10.1016/j.physa.2017.11.153>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

- Some recent results on high-dimensional representations in real and artificial neural networks are presented.
- Statistical-physics models for continuous attractor-based representations of space in the mammalian brain are proposed, with their phase diagram, and their properties are related to recent in vivo recordings in rodents.
- Some algorithms important for unsupervised machine learning are presented, in particular for dimensional reduction and feature extraction.

Statistical Physics and Representations in Real and Artificial Neural Networks

S. Cocco¹, R. Monasson², L. Posani¹, S. Rosay³, J. Tubiana²

¹ Laboratoire de Physique Statistique, Ecole Normale Supérieure and CNRS, PSL
Research, Sorbonne Universités UPMC, 24 rue Lhomond, 75005 Paris, France

² Laboratoire de Physique Théorique, Ecole Normale Supérieure and CNRS, PSL Research,
Sorbonne Universités UPMC, 24 rue Lhomond, 75005 Paris, France

³ Cognitive Neuroscience, SISSA, via Bonomea 265, Trieste, Italy

November 21, 2017

Abstract

This document presents the material of two lectures on statistical physics and neural representations, delivered by one of us (R.M.) at the Fundamental Problems in Statistical Physics XIV summer school in July 2017. In a first part, we consider the neural representations of space (maps) in the hippocampus. We introduce an extension of the Hopfield model, able to store multiple spatial maps as continuous, finite-dimensional attractors. The phase diagram and dynamical properties of the model are analyzed. We then show how spatial representations can be dynamically decoded using an effective Ising model capturing the correlation structure in the neural data, and compare applications to data obtained from hippocampal multi-electrode recordings and by (sub)sampling our attractor model. In a second part, we focus on the problem of learning data representations in machine learning, in particular with artificial neural networks. We start by introducing data representations through some illustrations. We then analyze two important algorithms, Principal Component Analysis and Restricted Boltzmann Machines, with tools from statistical physics.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 2 |
| 2 | Representation of space(s) in the hippocampus: model | 3 |
| 2.1 | Context and background | 3 |
| 2.1.1 | Zero-dimensional attractors: Hopfield model of associative memory | 3 |
| 2.1.2 | Place cells in the rodent hippocampus | 3 |
| 2.2 | A model for memorizing D-dimensional attractors (spatial maps) | 4 |
| 2.3 | Replica theory and phase diagram | 5 |
| 2.4 | Dynamics within one map and transitions between maps | 7 |
| 3 | Representation of space(s) in the hippocampus: decoding and data analysis | 8 |
| 3.1 | Decoding neural representations with effective Ising networks | 8 |
| 3.2 | Inference of effective Ising models from data | 11 |
| 3.3 | Back to model: the subsampling problem | 12 |
| 3.3.1 | Simulations: constructing reference and test sessions | 12 |
| 3.3.2 | Decoding Results | 13 |
| 3.3.3 | Inferred vs. true couplings | 13 |
| 3.4 | Analysis of multi-electrode recordings in CA1 | 14 |
| 4 | Representations of Data in Machine Learning | 17 |
| 4.1 | Introduction | 17 |
| 4.2 | Example 1: Dimensionality reduction | 17 |
| 4.3 | Example 2: Extracting latent features from data | 18 |

| | | |
|----------|--|-----------|
| 5 | Low-Dimensional Representations: Principal Component Analysis (PCA) | 19 |
| 5.1 | Mathematical reminder: Bayesian inference | 19 |
| 5.2 | Multivariate Gaussian variables | 20 |
| 5.3 | Principal components as minimal models of interacting variables | 21 |
| 5.4 | The retarded-learning phase transition | 22 |
| 5.5 | Incorporating prior information | 23 |
| 5.6 | Inverse Hopfield problem | 24 |
| 6 | Compositional Representations: Restricted Boltzmann Machines (RBM) | 26 |
| 6.1 | Definition and motivation | 26 |
| 6.2 | Sampling | 27 |
| 6.3 | Phenomenology of RBM trained on data | 27 |
| 6.4 | Statistical mechanics of RBM | 28 |
| 6.5 | Validation on data | 30 |

1 Introduction

In the early 80's, statistical physicists proved that ideas issued from their field could lead to substantial advances in other disciplines. Simulated Annealing, a versatile optimization procedure in which a fictitious sampling temperature is decreased until the minimum (ground state) of a cost function is reached, had major impact in applied computer science and engineering [1]. Attractor neural network models for memories [2], soon analytically solved with spin-glass techniques [3], emerged as one major conceptual tool in computational neuroscience. From a theoretical point of view, it became rapidly clear that statistical physics offered a powerful framework to deal with problems outside physics, in particular in computer science and theoretical neuroscience, involving many random, heterogeneous, strongly interacting components, which had remained very hard to tackle so far.

The purpose of the present document is to present some applications of statistical physics ideas and tools to the understanding of high-dimensional representations in neural networks. How the brain represents and processes information coming from the outside world is a central issue of computational neuroscience [4]. Experimental progress in electrophysiological and optical recordings make now possible to record the activity of populations of tens to thousands of neural cells in behaving animals, opening the way to study this question with unprecedented access to data and to ask new questions about brain operation on large scales [5]. Concomitantly, machine learning algorithms, largely based on artificial neural network architectures, have recently achieved spectacular performance in a variety of fields, such as image processing, or speech recognition/production [6]. How these machines produce efficient representations of the data and of their underlying distributions is a crucial question [7], far from being understood [8]. Profound similarities seem to emerge between the representations encountered in real and artificial neural networks [9] and between the questions raised in both contexts [10].

It is utterly hard to cover recent advances in such a diverse and vivid field, and the task is impossible in two lectures of two hours each. The material gathered here merely reflects the interests and, presumably, the ignorance of the authors more than anything else. The present notes focus on two applications of statistical physics to the study of neural representations in the contexts of computational neuroscience and machine learning. The first part is motivated by the representation of spaces, i.e. multiple environments, in hippocampal place-cell networks. An extension of Hopfield's attractor neural network to the case of finite-dimensional attractors is introduced and its phase diagram and dynamical properties, such as diffusion within one attractor or transitions between distinct attractors, are analyzed. We also show that effective, functional Ising models fitted from hippocampal multi-electrode recordings (limited to date to few tens of neurons) or from 'neural' data generated by spatially subsampling our model, share common features with our abstract model, and can be used to decode and to track the evolution of spatial representations over time. In a second part, we move to representations of data by machine learning algorithms. Special emphasis is put on two aspects: low-dimensional representations achieved by principal component analysis, and compositional representations, produced by restricted Boltzmann machines combining multiple features inferred from data. In both cases, we show how statistical physics helps unveil the different properties of these representations, and the role of essential control parameters.

2 Representation of space(s) in the hippocampus: model

2.1 Context and background

2.1.1 Zero-dimensional attractors: Hopfield model of associative memory

Statistical Mechanics and Neuroscience are not so far apart as they may seem at first sight. Indeed, brains are made of billions of neurons that are connected together. In many cases, brain functions are thought to be the outcome of collective states. This makes it a good playground for Statistical Mechanics. Here, we will focus on one particular brain function: memory.

In 1949, D. Hebb had the visionary intuition that memory could correspond to the retrieval of certain activity patterns in a network of interconnected neurons [11]. This *attractor hypothesis* goes as follows: (1) what is memorized are attractors of the network, *i.e.* activity states stable under the dynamical evolution rule; hence, recalling a memory corresponds to retrieving its activity pattern; (2) attractors are stored in the network couplings J_{ij} that govern the network dynamics and stable states; (3) a possible way to make an arbitrary pattern an attractor is to 'wire together neurons that fire together' in that pattern (the so-called 'Hebb rule').

In 1982, J.J. Hopfield [2] proposed a model based on Hebb's ideas in the case of zero-dimensional, or, equivalently, point attractors. This model, known as the Hopfield model, is strongly inspired by statistical physics models used in the context of the magnetic systems, such as the Ising model. It consists of a number N of binary neurons $\{s_i\}_{i=1\dots N} = \pm 1$ and stores a number P of configurations $\{\xi_i^\mu\}_{i=1\dots N, \mu=1\dots P} = \pm 1$ (point attractors), *e.g.* independently and uniformly drawn at random. The synaptic couplings that allow these configurations to be attractors are given by the Hebb rule:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \quad \forall i, j. \quad (1)$$

The last thing to define is the dynamics of the network. In the original paper [2], time was discretized and, at each time step t , neurons responded deterministically to their local fields, through the updating rule:

$$s_i^{t+1} = \text{sign}\left(\sum_j J_{ij} s_j^t\right). \quad (2)$$

Later studies, *e.g.* [12], incorporated stochasticity in the response, through a noise parameter T , so that the system obeyed detailed balance for the Gibbs distribution associated to the Hamiltonian

$$E_J[\mathbf{s}] = - \sum_{i < j} J_{ij} s_i s_j \quad (3)$$

at 'temperature' T .

In terms of biological relevance, the Hopfield model is of course extremely schematic. Yet, it captures many fundamental and robust aspects of neurons (in particular their linear summation of inputs, combined with a thresholding effect) and network (synaptic coefficients with values affected by the activity through the Hebb rule), while remaining, to a large extent, analytically tractable. The Hopfield model aroused a great excitement in the Statistical Mechanics community during the 80's, since it shared many common points with frustrated and disordered magnetic systems. Tools and methods from the statistical physics of disordered systems that had just been developed in the field of spin glasses [13] could therefore be used to derive analytically the properties of the Hopfield model [12]. The first question was to check whether the patterns $\{\xi_i^\mu\}_{i=1\dots N, \mu=1\dots P}$ were indeed attractive fixed points of the dynamics in eqn (2). The answer turned out to be positive (up to a small fraction of the N neurons) for small enough values of T and of the ratio $\alpha = P/N$ (in the double limit $N, P \rightarrow \infty$), *i.e.* for not too strong noise and memory load. Many aspects of this model were studied and refined, in particular to make it more biologically realistic. The reader is kindly referred to [14] for a detailed presentation of the literature. Rather, we will focus in an extension of this model to a different kind of attractors, that is finite-dimensional attractors.

2.1.2 Place cells in the rodent hippocampus

Let us now turn to real brains, more specifically, how space is represented in the brain [15]. Experimentalists use small electrodes that, implanted in the brain of awake animals, are able to record the simultaneous activity of a population of *single neurons*. In particular, in a brain area called hippocampus, O'Keefe & Dostrovsky have discovered the existence of 'place cells' when recording in rodents freely moving in an enclosure [16]. These neurons have the surprising property that they fire only when the animal is physically located in a precise region of space, hence their name. The region of activity corresponding to a place cell in

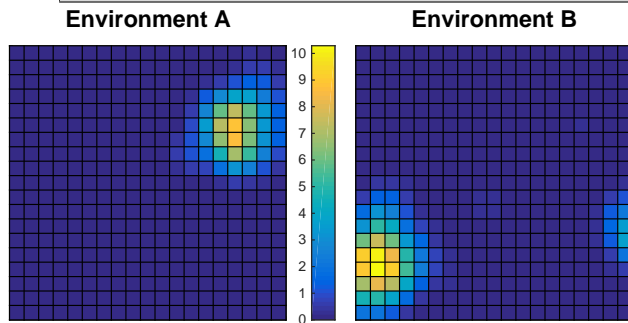


Figure 1: Remapping of place field for one recorded place cell in CA1 for a living rat exploring two square environments A and B with identical shapes. The size of the environments is 60×60 cm. The figure reports the average firing rate of the recorded cell when the rat is in each of the 3×3 -cm spatial bins; values in Hz, see color bar. Data from experiment by K. Jezek et al [17].

the environment defines its 'place field'. In the CA3 area of the hippocampus, a region with strong recurrent connection between pyramidal cells, the different place fields attached to a given place cell across different environments visited by the rodent seem to be totally uncorrelated — a property called global 'remapping'. In another hippocampal area, called CA1, remapping of place fields from one environment to another is generally weaker; the change in the activity of a place cell is characterized mainly by a modulation of its firing rate, a phenomenon called rate remapping, though global changes of the place fields as in CA3 may also be observed for some cells, see Fig. 1.

For many reasons, the hippocampus — more precisely its subregion CA3 — is often supposed to work as a *continuous* attractor neural network [18, 19]. It means that the attractors are not point configurations (of zero dimension, as in the Hopfield model), but attractors in one¹ or two dimensions²: each manifold corresponds to an environment, *i.e.* the collection of activity configurations of the hippocampal neural population characterizing the set of all positions in that environment. Apart from the dimensionality of the attractors, place cells share common points with the Hopfield model, such as the absence of correlation between attractors due to random remapping, and the Hebb rule that has some biological counterparts. Hence, it is appealing to extend the Hopfield model to continuous attractors.

2.2 A model for memorizing D-dimensional attractors (spatial maps)

We thus introduce a model for place cells in one- or two-dimensional spaces (the extension to higher dimensions is straightforward). As an extension of the Hopfield model, our model is based on binary neurons; other models with real-valued neural variable, *e.g.* firing rates, can be found in literature [21, 22]. The N place cells are modeled by binary units s_i equal to 0 (silent state) or 1 (active state)³. These neurons interact together through excitatory couplings J_{ij} . Moreover, they interact with inhibitory interneurons, whose effect is to maintain the total activity of the place cells to a fraction f of active cells (global inhibition). We also assume that there is some stochasticity in the response of the neurons, controlled by a noise parameter T . All these assumptions come down to considering that the network states are distributed according to the Gibbs distribution associated to the Hamiltonian (3), restricted to configuration of spins \mathbf{s} such that

$$\sum_i s_i = fN . \quad (4)$$

We want to store $L + 1$ environments in the coupling matrix. We call place field a position of space where a place cell preferentially fires. An environment ℓ is defined as a random permutation π^ℓ of the N neurons' place fields (assuming that the place fields are regularly arranged on a grid). This models the experimentally observed remapping of place fields from one map to the other⁴. With this definition, an environment is said to be stored when activity patterns localized in this environment are stable states of the dynamics. In other words, the configurations where active neurons have neighbouring place fields in this environment are equilibrium states. To make this possible, we assume a Hebbian prescription for the couplings J_{ij} that is a straightforward extension of the Hopfield synaptic matrix to the case of quasi-continuous attractors. This rule is illustrated in Figure 2, and is mathematically described as follows:

- additivity: $J_{ij} = \sum_{\ell=0}^L J_{ij}^\ell$ where the sum runs over all the environments.

¹The one-dimensional case would correspond to linear corridors.

²Though there is experimental evidence that place cells code also for rich contextual information [20], we consider only the spatial correlate of place-cell activity in the present document.

³We will hereafter use indifferently the terms "neuron", "place cell" and "spin", from the analogy with magnetic systems.

⁴In this basic version of the model, every place cell has a place field in every environment. The possibility of silent cells has been taken into account in [23]

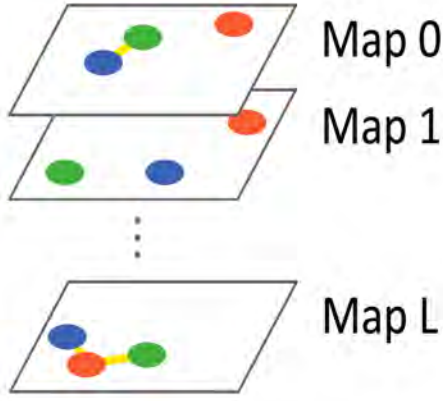


Figure 2: Remapping and connectivity rule in the model, illustrated with three units and $L + 1$ two-dimensional environments. The place field centers of the units are displayed respectively in red, blue and green. Thick yellow lines indicate the excitatory couplings between cells with neighbouring place fields in each environment. These place fields overlap; here, for the sake of clarity, only the centers of the place fields are represented.

- potentiation of excitatory couplings between units that may become active together when the animal explores the environment:

$$J_{ij}^\ell = \frac{1}{N} \text{ if } d_{ij}^\ell \leq d_c, \quad 0 \text{ if } d_{ij}^\ell > d_c, \quad (5)$$

where d_{ij}^ℓ is the distance between the place-field centers of i and j in the environment ℓ ; for instance, in dimension $D = 1$, $d_{ij}^\ell = \frac{1}{N} |\pi^\ell(i) - \pi^\ell(j)|$. d_c represents the distance over which place fields overlap. In practice, it is chosen so that, in each environment, each neural cell is coupled to a fraction w of the other cells (its neighbours); in dimension $D = 1$ again, we may choose $d_c = \frac{w}{2}$. The $\frac{1}{N}$ factor in eqn (5) ensures that the total input received by a cell remains finite as N goes to infinity, a limit case in which exact calculations become possible [24].

2.3 Replica theory and phase diagram

The aim of this calculation is to study the stable states of the network, and to find under which conditions these stable states correspond to a set of active neurons whose corresponding place fields are nearby in one of the environments. In other words, we want to know for which parameter values the Hebbian synaptic matrix (5) ensures the retrieval of the stored maps. The system under study enjoys both disordered (due to the random allocation of place fields in each map) and frustrated (from the competition between excitatory synapses and the global inhibition) interactions.

We start by computing the free energy of the system,

$$F = -T \log Z_J(T), \quad \text{where} \quad Z_J(T) = \sum_{\mathbf{s} \text{ with constraint (4)}} \exp(-E_J(\mathbf{s})/T). \quad (6)$$

This quantity depends a priori on the realization of the random permutations in each map. We assume that, in the large N limit, the free energy is self-averaging: its particular value for a given realization of the disorder is typically close to its average over all possible realizations of the disorder, which is thus a good approximation of F . The randomness of the remapping process is thus a key hypothesis for the model to be tractable. To compute the average of the logarithm of $Z_J(T)$ we use the replica method [13]: we first compute the n^{th} moment of $Z_J(T)$, and then compute its first derivative with respect to $n \rightarrow 0$.

Since we are interested in configurations where the place fields of the active neurons are spatially concentrated in one of the environments, we arbitrarily select one of the environments (called “reference environment”) and do the averaging over the remaining L other permutations; details about the calculation can be found in [23]. This choice is totally arbitrary because the difference between environments is eventually averaged out. In the reference environment, neurons are indexed in the same order as their place fields, which allows us to move from a microscopic activity configuration \mathbf{s} to a macroscopic activity density over continuous space

$$\rho(x) \equiv \lim_{\epsilon \rightarrow 0} \lim_{N \rightarrow \infty} \frac{1}{\epsilon N} \sum_{(x - \frac{\epsilon}{2})N \leq i < (x + \frac{\epsilon}{2})N} \overline{\langle s_i \rangle_J}, \quad (7)$$

where the overbar denotes the average over the random remappings while the brackets correspond to the average over the fast noise. For simplicity, we have assumed that the environment is one-dimensional here, but the above formula can easily be extended to higher dimensions. Note that our model is analytically tractable in the large N limit (thermodynamic limit), as each unit has an infinite number of neighbours it weakly interacts with. The mean-field approximation therefore becomes exact in the sense that order

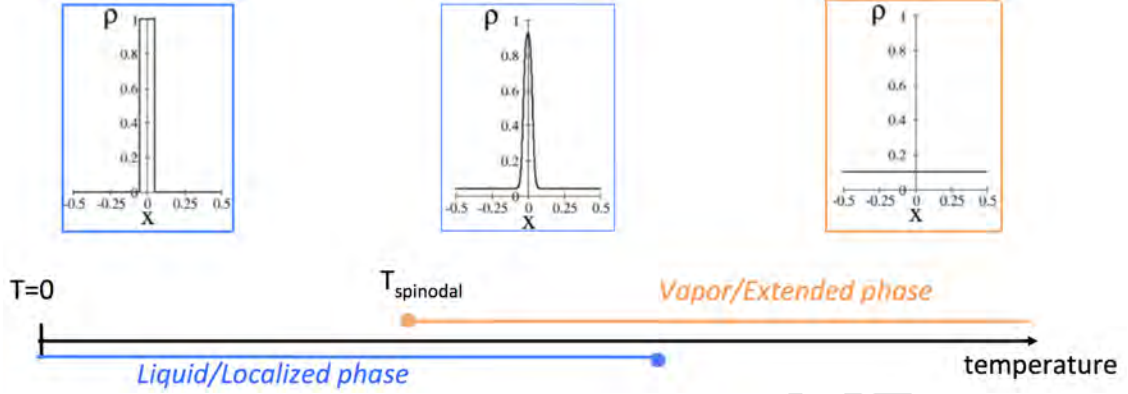


Figure 3: Phase diagram of Lebowitz & Penrose's theory of the liquid/vapor transition, in dimension $D = 1$ (periodic boundary conditions). Insets show the density of particles $\rho(x)$ as a function of position over space, $x \in [0; 1]$. Parameters: $f = 0.1$, $w = 0.05$. Note the coexistence between the homogeneous and bump states at intermediate temperatures. The location of the bump is arbitrary.

parameters such as the density in (7) exhibits no fluctuation when $N \rightarrow \infty$; however, contrary to standard mean field approaches, these order parameters depend on space [24].

The case of a single (reference) environment, i.e. $L = 0$, is strongly reminiscent of the theory of the liquid-vapor transition by Lebowitz and Penrose [24]: the continuous translational symmetry is spontaneously broken at low enough temperature, i.e. $\rho(x) \neq f$, and a liquid drop (bump of high density fluid) is surrounded by low-density vapor, see Fig. 3. this bump can then freely diffuse, and describes a finite-dimensional continuum of ground states.

In the $L > 0$ case, the n^{th} moment of $Z_J(T)$ is given by equation (20) in [23]. The averaged term depends on the configurations $\mathbf{s}^1, \dots, \mathbf{s}^n$ only through the overlap matrix with entries $q^{ab} = \frac{1}{N} \mathbf{s}^a \cdot \mathbf{s}^b$ [13, 23]. Then, to perform the $n \rightarrow 0$ limit, we consider the replica symmetric Ansatz, which assumes that the overlaps q^{ab} take a single value (over all replica indices $a \neq b$),

$$q \equiv \frac{1}{N} \sum_j \overline{(s_i^j)^2}, \quad (8)$$

which is the Edwards-Anderson parameter of spin glasses, characterizing the site-to-site fluctuations of the spin magnetizations [25]. This Ansatz is generally valid at high-enough temperature, when the Gibbs measure defined on the energy landscape is not too rough, see below. It allows us to compute the free energy as a function of the order parameters $\rho(x)$, $\mu(x)$ (chemical potential conjugated to $\rho(x)$), q and r (conjugated to q). $\mu(x)$ and r have simple interpretations. The effective field acting on neuron i , whose place field is located in $x = i/N$ in the reference environment, is the sum of two terms: a 'signal' contribution $\mu(x)$ coming from neighboring neurons j in the retrieved map (through the couplings J_{ij}^0), and a Gaussian noise, of zero mean and variance αr coming from the other maps $\ell \geq 1$ (see Fig. 14 in [23]). Here, $\alpha \equiv L/N$ denotes the load of the memory.

The four order parameters (two scalar, two functional) fulfill the following saddle-point equations obtained through extremization of the free-energy functional,

$$\begin{aligned} r &= 2(q - f^2) \sum_{k \geq 1} \left[\frac{k\pi}{\sin(k\pi w)} - \beta(f - q) \right]^{-2}, \quad q = \int dx \int Dz [1 + e^{-\beta z \sqrt{\alpha r} - \beta \mu(x)}]^{-2}, \\ \rho(x) &= \int Dz [1 + e^{-\beta z \sqrt{\alpha r} - \beta \mu(x)}]^{-1}, \quad \mu(x) = \int dy J_w(x - y) \rho(y) + \lambda, \end{aligned} \quad (9)$$

where $\beta \equiv 1/T$, $Dz = \exp(-z^2/2)/\sqrt{2\pi}$ is the Gaussian measure, and λ is determined to enforce the fixed activity level constraint $\int dx \rho(x) = f$, see (4). The precise expression of r depends on the eigenvalue spectrum of the J_{ij}^0 matrix. Changing the hard cut-off d_c to a smooth e.g. exponential decay of the coupling with the distance d_{ij}^ℓ between the place-field centers would change the expression for r , but would not affect the overall behaviour of the model.

We find three distinct solutions to these coupled equations:

- a paramagnetic phase (PM), corresponding to high levels of noise T , in which the average local activity is uniform over space, $\rho(x) = f$, and neurons are essentially uncorrelated, $q = f^2$.
- a 'clump' phase (CL), where the activity depends on space, i.e. $\rho(x)$ varies with x and is localized in the reference environment. This phase corresponds to the 'retrieval phase' where the environment

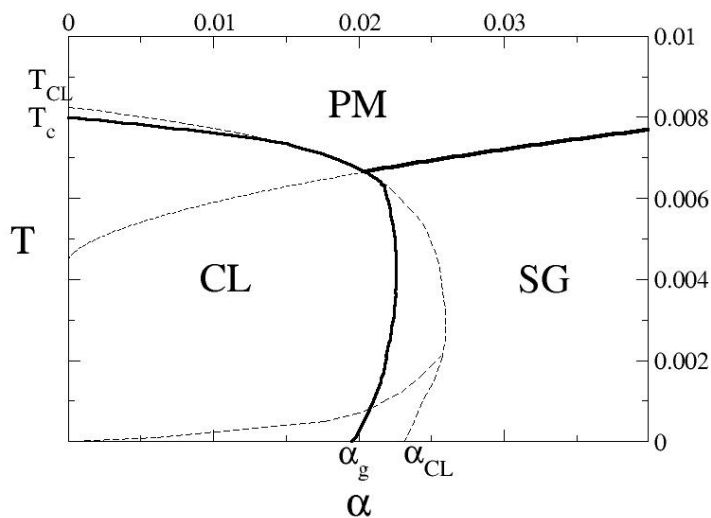


Figure 4: Phase diagram in the (α, T) plane in $D = 1$ with $f = 0.1$ and $w = 0.05$. Thick lines: transition between phases. Dashed-dotted line: $T_{\text{PM}}(\alpha)$. Thin dashed line: CL phase's longitudinal stability regions. Dotted line: CL phase's RSB line. α_{CL} : storage capacity at $T = 0$ of the replica-symmetric clump phase. α_g : CL-SG transition load at $T = 0$. T_{CL} : temperature of loss of stability of the clump at $\alpha = 0$. T_c : CL-PM transition temperature at $\alpha = 0$. $T_{\text{PM}} = T_{\text{PM}}(\alpha = 0)$ (see text).

is actually memorized. In fact, all the $L + 1$ environments are memorized since any of them could be chosen as the reference environment. Note that the value of x (center of the bump of activity) is totally arbitrary, as all positions are equivalent after averaging over the permutations.

- a glassy phase (SG), corresponding to large loads α , in which the local activity $\langle s_i \rangle$ varies from neuron to neuron ($q > f^2$), but does not cluster around any specific location in space in any of the environments ($\rho(x) = f$ after averaging over remappings). In this SG phase the crosstalk between environments is so large that none of them is actually stored in the network activity. In the SG phase, contrary to the CL phase, no environment is memorized. This is the 'black-out catastrophe' [14] already described in the Hopfield model, in which retrieval also takes place in an all-or-nothing fashion.

We now need to determine which solution is selected as functions of α , T , that is, the phase of lowest free energy that will be thermodynamically favored, as well as the domains of existence (stability) of those three phases against longitudinal and replicon modes [26], and the transition lines between them. To study the stability, we write the Hessian of the free energy and study its eigenvalues in the longitudinal and replicon sectors. Then, the transition between two phases is the line where the free energies in both phases equalize. We have done these calculations in the one-dimensional case, as detailed in ref. [23]. The outcome is the phase diagram shown in Fig. 4, displaying the three phases domains in the (α, T) plane:

- the paramagnetic solution exists for all α , T and is stable for $T > T_{\text{PM}}(\alpha)$ displayed with the dot-dashed line in Fig. 4.
- the glassy phase exists for $T < T_{\text{PM}}(\alpha)$, and is always replica-symmetry broken; we expect replica symmetry breaking to be continuous in this region, as in the celebrated Sherrington-Kirkpatrick model [13].
- the longitudinal stability of the clump phase is computed numerically and shown with the thin dashed line in Fig. 4. The clump is stable against replicon modes except in a little low- T high- α region (dotted line). An interesting feature of the clump phase stability domain is the reentrance of the high- α boundary.

We have checked this analytically-derived phase diagram by Monte Carlo simulations. For a detailed comparison of this phase diagram with the one of the Hopfield model, and how the dimensionality of the attractors plays a role, see [27].

2.4 Dynamics within one map and transitions between maps

The phase diagram above informs us on the stable states of the model. For moderate temperature and memory load, i.e. in the CL phase, thermodynamically stable states have an activity spatially localized somewhere in one of the maps. But this does not constrain *which* map is retrieved and *where* in this map, i.e. what is the position intersecting the place fields of the active cells. Indeed, under the influence of noise, the bump of activity can move around in a given map, and also jump to another map. We have studied both these dynamics within one map and between maps, respectively in [28] and [29].

Within one map, we have shown formally that, in the case of a single continuous attractor (one map, *i.e.* $\alpha = 0$), the bump of activity behaves like a quasi-particle with little deformation. This quasi-particle undergoes a pure diffusion with a diffusion coefficient that can be computed exactly from first principles, *i.e.* from the knowledge of microscopic flipping rates of spins in Monte Carlo simulations. The diffusion coefficient scales as $1/N$, see eqn (31) in [28]. When imposing a force on the spins, see Section 3.3.1, the activity changes so as to move the bump. An illustration is shown in Fig. 5. It can be shown analytically that the mobility of the bump and its diffusion coefficient obey the Stokes-Einstein relation.

In the presence of multiple maps, the disorder in the couplings due to the additive storage creates an effective free-energy landscape for the bump of activity in the reference environment. The free-energy barriers scale typically as \sqrt{N} , and are correlated over space lengths of the order of the bump size, see [28]. In one dimension, the bump therefore effectively undergoes Brownian motion in the Sinai potential, with strongly activated diffusion. In higher dimension, diffusion is facilitated with respect to the 1D case, as can be observed in simulations.

In addition to moving in the reference environment, the bump can also spontaneously jump between maps. Fast transitions between maps, evoked by light inputs, have been observed by K. Jezek and colleagues in the so-called 'teleportation experiment' [17]. Understanding these transitions in the framework of a simple model provides insight on the mechanisms involved in the biological system. Map-to-map transitions can be studied with replica theory again, but in a more subtle framework, where solutions with non-uniform activities in two maps (and not only one as in eqn. (9) above) are searched for. There are two scenarios for spontaneous transitions between spatial representations, see Fig. 4 in [29]:

- through a mixed state, which gives bumps of activity in both maps; these bumps are weaker than the one in the CL phase in a single reference environment. This scenario is preferred (has lower free-energy cost) at low T . Transitions take place at special 'confusing' positions in both environments, where both maps locally resemble most.
- through a non-localized state, *i.e.* through the PM phase. Owing to the liquid-vapor analogy, the bump of activity in map A evaporates, and then condensates in map B. This scenario is preferred (has lower free-energy cost) at high T (but sufficiently low to make the CL phase thermodynamically favorable with respect to PM, see phase diagram in Fig. 4).

We show in Fig. 6A the rate of transitions between maps computed from Monte Carlo simulations, see Supplemental Material in [29] for details. We observe that the rate increases with temperature, and diminishes with the load and the system size. According to Langer's nucleation theory [30], we expect the rate to be related to the free-energy barrier ΔF between CL phase in environment A and the CL phase in environment B through (see formula 3.37 in [30]),

$$R = \kappa \sqrt{\frac{T}{2\pi N |\lambda_-|}} \mathcal{V} \exp(-N \Delta F/T), \quad (10)$$

where κ is the growth rate of the unstable mode at the transition state, λ_- is the unique negative eigenvalue of the Hessian of the free-energy at the transition state, \mathcal{V} is the volume of the saddle-point subspace (resulting from the integral over continuous degrees of freedom leaving the saddle point configuration globally unchanged). Hence, we expect the measures of the rates obtained for different system size to collapse onto each other upon the following rescaling:

$$R \rightarrow -\frac{\log(R\sqrt{N})}{N}. \quad (11)$$

This scaling is nicely confirmed by numerics, see Fig. 6B. We observe the collapse to a limiting curve, related to the barrier height through $\Delta F/T$. Note that ΔF is itself a function of temperature T , calculated in [29].

3 Representation of space(s) in the hippocampus: decoding and data analysis

3.1 Decoding neural representations with effective Ising networks

The problem of decoding which brain state is internally represented from the observation of neural activity has a natural application in experiments involving the simultaneous recording of a neural population [31]. The *decoding* problem can be tackled by learning statistical properties of a known set of brain states and classifying new observations accordingly, a problem which is deeply connected to high-dimensional classification in machine learning. An example is provided by the environmental memory in the hippocampus. Activity is

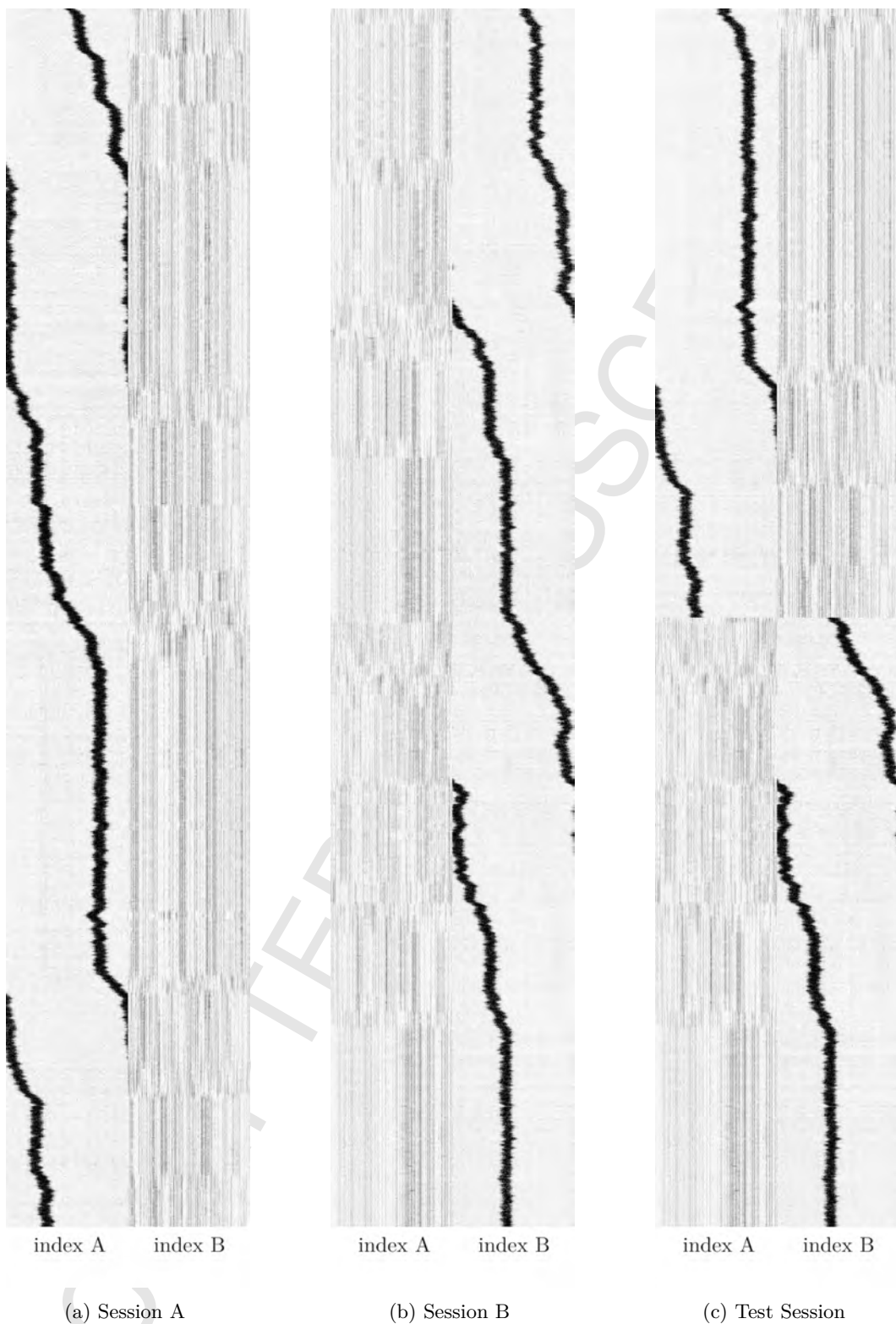


Figure 5: Monte Carlo simulation sessions of our memory model in the case of two 1D environments (random permutations), denoted by A and B. X-axis: states of the system $s(t)$ (black dots correspond to active neurons $s_i = 1$ and white dots to silent cells, $s_i = 0$), with neurons ordered in increasing order of their place field centers in the A (left part of columns) or B (right part of columns) permutations. Y-axis: time in MC rounds, increasing from top to bottom. The bump is forced to move rightwards with an external force, see [28]. In columns (a) and (b), the system is initialized with a localized bump of activity in environments, respectively, A and B. Column (c): Test simulations composed of the second halves of simulations reported in (a) and (b) used for decoding purposes, see text. Parameter values: $T = 0.006$, $N = 1000$, $w = 0.05$, $f = 0.1$.

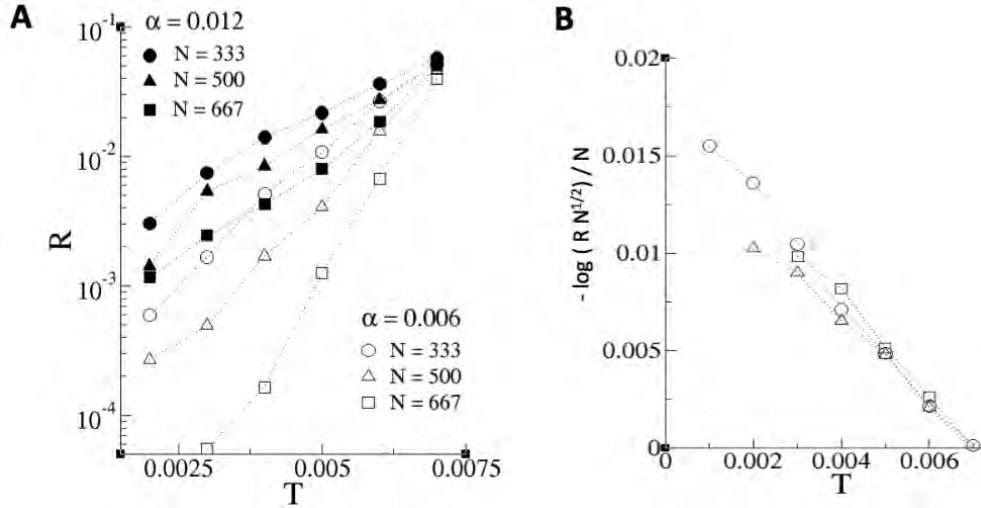


Figure 6: Map-to-map spontaneous transitions. **A.** rate R of transitions computed from Monte Carlo simulations, as a function of the temperature T , see [29]. Parameters: $f = 0.1$, $w = 0.05$. **B.** Replotting the same data as in panel **A** for $\alpha = 0.006$ after transformation in eqn (11) allows us to estimate the ratio of the free-energy barrier over temperature, see eqn (10).

recorded while the animal explores a set of environments. Each environment is associated to a memorized cognitive map, and these *reference sessions* are then used to learn models of the activity associated to each map. In turn, these models can be used to decode activity recorded during a controlled *test session*, i.e. decide which map this activity is associated to. During the test session, environmental cues are manipulated by the experimentalist, and decoding the internal response to the change of experimental conditions allows us to investigate the induced dynamics over internal representations [17, 32].

We can tackle the decoding problem by inferring a probability density function over the neural patterns \mathbf{s} for each brain state M , $P(\mathbf{s}|M)$ ⁵. These probability distributions can be used to decode the internal state M given an observation (a neural pattern) \mathbf{s} in the test session. More precisely, M is decoded by maximizing the log-likelihood

$$\mathcal{L}(M|\mathbf{s}) = \log P(\mathbf{s}|M). \quad (12)$$

This inference framework relies on the definition of a parametric probability function, whose parameters are inferred by solving the corresponding *inverse problem* from reference data. According to the maximum entropy principle, our choice is to use the family of graphical models [33–35] as parametric probabilistic functions. Depending on the reference sample size and/or the complexity of representations we can invert the Independent model, which accounts for the different average activations of neurons in different brain states, or make a step further and include correlations between neuron activities, defining an Ising model for each state M .

$$P(\mathbf{s}|M) = \frac{\exp\left(\sum_i h_i^M s_i + \sum_{i<j} J_{ij}^M s_i s_j\right)}{\mathcal{Z}^M(h, J)} \quad (13)$$

where \mathcal{Z}^M is a normalization constant

$$\mathcal{Z}^M(h, J) = \sum_{\mathbf{s}} \exp\left(\sum_i h_i^M s_i + \sum_{i<j} J_{ij}^M s_i s_j\right). \quad (14)$$

The core steps of the Ising decoding procedure are:

- (a) *Reference session.* For each brain state M , (a) collect samples of neural pattern in a known brain state M (reference session), and compute the frequencies p_i^M and pairwise joint frequencies p_{ij}^M of the recorded neurons; (b) find the Ising model that reproduces the same quantities on average, i.e. such that $\langle s_i \rangle = p_i^M$ and $\langle s_i s_j \rangle = p_{ij}^M$, where $\langle \cdot \rangle$ denotes the average over the probability distribution $P(\mathbf{s}|M)$. This is a highly non-trivial computational problem, reviewed in Section 3.2.

⁵For definiteness, we will hereafter consider the neural configuration at a given time, $\mathbf{s} = \{s_1, s_2, \dots, s_N\}$, to be the set of binarized neural activities of the neurons $i=1\dots N$ under consideration, with $s_i = 0$ if neuron i is silent, 1 if it is active, see Section 3.2 for more details.

- (b) *Test session.* Given a neural pattern from the test session \mathbf{s}^t , compute the log-likelihood of each brain state, and decode the internal state as the most likely one

$$M^t = \underset{M}{\operatorname{argmax}} \mathcal{L}(M|\mathbf{s}^t). \quad (15)$$

Within this framework we can therefore decode the neural representation from the observed neural pattern. This procedure has been applied to experimental data from the hippocampus, showing good performance in retrieving the explored environment from neural activity [32]. Similar procedures have been successfully applied to other brain regions, see for instance [36–39].

Before applying the decoding procedure in the context of the representations of space, let us review how the effective Ising model in eqn (13) can be fitted from data.

3.2 Inference of effective Ising models from data

We discuss in this Section the problem of the inference of a graphical model from data [34, 38, 40–42]. Data are defined here as a set of B recorded configurations of N variables, $\mathbf{s}^b = \{s_1^b, s_2^b, \dots, s_N^b\}$, with $i = 1, \dots, N$; s_i^b denotes the value of the variable at site i in configuration b . Variables s_i^b can be real valued, binary, or multi-categorical (Potts state). Two cases of great practical interest are:

- Neurons can be described by binary variables $s_i = 0, 1$, expressing whether they are silent or active, i.e. emit a spike. Spiking times obtained from multi-electrode recordings [38, 43–45] can be processed into a series of $B = T/\Delta t$ recorded neural configurations \mathbf{s}^b , by dividing the recording time T in small time windows $b = 1 \dots B$ of duration Δt ; then, s_i^b is equal to 1 if neuron emits one or more spikes in time bin b , to 0 if it remains silent.
- Amino acid s_i at site i in a protein sequence can take 20 different values. Data configurations are sequences from a family of homologous proteins, assumed to share a common 3D fold and biological function, collected in protein databases [46–50].

For the sake of simplicity, we assume hereafter that variables take binary values, $s = 0, 1$. We further assume that the distribution over configurations \mathbf{s} of N such variables is given by the Ising model defined in eqn (13); to lighten notations, we will drop the M subscript hereafter. The model is parametrized by N fields h_i and $\frac{1}{2}N(N-1)$ couplings J_{ij} .

We assume that the different data configurations are independently drawn from $P(\mathbf{s}|h, J)$ in eqn (13). Hence, the probability of the data reads

$$\prod_{b=1}^B P(\mathbf{s}^b|h, J) = \exp[-B S(h, J)] \quad (16)$$

where the cross-entropy S is

$$S(h, J) = \log Z(h, J) - \sum_i h_i p_i - \sum_{i<j} J_{ij} p_{ij} \quad (17)$$

depends on the data through the single-site and pairwise frequencies

$$p_i = \frac{1}{B} \sum_b s_i^b \quad \text{and} \quad p_{ij} = \frac{1}{B} \sum_b s_i^b s_j^b. \quad (18)$$

The best values for the fields and the couplings are the one minimizing S . While S is a convex function of its arguments, the minimum is not guaranteed to be finite. For instance, the minimum of S is realized at $J_{12} = -\infty$ when neurons 1 and 2 never spike together ($p_{12} = 0$). This problem can be avoided by including a prior, also called regularization, over the fields and couplings. Usual regularization schemes include adding the L_1 and/or L_2 norms of the couplings, to avoid small nonzero couplings or couplings with very large, unrealistic values. Another regularization scheme consists of imposing a small rank for the coupling matrix, see Section 5.6.

The computational problem in the minimization of eqn (17) is the calculation of the partition function $Z(h, J)$, which is generally intractable as it involves a summation over the 2^N configuration of the systems. Some methods to solve the inverse Ising problem bypass the calculation of Z , such as Boltzmann Machine algorithm [40], the Pseudo-Likelihood approximations [34, 41], the minimum probability flow [51], or resort to approximate expressions for Z , e.g. mean field [52], high-temperature expansions [53], and adaptive cluster expansions [42, 54].

Once the Ising model has been inferred, it can be used for various tasks:

- Extract structural information on the connectivity/coupling matrix between the variables. In the case of neurons, this *functional connectivity* is not physiological (synaptic), but is an effective set of couplings depending on the brain state [31, 36, 55]. In protein covariation analysis, it has been shown that large couplings often coincide with amino acids in contact on the three-dimensional structure of the protein [48, 49].
- Use the inferred model to score new configurations and decide if they are compatible with the data in the training data set. We will see a direct application in Sections 3.3 and 3.4.
- Generate new configurations through Monte Carlo simulations. This can be very useful to obtain in silico data with the same features as the ones in the training set, for instance, new proteins with the same structure or function as natural proteins.

3.3 Back to model: the subsampling problem

The theoretical model for spatial memory in place-cell populations from Section 2.2 shows remarkable features compatible with the recall of brain states at the level of population activity [56, 57]. The existence of a *clump phase*, in which the system is maintained in a local minimum of self-sustained localized activity, is compatible with the attractor-neural-network (ANN) general paradigm of cognitive functions being represented by collective states of the neural network. The presence of spontaneous transitions from one representation to another is consistent with the flickering phenomena triggered by weak inputs observed in the rat hippocampus [17], where spatial memory is thought to be stored and retrieved [58–60]. However, theoretical results were obtained in the limit of very large systems, which is also the case of real brain regions ($\sim 10^9$ neurons), while electrophysiological setups permit to record simultaneously a much smaller ($< 10^2$) number of neurons. It is natural to wonder to what extent such a small number of neurons could provide information about the collective state of the whole population.

Hereafter, we describe an attempt to draw a parallel between experimental conditions of multi-array recordings and the theoretical model for environment memory in the attractor neural network framework. We first design a Monte Carlo simulation that mimics an experiment with two memorized environments, referred to as A and B. We simulate single-environment *reference sessions* by forcing the activity to explore local minima corresponding to the memorized environments in a system with a relatively large number ($N = 1,000$) of neurons. We then address the question if a small, randomly selected, set of neurons (here, $N_{sam} = 33$ over 1,000) could provide enough information to perform the decoding procedure and infer the time course of the spatial representations from neural activity. As place cells are non topographical, i.e. cells that are physically nearby in the hippocampus can have distant place fields, recording a spatially located population of cells can be thought of to be equivalent to a random subsample in the place-field abstract space.

The decoding task is finally performed on the test session using Ising and independent models learned from the reference sessions, and their decoding capability is tested in a classification problem on a test session composed by samples from both states. The relationship between true and inferred coupling is then analyzed.

3.3.1 Simulations: constructing reference and test sessions

Monte Carlo simulations are conducted as follows:

- First we define two 1D environments, hereby referred to as A and B, through their two random place-field permutations, denoted by π^A, π^B .
- From these two environments, two coupling matrices $J^M, M \in \{A, B\}$, are created using learning prescription described in eqn (5):

$$J_{ij}^M := \begin{cases} \frac{1}{N} & \text{if } \frac{1}{N} |\pi^M(i) - \pi^M(j)| \leq \frac{w}{2}, \\ 0 & \text{otherwise.} \end{cases} \quad (19)$$

- A unique coupling matrix J is then constructed as point-sum of the two single-environment matrices: $J_{ij} = J_{ij}^A + J_{ij}^B$.
- simulations are performed, with $n = 10^4$ Monte Carlo steps, each one starting from an initial neuronal condition localized in one of the two reference environments M . To maintain the total activity constant, we select, at each algorithm step, one active spin $s_i = 1$ and one silent spin $s_j = 0$. The flip trial is then defined as the joint flip of these spins.

- an additional small force is added to make the bump exhaustively explore the one-dimensional map, by an asymmetric term in the energy. This results in a left-right asymmetry in the Monte Carlo acceptance rule:

$$\Delta E = \sum_{k \neq i, j} (J_{ik} - J_{jk}) s_k + A^M(i, j) \quad (20)$$

with $A^M(i, j)$ being a right-pulling force in the environment M , namely

$$A^M(i, j) := \frac{A}{fN^2} \times (\pi^M(i) - \pi^M(j) + N\epsilon_M(i, j)) \quad (21)$$

where A controls the magnitude of the pulling force, $\pi^M(i)$ is the position occupied by the place field of neuron i in environment M , and $\epsilon_M \in \{-1, 0, 1\}$ ensures periodic boundary conditions.

Two simulations, one for A and one for B, are conducted. Parameters are carefully chosen such that the *clump* phase is maintained: the bump thoroughly explores the environment, and no spontaneous transitions occur. In other words, the same system is sampled in one of the two maps during the whole simulation; this mimics the single-environment exploration of the rodent during training sessions [17]. Two *reference sessions* are defined using the first half (5,000 steps) of each simulation, and a *test session* is constructed by concatenating the second halves, for a total of 10,000 total time steps. Parameters used in the following analysis are: $T = 0.006$, $N = 1000$, $w = 0.05$, $f = 0.1$.

3.3.2 Decoding Results

As a measure of decoding precision we use the true positive rate (TPR), i.e. the overall fraction of correctly-classified neural pattern.

$$\text{TPR} := \frac{\# \text{ correctly classified time steps}}{\# \text{ total time steps}} \quad (22)$$

We obtain

$$\begin{aligned} \text{Ising model} : \quad & \text{TPR} = 0.928 \\ \text{Independent model} : \quad & \text{TPR} = 0.491 \end{aligned} \quad (23)$$

The difference between the use of independent and Ising model, shown in Fig. 7, is remarkable. The independent model, in which all couplings are set to zero, accounts only for the average firing rates of the cells. It shows no decoding capability at all, with a TPR equal to 0.49 (compatible with random guessing). This could be expected from the fact that the localized bump of activity, which represents the position of the rat within the retrieved a map, moves along the entire environment during reference sessions. Hence the average activity of all cells is close to f in both maps. The independent model, which only uses information on averages to decode the activity, is therefore unable to achieve useful discrimination.

Conversely, the Ising model exhibits an impressive performance in the decoding task. As shown in Fig. 7b, the time course of the likelihood difference $\Delta \mathcal{L}$ allows us to unambiguously decode the spatial representation as a function of time. This difference is also clear from the scatter plot of the likelihoods in the test session, which shows a well-separated pattern in the plane, contrary to the Independent model (Fig. 8).

3.3.3 Inferred vs. true couplings

The application of inference routines to a simulated neural network allows us to investigate the relationship between functional couplings, i.e. the inferred J_{ij} in the inverse Ising model, and the real coupling strength, defined in eqn. (19). We show in Fig. 9 the couplings inferred between the neurons as functions of the distances between their place-field centers in each map. We observe that:

- Couplings decay very rapidly with the distance, on a typical scale compatible with both wN and fN , and the width of the bump; Note that w, f have similar values in the simulations. At long distances, couplings are independent of distance, and equal to a negative value. The presence of many long-range inhibitory couplings, clearly visible in the histograms of Fig. 10, is a natural consequence of constraint (4) on the level of activity.
- The magnitude of coupling at small distances, $\sim 2 - 3$ in Fig. 9, is much larger than the one of the 'true' couplings in the model, equal to $J^0 = \frac{1}{TN} = 0.167$. This suggest that the inferred couplings are *effective*, and would coincide with the true couplings only in the limit of perfect spatial sampling ($N_{sam} = N$).

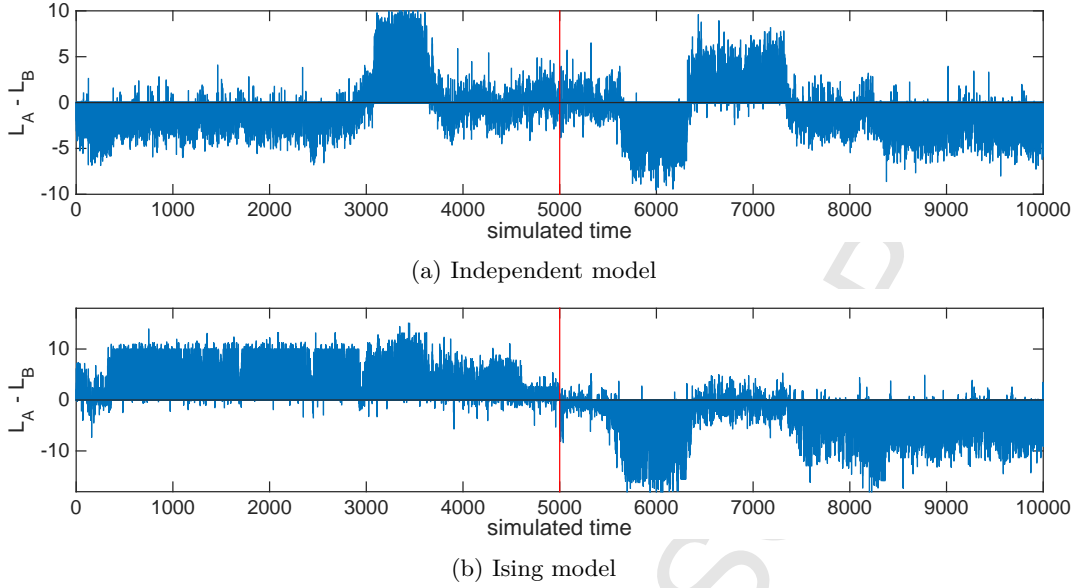


Figure 7: Log-likelihood difference $\mathcal{L}_A(t) - \mathcal{L}_B(t)$ along the test session using independent model and Ising model on the montecarlo test session. The first half of the test session is sampled from environment **A**, the second half from environment **B**.

To better understand the value of the inferred couplings, let us compute the statistical moments of the neurons. As explained above in the description of the independent-cell models, all neurons have the same average activity in both environments:

$$p_i^A = p_i^B = f, \quad \forall i. \quad (24)$$

We can also estimate easily the joint probability that two neurons are active. In the large- N limit, the true couplings vanish as they scale as $1/N$. Hence, spins become two-by-two independent in a ground state of the Hamiltonian, that is, when the bump is centered around a given position x . Conditioned to x , and defining the position of the place-field center of cell i in map M through

$$x_i = \frac{\pi^M(i)}{N}, \quad (25)$$

we have

$$\langle s_i \rangle_x = \rho(x_i - x), \quad \langle s_j \rangle_x = \rho(x_j - x), \quad \langle s_i s_j \rangle_x - \langle s_i \rangle_x \langle s_j \rangle_x \sim \frac{1}{N}, \quad (26)$$

where ρ is implicitly centered in 0 in the above expression.

However, we have to average over the position x of the bump that moves across the environment (Fig. 5). Doing so, we obtain the pairwise activity, see eqn (36) and Fig. 12 in [23]:

$$p_{ij}^M = \int dx \rho(x_i + x) \rho(x_j + x) \quad \forall i, j. \quad (27)$$

This effective matrix of pairwise activities therefore depends on the map, which explains why the Ising model, contrary to the Independent model, is map-specific and can efficiently decode the representation. However, the effective correlation between neurons, $p_{ij}^M - f^2$, does not scale as $\frac{1}{N}$: the Ising couplings are thus effective interactions, not simply related to the true couplings in the model. We expect this statement to hold also for the functional couplings inferred from real recordings and their physiological, synaptic counterparts.

3.4 Analysis of multi-electrode recordings in CA1

The inference routines and the test-session validation described above are directly applicable to micro array recordings of neural activity in vivo. As previously introduced, one can record the brain area activity in a collection of stable memory states, build models from these activities, and use them to decode representations (neural states) in a successive test session.

A good testing ground for this analysis is spatial memory in the rat hippocampus. Once environments have been memorized by the animal, it is relatively easy to collect samples of one memory state by letting the rat explore the corresponding environment. In a recent experiment, conducted by Jezek et al. [17], environmental conditions (light cues) are abruptly changed to trigger instabilities in the evoked spatial maps in the test

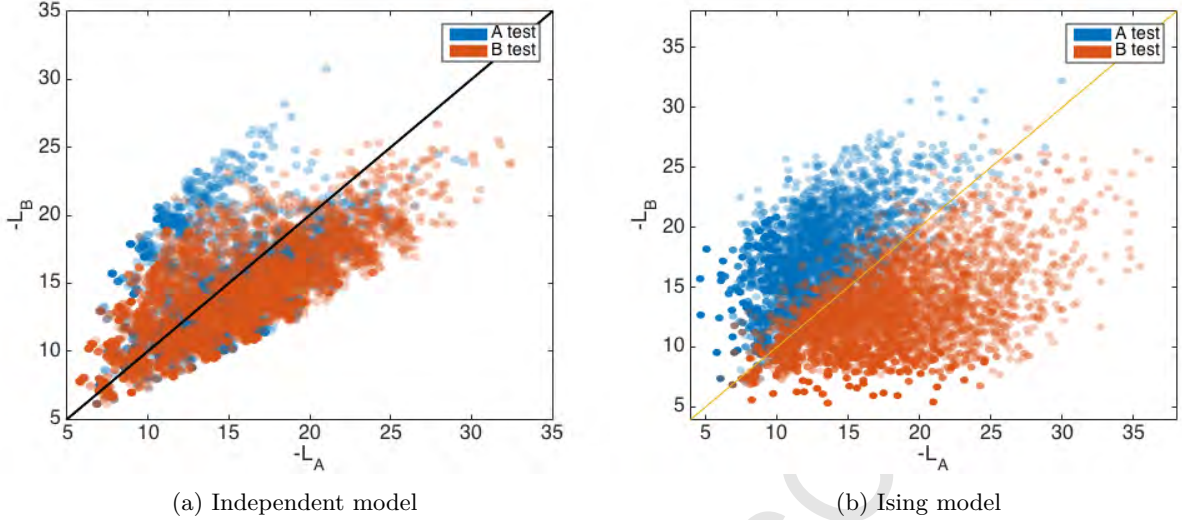


Figure 8: Log-likelihood scatters computed from the Independent (a) and Ising (b) models. Each dot represents the value of $-\mathcal{L}_A$ and $-\mathcal{L}_B$ for each neural configuration \mathbf{s}^t during the Monte Carlo test session.

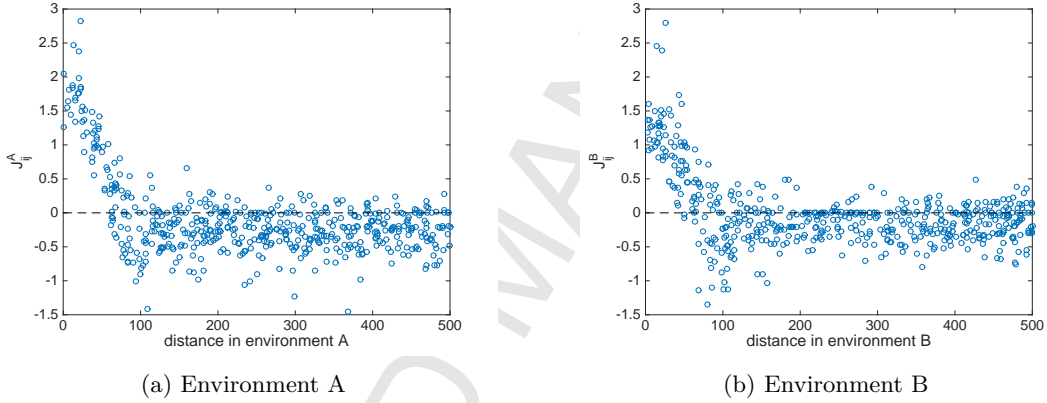


Figure 9: Inferred coupling J_{ij} vs. distance $|\pi^M(i) - \pi^M(j)|$ between the place-field centers of the corresponding neurons in environment $M = A$ (left) and $M = B$ (right).

session. By decoding which representation is being expressed during the test session as a function of time, we can investigate the response and fast dynamics of the memory state in the hippocampal network.

The Ising inference method has been used in this context to decode which map, denominated A or B, is retrieved during test sessions with unknown environmental conditions. A test session recorded from hippocampal CA1 containing an environment-switch (teleportation) event is shown in Fig. 11. Contrary to the decoding analysis performed on the subsampled theoretical model, the Independent-cell model shows good performance in decoding task on real recordings:

$$\begin{aligned} \text{Ising model: } & \text{TPR} = 0.85 \\ \text{Independent model: } & \text{TPR} = 0.81 \end{aligned} \quad (28)$$

One simple possible reason is that, as a result of rate-remapping taking place in CA1, the average firing rates of cells are different in both maps. These discrepancies can be exploited by the Independent-cell model to adequately decode the loaded map as a function of time in the test session. For a more detailed analysis and comparison of inference methods applied to hippocampal data see [32, 62].

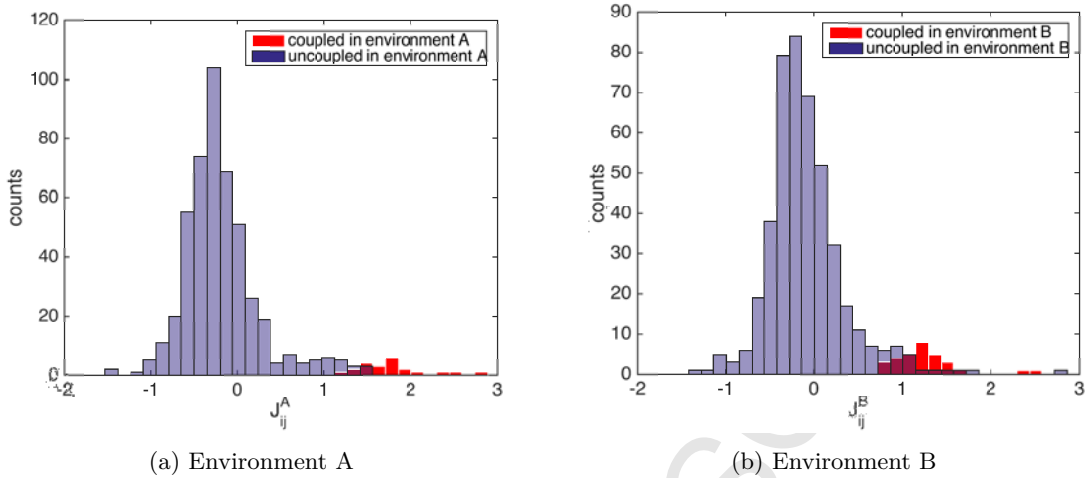


Figure 10: Relationship between true couplings and inferred couplings. In purple, histogram of inferred couplings. In red, inferred couplings corresponding to truly connected neurons in the environment.

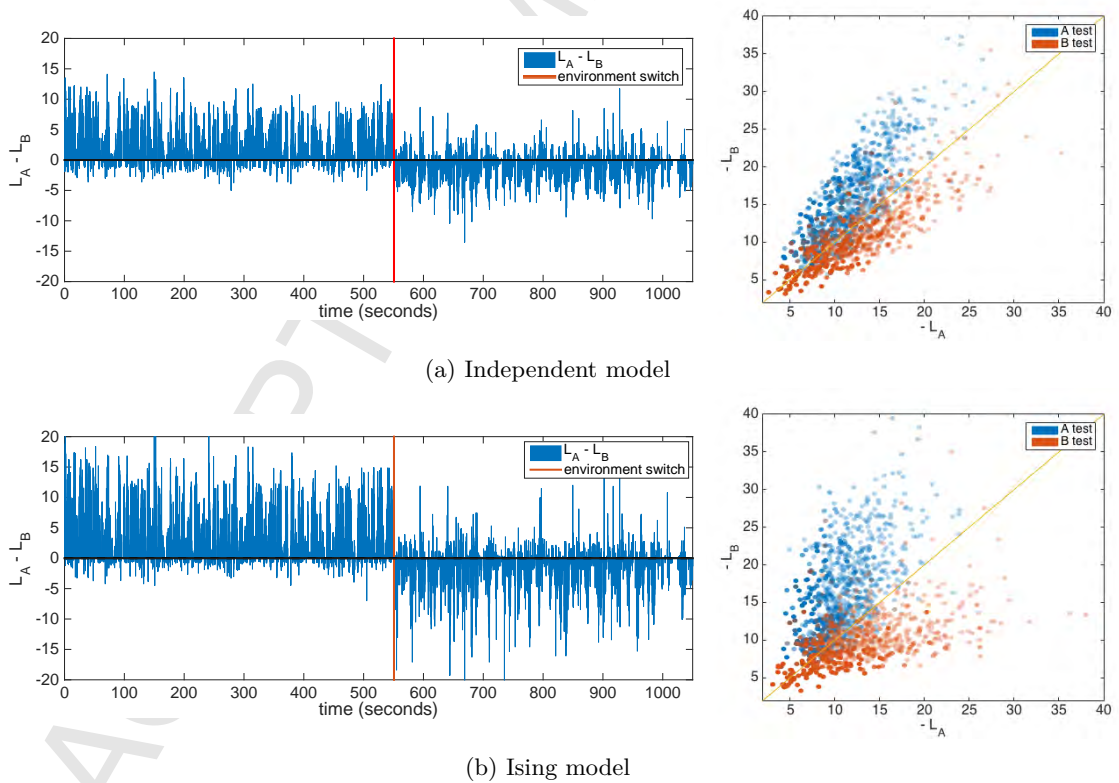


Figure 11: Left: Log-likelihood difference $\mathcal{L}_A(t) - \mathcal{L}_B(t)$ along the test session using Ising and independent decoder on CA1 hippocampal data. The environmental conditions are abruptly changed from A to B in correspondence to the red line. Right: log-likelihood scatters, see Fig. 8 for explanations.

4 Representations of Data in Machine Learning

4.1 Introduction

We start by the definition of a data representation. Suppose we are given a set of P data samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(P)}$ of a N -dimensional random variable \mathbf{X} having joint density $\rho(\mathbf{X})$. A data transformation is a deterministic transformation from the multidimensional vector space of data into another one:

$$F : \mathbf{x} \in \mathbb{R}^N \rightarrow \mathbf{x}' = \mathbf{F}(\mathbf{x}) \in \mathbb{R}^M, \quad (29)$$

where M can be larger or smaller than N . In general, F is assumed to be differentiable, but is not necessarily invertible. We say that the random vector \mathbf{X}' is a representation of the original random vector \mathbf{X} . Changing the representation of a random variable can be often extremely helpful in data science because: i) it allows for better visualization and understanding of the process that generated the data; ii) the performance of machine-learning algorithms, such as classification or clustering methods heavily depends on the choice of representation used.

Although it is not obvious that a given representation is good, it is clear that many, many representations are useless: if $F(\mathbf{X}) = 0, \forall \mathbf{X}$, then \mathbf{X}' is a trivial random variable, and does not carry any information about \mathbf{X} . More generally, it is clear that any transformation F that does not vary strongly across the support of \mathbf{X} is of little use. On the opposite, $F = Id$ is not of much use either, since the properties of the data distribution have not changed. Typically, a good data representation \mathbf{X}' must have helpful properties that \mathbf{X} does not have, such as low dimensionality, independence between components or sparse values, while carrying information on the original random vector \mathbf{X} . Thus, the transformation F must depend on $\rho(\mathbf{X})$ and should be learnt. Once learnt, a data representation can often shed light on how the data was generated: one can find so-called 'features', i.e. frequent collective modes of variation in the data, find a partition into classes, discover outliers...

Moreover, a good data representation can significantly improve the performance of subsequent machine learning tasks, by retaining only useful information about the data sample. For instance, in the so-called deep neural networks, one learns a sequence of data transformations, e.g. to predict label from an image. By using non-linearities and so-called pooling architectures, the learnt intermediate representations of the data can become invariant, e.g. w.r.t. noise, shifts, rotations [63]. Deep neural networks have brought remarkable breakthrough in many areas, such as visual and speech recognition, natural language processing,... [6, 7].

We now illustrate these concepts with two examples of great relevance in applications.

4.2 Example 1: Dimensionality reduction

One important subclass of data transformations is Dimensionality Reduction. One aims at compressing a random vector \mathbf{X} of typically high dimension N , into a smaller random vector \mathbf{X}' of dimension $M < N$, e.g. $M = 2$ or 3 , while keeping as much information as possible about \mathbf{X} . Such compression is motivated by the fact that data very often lie in or close to a subspace of much lower dimension than N . This is the so-called 'manifold hypothesis'. Indeed, consider for instance a data set constituted by pictures of a person's face, taken in many different positions; each picture is made of, say, 1000×1000 pixels. It is clear that this data set is a very small subset of all possible 1000×1000 colored pictures, which define a $3 \cdot 10^6$ -dimensional vector. The reason is that, for a given face, there are only ~ 50 varying degrees of freedom (the position of all muscles), a very small number compared to 10^6 [64]. Hence, all data points lie in a (non-linear) manifold, of very low dimension M compared to N . More generally, the variability in the data often comes from a small number of explanatory latent factors that affect all components, and we would like to recover them.

In all generality, we do not have good and general methods to learn functions that turns an image into this kind of 'muscle positions' representation. Some simpler dimensionality reductions can nonetheless be learnt and be extremely useful. For instance, dimensionality reduction can be obtained through a simple linear transformation:

$$\mathbf{X}' = W \mathbf{X}. \quad (30)$$

where the weight W is a $M \times N$ rectangular matrix that must be trained on the data in order to retain as much information as possible from \mathbf{X} . An interesting choice of matrix W is obtained by the Principal Components Analysis (PCA) algorithm: the rows $W_{i,\cdot}$ are the eigenvectors corresponding to the i 'th largest eigenvalues of the empirical data covariance matrix $C_{ij} = \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle$, where the average is computed over the data; this choice will be justified in Section 5. Such transformation mainly serve two purposes. The first one is to provide a better understanding of the data by visualizing it: one computes a 2 or 3 dimensional-representation of the data distribution ρ ; then each data point is represented in a 2 or 3D space. For example, one can compute the 2D PCA representation 28×28 images of digits from the MNIST handwritten digits dataset, vectorized as 784 dimensional vectors, see Fig. 12; the scatter plot shows two distinct clusters,

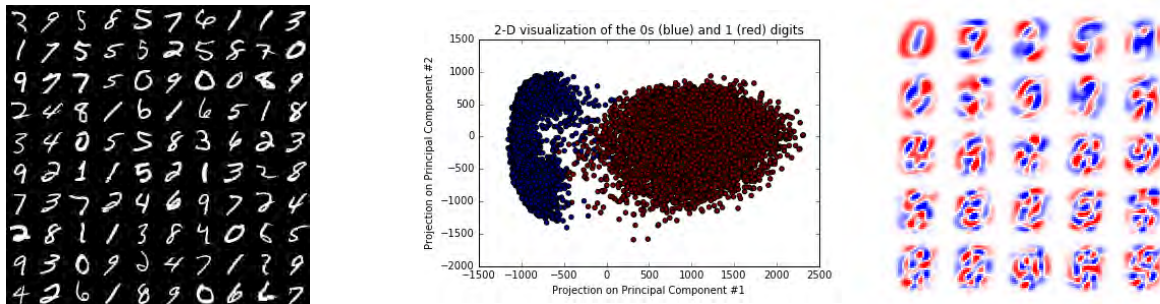


Figure 12: (a) 100 digits extracted from the MNIST dataset; each digit is made of 28×28 pixels. (b) A 2-dimensional PCA representation of the MNIST handwritten digits data set. Each point is a different image with x and y coordinates being the value of the first and second components of the representation. Here, only the digits 0's (dark blue; left cluster) and 1's (dark red; right cluster) are represented. (c) Visualization of the weight matrix W . Each image is a principal component vector $W_{i,\cdot}$; blue (resp. red) pixels denote large positive (resp. negative) values. the PCA representation is obtained by computing the set of overlaps between an image and each principal component vector

corresponding to two digit types (0s and 1s). A more interesting illustration is the interpretation of molecular dynamics simulation of complex systems, made of many strongly interacting and heterogeneous microscopic components. Observing the dynamics of such systems, e.g. a protein described at the atomic level, amounts in practice to look at thousands of correlated time series. Principal component analysis offer low-dimensional projections of these time traces, and allows one to visualize collective motions underlying the evolution of the system, see [65] for a recent review on applications to biomolecules, including nucleic acids and proteins. The second purpose of dimensionality reduction is to overcome the so-called curse of dimensionality. In very high dimensional spaces, most datasets sample only very sparsely the vector space \mathbb{R}^N . Consider for instance the following supervised learning problem. We are given a training data basis of 10,000 100×100 grayscale (normalized between 0 and 1) images of cats and dogs, with binary labels attached, and we want to train a parametric model to classify whether images are cats or dogs. At this point, it is useful to think that this classification task is essentially an interpolation problem: there exist a mathematical function $\theta : \mathbf{X} \rightarrow y \in \{0, 1\}$ that assigns 0 to cats and 1 to dogs. We observe pairs of values $(\mathbf{X}^i, y^i = \theta(\mathbf{X}^i))$, with $i = 1 \dots 10,000$, and want to interpolate the values of θ for new test images. This interpolation problem would be trivial if the input space was densely sampled, e.g. if for any point in \mathbb{R}^N there would be a training data point at distance $\leq \epsilon$. In practice, it is impossible because the latter condition requires about ϵ^{-N} data points, which is out-of-reach when N is large.

One possible way-out is to first learn a new data representation of lower dimension, $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, e.g. using PCA, and then train a classification model of the form: $y = \theta(\mathbf{x}')$. If the low dimensional representation keeps relevant information about the nature of the image, then learning can be performed. One popular application of PCA for supervised learning is the 'eigenface' face recognition algorithm. A PCA representation is trained on a data set of faces, before applying supervised learning [66]. The eigenface algorithm is considered among the first successful face recognition algorithms.

4.3 Example 2: Extracting latent features from data

The variability in real-world data, such as images, can often be decomposed into a set of largely independent modes of variation. For instance, two faces are different because some of their parts are different: nose, ears, lips... At a lower level of description, an image can contain or not an edge at a given location, or at some angle or scale, and two different images have different set of activated edges. Extracting these so-called 'features' is of particular interest for machine learning, in particular for classification, because the decision function $y = \theta(X)$ that must be learnt may be expressed more easily as a function of these 'features' X' than from the raw pixels X . For instance, one could achieve better results by expressing $\theta(X')$ as a linear function of X' , instead of a higher order polynomial of X . Moreover, the learnt representations have interesting statistical properties, such as low statistical dependence between modes, invariance with respect to irrelevant perturbations of the data such as corruption by noise... that can be used for denoising. Some notable algorithms for unsupervised feature extraction are Independent Component Analysis (ICA) [67], sparse autoencoders [68], and sparse dictionary learning [69]. We display in Fig. 13 the features learnt by ICA applied to the MNIST digits data set. The features learnt correspond to individual handwritten strokes, unlike PCA where the principal component do not have a simple interpretation. Interestingly, the features

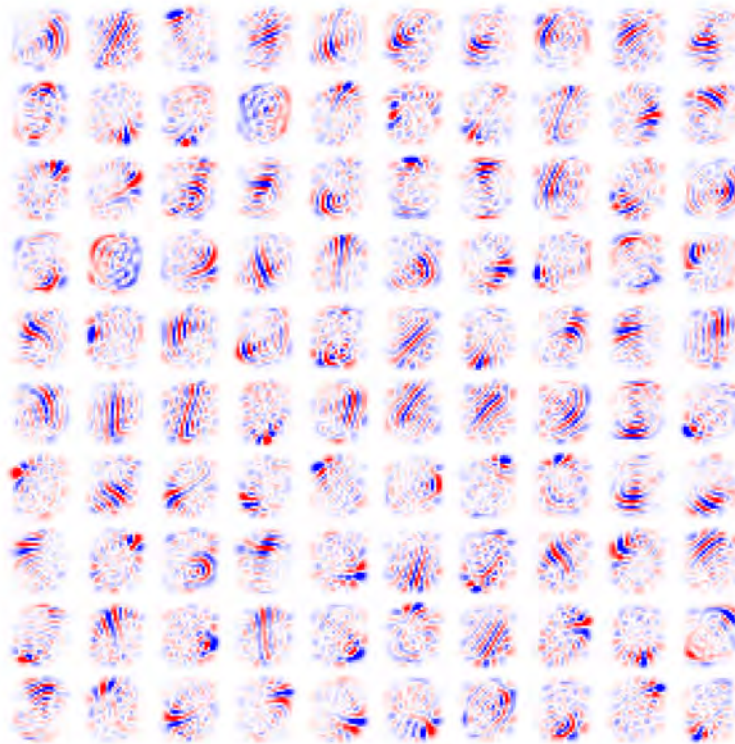


Figure 13: Features learnt by Independent Component Analysis on MNIST

found by sparse dictionary learning applied to natural images dataset qualitatively match very well the receptive fields of neurons in the visual cortex of mammals, such as in monkey [70,71]. Feature extraction carried out in the brain bear strong analogies with machine-learning procedures [9].

5 Low-Dimensional Representations: Principal Component Analysis (PCA)

In this section, we focus on the PCA transformation introduced in Section 4.2. To cast PCA in a Bayesian framework, we start with a basic reminder about Bayes's approach to inference.

5.1 Mathematical reminder: Bayesian inference

We observe a data sample σ , and would like to fit these data with a model parametrized by some variable τ . We assume that both σ and τ are random variables, with a joint distribution $p(\sigma, \tau)$. According to the definition of conditional probabilities, we may write

$$p(\sigma, \tau) = p(\sigma|\tau) \times p(\tau) = p(\tau|\sigma) \times p(\sigma) . \quad (31)$$

In the first equality, $p(\sigma|\tau)$ is the the probability of the data given the model parameters, also called *likelihood* of the model parameters given the data. The second term, $p(\tau)$, is the *prior distribution* over model parameters. The expression can be rewritten using the posterior distribution of the parameters given the observations $p(\tau|\sigma)$ (which can be maximized, sampled from...), and the overall probability $p(\sigma)$ of the data to be generated by the class of models under consideration. This posterior distribution is given by Bayes formula:

$$p(\tau|\sigma) = \frac{p(\sigma|\tau)p(\tau)}{p(\sigma)} , \quad (32)$$

which is simply derived from eqn (31). One historical application of the Bayesian inference formula is Laplace's statistical proof that boys and girls have different birth rates. Laplace had access to the number of boys and girls born in Paris between 1745 and 1770: $\sigma = 245,945$ girls out of $P = 245,945 + 251,527 =$

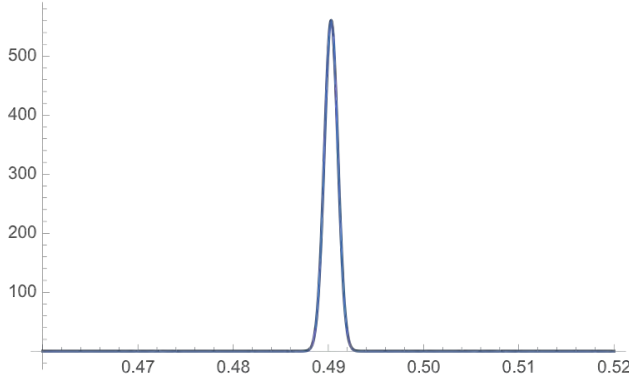


Figure 14: The posterior probability density of the female birth rate τ for Laplace's birth rate problem according to eqn (34). Notice that $p(\tau = 0.5)$ is very small, but non zero.

497,472 babies born during this time period. Although the numbers of male and female births are different, it is not possible to know a priori whether the discrepancy comes from a statistical fluctuation or from a systematic difference in birth rates. Laplace assumed that each birth was a realization of an independent and identically distributed random variable, giving a girl with probability τ and a boy with probability $1 - \tau$. Under this basic assumption, σ follows a binomial distribution $\mathcal{B}(P, \tau)$, with a likelihood:

$$p(\sigma|\tau) = \binom{P}{\sigma} \tau^\sigma (1 - \tau)^{P - \sigma}. \quad (33)$$

Assuming a uniform density prior $p(\tau) = 1$ over $\tau \in [0; 1]$, the posterior distribution reads

$$p(\tau|\sigma) = C \tau^\sigma (1 - \tau)^{P - \sigma}, \quad (34)$$

where C is a normalization constant, and is shown in Fig. 14. It is then easy to calculate the mean value and the standard deviation of τ with the posterior distribution, with the results $\text{Mean}(\tau) = 0.490291$ and $\text{Std}(\tau) = 0.007117$. The probability that τ is actually larger or equal to $\frac{1}{2}$ is given by the integral of $p(\tau|\sigma)$ over the $\tau \in [\frac{1}{2}; 1]$ interval, and is approximately equal to 10^{-42} . This extremely small value makes it very unlikely that the discrepancy between the large numbers of female and male births is due to a pure statistical fluctuation.

5.2 Multivariate Gaussian variables

A popular example of parametric model is the multivariate Gaussian distribution, used to model sets of continuous random variables exhibiting correlations. Hereafter, we will assume that all random variables have zero mean for the sake of simplicity. Given a vector of N variables $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_N)$, we write:

$$\rho(\boldsymbol{\sigma}|\boldsymbol{\tau}) = \frac{\sqrt{\det \boldsymbol{\tau}}}{(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{\sigma}^T \cdot \boldsymbol{\tau} \cdot \boldsymbol{\sigma}\right) \quad (35)$$

where $\boldsymbol{\tau}$, called *precision matrix* is a symmetric, positive definite matrix that encodes the inter-dependencies between variables. Its off-diagonal entries can be interpreted as (minus) the couplings between the variables. For instance, with $N = 2$, $\tau_{12} > 0$ means that the configurations (a, b) and $(-a, -b)$ are more likely than the configurations $(-a, b)$ and $(a, -b)$ (when $a, b > 0$), leading to a positive correlation between σ_1 and σ_2 . Given a data set of P samples $\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(P)}$, one can compute analytically the maximum likelihood estimator of the precision matrix:

$$\boldsymbol{\tau}^{MLE} = \operatorname{argmax}_{\boldsymbol{\tau}} \sum_{s=1}^P \log \rho(\sigma^{(s)}|\boldsymbol{\tau}). \quad (36)$$

To solve for $\boldsymbol{\tau}^{MLE}$, the gradient of the right hand side in the above equation reads

$$\frac{\partial}{\partial \tau_{ij}} \sum_{s=1}^P \log \rho(\sigma^{(s)}|\boldsymbol{\tau}) = -\frac{1}{2} \sum_{s=1}^P \sigma_i^{(s)} \sigma_j^{(s)} + \frac{P}{2} (\boldsymbol{\tau}^{-1})_{ji}. \quad (37)$$

We recognize that first term is the empirical data covariance matrix, C . The gradient vanishes -and it is easy to show that this corresponds to a global maximum- when

$$\boldsymbol{\tau}^{MLE} = C^{-1}. \quad (38)$$

The inversion can be performed numerically as long as $P \geq N$; for $P < N$, the data covariance matrix is not full rank. However, finite sampling effects of order $\frac{1}{\sqrt{P}}$ in C result in error on $\boldsymbol{\tau} = C^{-1}$ of the order of $\sqrt{\frac{N}{P}}$. Thus, for large-dimensional data sets, i.e. when the ratio N/P is of the order of unity, we expect inference to be plagued with errors.

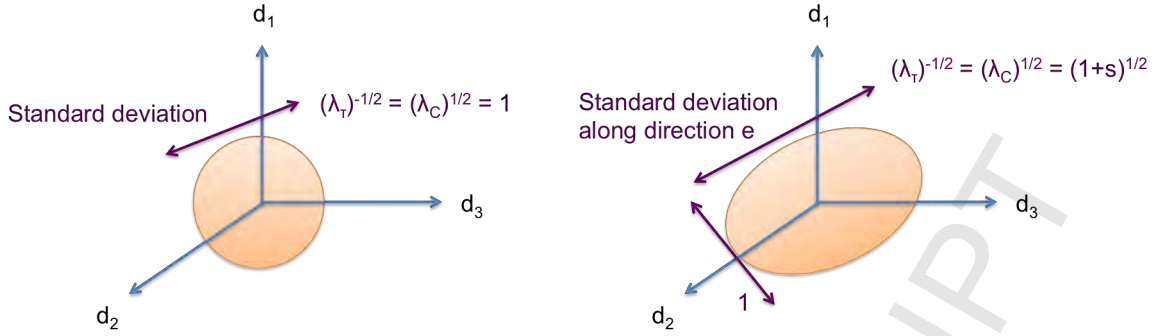


Figure 15: Probability density contours for (left) the null model ($\tau = Id$) and (right) the principal-component model of eqn (39).

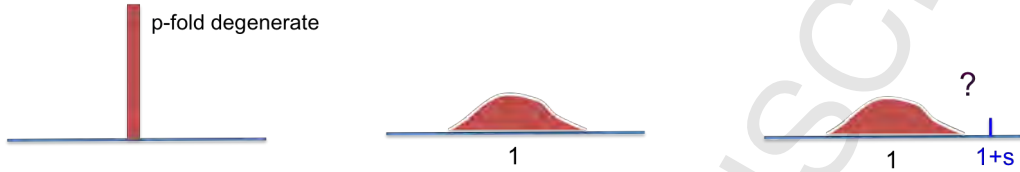


Figure 16: Distribution of the eigenvalues of the empirical covariance matrix for: (a) the null model with infinite sampling; (b) the null model with finite sampling; (c) the principal component model with finite sampling

5.3 Principal components as minimal models of interacting variables

The simplest model distribution over vector of random variables, each components being normalized to have zero mean and unit variance, is the independent one, which corresponds to $C = \tau = Id$. In this case, we have $p(\sigma|\tau) \propto \exp(-\frac{1}{2} \sum_i \sigma_i^2)$, and the resulting distribution is isotropic, see Fig. 15(left). A minimal non-trivial model is obtained by breaking this isotropy. We assume there exists a specific direction, denoted by $|e\rangle$, in the N dimensional space with a larger variance:

$$\tau = Id - \frac{s}{1+s} |e\rangle\langle e| \iff C = \tau^{-1} = Id + s |e\rangle\langle e|, \quad (39)$$

where $s > 0$. In this expression, $|e\rangle$, the *principal component* can be interpreted as a *collective mode* of variation of the data. Indeed, the random variable $\sigma_e = \sum_{i=1}^N e_i \sigma_i$ has variance $V_e = \langle e|C|e\rangle = 1 + s$ larger than 1, whereas it would be 1 if the σ_i were independent. The σ_i variables correlate in a way that makes σ_e have large variance, see Fig. 15(right).

Given a data set, maximum likelihood estimation can be performed analytically to infer the principal component $|e\rangle$. The likelihood writes

$$\rho(\sigma^{(s)} | |e\rangle) = \frac{\sqrt{\det \tau}}{(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{1}{2} \sum_{i,j} \sigma_i^{(s)} \tau_{ij} \sigma_j^{(s)}\right). \quad (40)$$

The $|e\rangle$ -dependent part of the log-likelihood is simply

$$L = \frac{s}{2(1+s)} \sum_{i,j} e_i e_j \left(\sum_s \sigma_i^{(s)} \sigma_j^{(s)} \right). \quad (41)$$

Hence, the MLE for the direction $|e\rangle$ (assumed to be normalized) is the top eigenvector (with largest eigenvalue) of the empirical covariance matrix $C = \frac{1}{n} \sum_s \sigma_i^{(s)} \sigma_j^{(s)}$. Although the inference can be performed easily for any covariance matrix, we do not expect that the inferred vector is always statistically significant, according to the discussion at the end of Section 5.2. For instance, even if the data are generated according to the null model $\tau = Id$, the empirical covariance matrix has a largest eigenvalue (> 1) due to finite sampling (Fig. 16). Similarly, if the data is generated according to the principal-component model but s is 'small' and P is moderate, the largest eigenvector of the empirical correlation matrix may be far away from $|e\rangle$ (Fig. 16). In the next section, we report analytical results derived using random matrix theory and statistical mechanics tools telling us when inference is possible.

```

In[1]:= lamedge[r_, s_] := (1 + s Sqrt[r]) ^ 2
rho[lam_, r_] := Sqrt[(lamedge[r, +1] - lam) (lam - lamedge[r, -1])] / (2 Pi r lam)
graph[r_] := Plot[rho[lam, r], {lam, lamedge[r, -1], lamedge[r, +1]}, PlotRange -> All]
Show[graph[.01], graph[.1], graph[.4], graph[.8]]
Show[graph[1.5], graph[2.], graph[4.], graph[8.]]

```

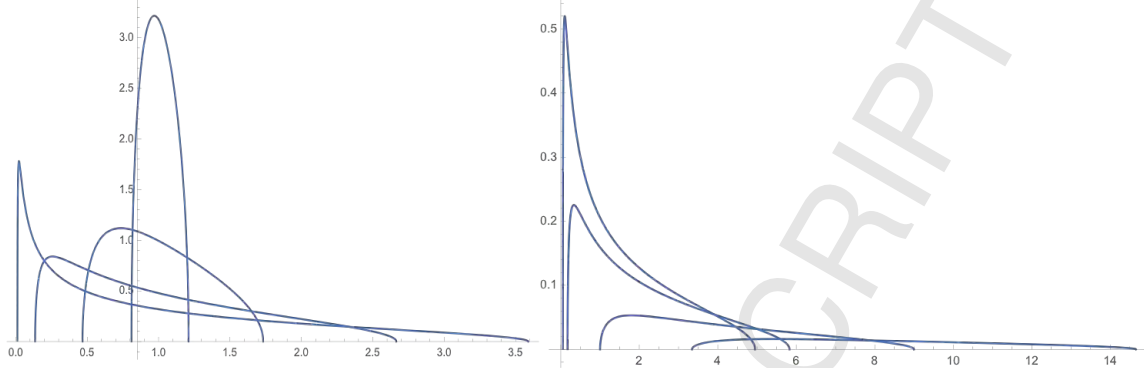


Figure 17: The Marcenko-Pastur distribution of eigenvalues, $\rho(\lambda)$, for various values of the noise level r , reported in the Mathematica script on the top. Left: $r < 1$. Right: $r > 1$; the Dirac peak in $\lambda = 0$ is omitted.

5.4 The retarded-learning phase transition

We first study the empirical covariance matrix $C_{ij} = \frac{1}{n} \sum_s \sigma_i^{(s)} \sigma_j^{(s)}$, and its spectrum when the data are generated according to the null model $\tau = Id$. We are interested in particular in the empirical density of eigenvalues:

$$\rho(\lambda) = \frac{1}{N} \sum_{\mu=1}^N \overline{\delta(\lambda_{\mu} - \lambda)}, \quad (42)$$

where $\{\lambda_1, \dots, \lambda_N\}$ is the set of eigenvalues of C . The overbar denotes the average over the realizations of the N data samples (s). Note that, in the large P, N limits with a fixed ratio $r \equiv \frac{N}{P}$, we expect that the spectrum attached to a random realization will coincide with the average spectrum ρ with high probability. The probability density of eigenvalues can be computed analytically using random matrix theory tools, in the limit case where the dimension N and number of data points P both go to infinity, and at fixed noise level r [72]. The result is the so-called Marcenko-Pastur distribution:

$$\rho(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi r \lambda}, \quad \text{with } \lambda_{\pm} = (1 \pm \sqrt{r})^2. \quad (43)$$

The expression above is valid for $r < 1$; for larger r , the covariance matrix is not full rank, and there is also a Dirac peak of mass $1 - \frac{1}{r}$ in $\lambda = 0$. The distribution of eigenvalues is plotted in Fig. 17 for various values of the noise level r . For very good sampling $r \rightarrow 0$, Wigner semi-circle law is recovered around $\lambda = 1$, as the different entries of the correlation matrix become essentially uncorrelated. Interestingly, the spectrum density can be quite wide when r is small: for instance, for $r = 1$, eigenvalues can be as large as $\lambda = 4$. As a consequence, these 'sampling noise' eigenvectors can screen away true principal components if s is not too large.

The same computation as above can be carried out for the principal-component model with $s > 0$, and shows the existence a phase transition, see Fig. 18 [73, 74]:

- If $r < s^2$ (weak noise regime), the largest eigenvalue is well above the 'bulk' of eigenvalues due to finite sampling, and the largest eigenvector $|v_1\rangle$ has a finite overlap $\langle e|v_1\rangle$ with the principal component $|e\rangle$.
- If $r > s^2$ (strong noise regime), the principal eigenvalue is inside the Marcenko-Pastur 'bulk' of eigenvalues, and $|v_1\rangle$ is merely noise, i.e. the overlap $\langle e|v_1\rangle$ vanishes in the large size limit.

In summary, recovering the principal component is impossible unless $P^* \sim \frac{N}{s^2}$ examples at least are presented, after which the error decays monotonously; hence the name of retarded learning coined in a slightly different context in [75]. This computation can be generalized to any finite number of eigenvalues $K > 1$, associated to the set $s_1 > s_2 > \dots > s_K$; each time the noise level r crosses s_k^2 , one more eigenvalue pops out of the noisy bulk of eigenvalues, and the corresponding eigenvector is informative about the k^{th} principal component to be inferred. A practical application of this computation is to serve as a guideline for how many principal components one should keep when PCA is used for dimensionality reduction. For instance, one can choose to keep only the eigenvalues that are larger than λ_+ , the bulk top eigenvalue in eqn (43).

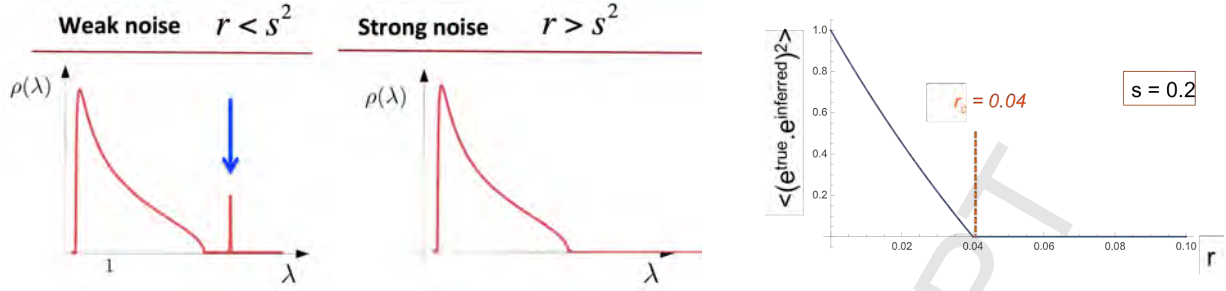


Figure 18: The retarded-learning phase transition. Left panels: spectrum of eigenvalues of the empirical correlation matrix in the principal-component model in the cases of weak noise (left, $r < s^2$) and of strong noise (right, $r > s^2$). Right panel: average squared overlap between the top components of the true and empirical correlation matrices as a function of the noise level r , for $s = 0.2$.

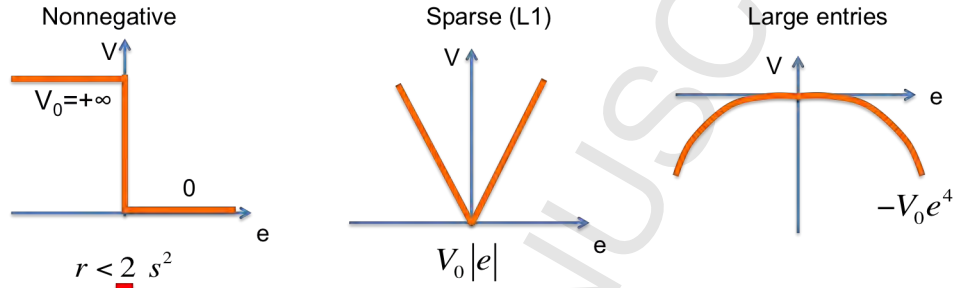


Figure 19: Prior potentials $V(e)$ used for learning top components. In the nonnegative case (left), the top component can be inferred when $r < 2s^2$ in the presence of the prior, which is twice bigger than the maximum-likelihood threshold, $r = s^2$. Other prior potentials include the L_1 regularization (middle), and a potential favoring large components (right), see [79].

5.5 Incorporating prior information

We have seen in the previous Section that, when $r < s^2$, it is possible to extract a vector with a finite scalar product with the top component $|e\rangle$. One natural question is whether exploiting prior information about the structure of the top component can help us increase this threshold, i.e. find out the top component with less data. We assume a prior distribution for the entries of the top component:

$$P(|e\rangle) = \prod_i P(e_i), \quad \text{with} \quad P(e_i) \propto \exp[-V(e_i)]. \quad (44)$$

Several expressions of interest can be considered for the potential V , with the representative curves shown in Fig. 19. We may for instance know that the top component has all its components e_i positive or zero.

$$V(e) = \begin{cases} +\infty & \text{if } e < 0 \\ 0 & \text{if } e \geq 0 \end{cases} \quad (45)$$

This can be useful in practice if we look for a collective excitatory mode, e.g. in gene expression data [76, 77]. At first sight, it seems trivial to find a vector with a positive dot product with $|e\rangle$, as a good candidate is $|v\rangle = \frac{1}{\sqrt{N}}(1, 1, \dots, 1)$. However, since $|e\rangle$ may be arbitrarily sparse (have all components equal to zero but a finite number), there is no guarantee that $\langle e|v\rangle$ is actually finite in the large N limit. It was recently shown that maximizing (41) under the condition $e_i \geq 0, \forall i$ leads to an estimate of the top component with a positive dot product as long as $r < 2s^2$ [78], see Fig. 19. Hence, incorporating prior information about the non-negativity of the entries of the top components allows for doubling the noise level. Other cases of interest are the L_1 regularization,

$$V(e) = V_0|e| \quad (46)$$

which favors sparsity. Recently, motivated by the study of covariation in protein families and the search for eigenvectors of the residue-residue correlation matrix with strong components on sites in contact on the 3D structure [80], we have considered the following potential [79, 81],

$$V(e) = -V_0e^4, \quad (47)$$

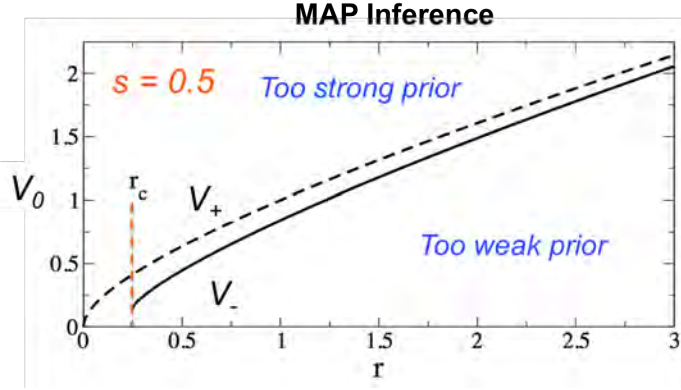


Figure 20: The PCA retarded-learning phase diagram using priors for large entries in eqn (47), for $s = 0.5$. At fixed r , inference of the top component is possible if the strength of the large-component prior, V_0 , is comprised between V_- and V_+ . For $V_0 < V_-$, the prior is too weak, and the situation is similar to maximum likelihood decoding. For $V_0 > V_+$, the prior is too strong: the inferred vector will have few large entries e_i , as required, but the sites i carrying these large entries will not match their counterparts in the true top component.

which favors large components. As the total normal of $|e\rangle$ is still fixed to unity, only a finite number of components can be large (and finite). Note that the cost of weak components is very small, hence this potential does not enforce any sparsity constraint, contrary to eqn (46). Instead of maximizing the likelihood, we now maximize the full a posteriori probability $P(|e\rangle|C) \propto P(C||e) \times P(|e\rangle)$. The optimization can not be performed analytically anymore, but the analysis of the typical properties of the solution can be analyzed with the replica method [82]. For the prior shown in eqn (47), small values of V_0 can reduce the learning lag, whereas too large values impeach learning, see Fig. 20.

5.6 Inverse Hopfield problem

PCA has also a strong connection with the inverse Ising problem of Section 3.2, when the coupling matrix J is constrained to have low rank, typically $\ll p$. This assumption may help avoid overfitting the data. We therefore write the interaction matrix as follows [80, 83],

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^k \xi_i^\mu \xi_j^\mu - \frac{1}{N} \sum_{\mu=1}^{\hat{k}} \hat{\xi}_i^\mu \hat{\xi}_j^\mu. \quad (48)$$

Here, k and \hat{k} are, respectively the numbers of positive and negative eigenvalues of J , and the total rank is $k + \hat{k}$. The interaction matrix in eqn (48), together with the Gibbs measure in eqn (13), define a generalized Hopfield model, made of the standard *attractive* patterns ξ^μ and of *repulsive* pattern $\hat{\xi}^\mu$. To make the meaning of these patterns more explicit, we rewrite the probability distribution (13) of this generalized Hopfield model:

$$P(\mathbf{s}|h, \xi, \hat{\xi}) = \frac{1}{Z(h, \xi, \hat{\xi})} \exp \left(\mathbf{h} \cdot \mathbf{s} + \frac{1}{2N} \sum_{\mu=1}^k (\xi^\mu \cdot \mathbf{s})^2 - \frac{1}{2N} \sum_{\mu=1}^{\hat{k}} (\hat{\xi}^\mu \cdot \mathbf{s})^2 \right), \quad (49)$$

where \cdot denotes the scalar product (summation over components i). The meaning of the patterns is transparent: they define favored, for attractive patterns, or disfavored, for repulsive patterns, directions in the space of configurations \mathbf{s} , along which the probability increases or decreases quadratically.

Attractive and repulsive patterns may be inferred through minimization of the cross-entropy defined in eqn (17). To the lowest order in ξ_i^μ/\sqrt{N} and $\hat{\xi}_i^\mu/\sqrt{N}$, one finds that the fields and patterns minimizing S are given by [83]:

$$\begin{aligned} h_i &= \log p_i \\ \xi_i^\mu &= \sqrt{1 - \frac{1}{\lambda^\mu}} \frac{v_i^\mu}{\sqrt{p_i(1-p_i)}} \quad (\mu = 1, \dots, k) \\ \hat{\xi}_i^\mu &= \sqrt{\frac{1}{\lambda^{N-\mu}} - 1} \frac{v_i^{N-\mu}}{\sqrt{p_i(1-p_i)}} \quad (\mu = 1, \dots, \hat{k}) \end{aligned} \quad (50)$$

where $\lambda^1 \geq \lambda^2 \geq \dots \geq 1 \geq \dots \geq \lambda^{N-1} \geq \lambda^N$ are the eigenvalues of the Pearson correlation matrix,

$$C_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i) p_j(1-p_j)}}, \quad (51)$$

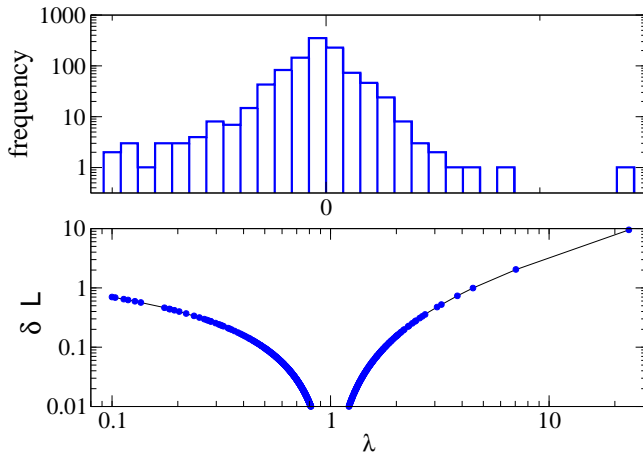


Figure 21: Top: Eigenvalue spectrum of the Pearson correlation matrix for the sequences of the trypsin inhibitor family (PF00014); the noise ratio is $r \simeq 0.5$, hence the edges of the Marcenko-Pastur spectrum are approximately $\lambda_- = 0.09$ and $\lambda_+ = 2.9$, see eqn (43). Bottom: pattern contributions to the log-likelihood of the inferred Hopfield model for the patterns corresponding to the eigenvalues along the x -axis. The most-contributing patterns are attractive patterns corresponding to the largest eigenvalues and repulsive patterns corresponding to the smallest eigenvalues. See [80] for more details.

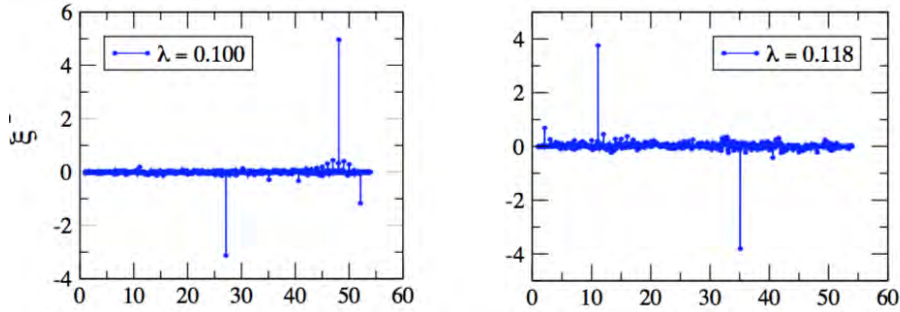


Figure 22: Two of the smallest-eigenvalue repulsive patterns obtained for the trypsin inhibitor family (PF00014). x index: $i + s/20$, where i is the site index and $s = 1 \dots 20$ is the amino acid (or gap) index. Each pattern is localized on essentially two components, corresponding to two sites i_1, i_2 in contact through a cysteine-cysteine bridge. Importantly, the two sites are close on the 3D fold but distant along the protein backbone. See [80] for more details.

and the \mathbf{v}^μ are the associated eigenvectors (with squared norms equal to N)⁶.

The above procedure is strongly reminiscent of PCA. Formula (50) shows, however, that the patterns do not coincide with the eigenvectors of C due to the presence of p_i -dependent terms. Furthermore, the presence of the λ^μ -dependent factor discounts the patterns corresponding to eigenvalues close to unity. This effect is easy to understand in the case of independent spins: in the limit of perfect sampling ($B \rightarrow \infty$), C coincides with the identity matrix, hence $\lambda^\mu = 1, \forall \mu$, and the patterns and the couplings vanish as they should. In the general case of coupled spins, the sum of the eigenvalues of C is equal to N (since $C_{ii} = 1, \forall i$). Therefore, the largest and smallest eigenvalues are guaranteed to be, respectively, above and below unity, and the corresponding attractive and repulsive patterns are real valued.

Inserting expression (50) in the cross-entropy (17), we obtain the contribution (per data configuration) of pattern μ to the log-likelihood,

$$\delta \mathcal{L}^\mu = \frac{1}{2} (\lambda^\mu - 1 - \log \lambda^\mu), \quad (53)$$

a quantity which is strictly positive for $\lambda^\mu \neq 1$, see Fig. 22. This expression helps select most relevant patterns, in decreasing order of their contributions $\delta \mathcal{L}^\mu$. This is in analogy with PCA when one selects the signal eigenvectors as the ones detaching most from the spectrum of the Marcenko-Pastur distribution, see Section 5.4. However, at difference with PCA, where only top eigenvalues (large λ^μ) and attractive

⁶Note that the Hopfield model is, by construction, invariant under global rotations in the pattern index space, e.g. any rotation \mathcal{O} of all the attractive patterns in the k -dimensional space:

$$\xi^\mu \rightarrow \sum_\nu \mathcal{O}^{\mu,\nu} \xi^\nu. \quad (52)$$

In other terms, the patterns are defined up to a rotation and are not unique; the gauge chosen in eqn (50) corresponds to orthogonal patterns in site space. Obviously, the couplings J_{ij} are gauge-invariant.

patterns are taken into account, selected patterns in the inverse Hopfield model are on both ends of the spectrum. Small eigenvalues, much below $\lambda = 1$, can give large contributions to the log-likelihood (Fig. 22). In applications to the study of amino-acid covariation in protein families, repulsive patterns can be shown to be localized on a small number of sites; they are much more informative about structural constraints in the protein, e.g. on the pairs of amino acids in contact, than attractive patterns [80].

6 Compositional Representations: Restricted Boltzmann Machines (RBM)

6.1 Definition and motivation

A Restricted Boltzmann Machine (RBM) is a graphical model, i.e., a probability distribution over a multidimensional data set, similar to the multivariate Gaussian distribution or the Boltzmann Machine distribution. It is constituted by two sets of random variables, a visible layer (\mathbf{v}) -the data layer- and a hidden layer (\mathbf{h}), which are coupled together, see Fig. 23. The joint probability distribution of the visible and hidden unit configurations, $\mathbf{v} = (v_1, v_2, \dots, v_N)$ and $\mathbf{h} = (h_1, h_2, \dots, h_M)$, is the Gibbs distribution

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})}, \quad (54)$$

defined by the energy

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^N \mathcal{U}_i(v_i) + \sum_{\mu=1}^M \mathcal{U}_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu} v_i h_\mu, \quad (55)$$

where the $\mathcal{U}_i, \mathcal{U}_\mu$ are unary potentials that control the marginal distributions of the variables v_i, h_μ , and the weight matrix $w_{i\mu}$ couples the visible and hidden layers. Depending on the choice of the potentials, the visible and hidden variables can be binary or continuous. The visible potentials \mathcal{U}_i is in general chosen based on the data we want to model; for example if $v \in [0, 1]$, then $\mathcal{U}_i(v_i) = -g_i v_i$, where the field g_i is a parameter of the model. The hidden potentials \mathcal{U}_μ can be chosen arbitrarily as long as sampling is feasible. Some useful examples are:

- The Bernoulli potential: $\mathcal{U}_\mu(h_\mu) = -g_\mu h_\mu$ with $h_\mu \in [0, 1]$;
- The Quadratic potential: $\mathcal{U}_\mu(h_\mu) = \frac{1}{2} h_\mu^2$, $h_\mu \in \mathbb{R}$;
- The ReLU potential: $\mathcal{U}_\mu(h_\mu) = \begin{cases} \frac{1}{2} h_\mu^2 + \theta_\mu h_\mu & \text{if } h_\mu \geq 0 \\ +\infty & \text{if } h_\mu < 0 \end{cases}$.

By marginalizing over the hidden units, one can compute the probability distribution over the visible layer:

$$P(\mathbf{v}) = \int \prod_{\mu} dh_{\mu} P(\mathbf{v}, \mathbf{h}) \equiv \frac{1}{Z_{eff}} \exp[-E_{eff}(\mathbf{v})] \quad (56)$$

The marginal distribution in eqn (56) can be expressed analytically in terms of the weight matrix and the potentials. Training an RBM consists in fitting this marginal distribution to the data by maximum likelihood. Unlike multivariate Gaussian distributions studied in the previous Section, the optimal RBM must be found numerically, e.g. using approximate stochastic gradient ascent over the likelihood [84].

We stress that, in contradistinction with Boltzmann Machines (Ising models) or multivariate Gaussian distributions, there are no direct couplings between pairs of units in the same layer (hence, the name

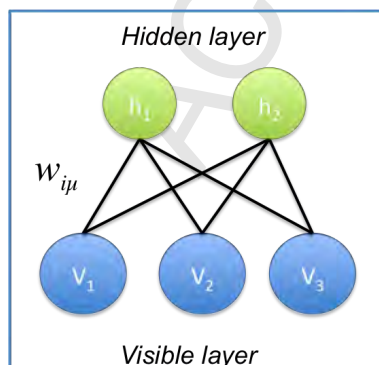


Figure 23: Architecture of Restricted Boltzmann Machines. A RBM is defined on a bidirectional bipartite graph, with a visible (\mathbf{v}) layer that represents the data, connected to a hidden (\mathbf{h}) layer supposed to extract statistically meaningful features from the data and, in turn, to condition their distribution. There are N visible units indexed by i , and M hidden units indexed by μ . The connections between visible and hidden units are denoted by $w_{i\mu}$.

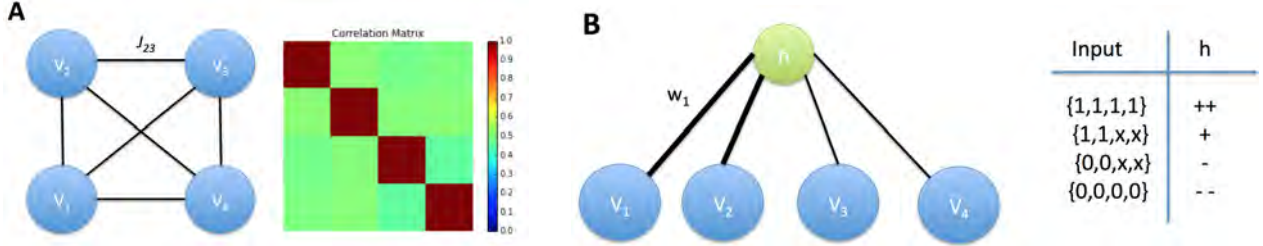


Figure 24: How to model correlations among a set of variables. **A.** Boltzmann Machine approach: The matrix of pairwise correlations between variables is computed from data, and a network of couplings is inferred to reproduce those correlations. **B.** Restricted Boltzmann Machine approach: observed correlations are due to one or more common input(s), whose values drive the configurations of the variables. A network of connection between the visible layer (support of data configurations) and a layer of hidden units (support of common inputs) is found to maximize the probability of the data items. The rightmost column indicates the magnitude h of the hidden unit as a function of the visible configuration.

restricted). RBM can nonetheless model correlations between visible variables, as the latter can be indirectly correlated through the hidden layer. Informally speaking, instead of explaining the correlations between several visible units through a set of adequate couplings, we interpretate them as collective variation driven by the common inputs (the hidden units) shared by these visible units, see Fig. 24. The hidden units thus represent collective modes of variation of the data.

Another way to state this is to observe that, as one marginalizes over the hidden layer, effective couplings between visible layer units arise. For instance, it is easy to show that for Gaussian hidden units, i.e. for the Quadratic potential \mathcal{U}_μ in the list above, the marginal distribution over the visible layer is:

$$E_{eff}(\mathbf{v}) = - \sum_i g_i v_i + \frac{1}{2} \sum_\mu \left(\sum_i w_{i\mu} v_i \right)^2 \quad (57)$$

In that case, we recognize a pairwise effective Hamiltonian, the Hopfield model with M patterns [2, 85]. In general, non-quadratic hidden-unit potentials generate effective Hamiltonian for the visible units with high-order interactions. The presence of couplings to all orders produced from a unique set of $N \times M$ connections $w_{i\mu}$ has deep effects on the sampling dynamics of RBM.

6.2 Sampling

The connection with data representation algorithms is best seen when considering the sampling scheme. Since there are no connections within a layer, the hidden layer units are conditionally independent given the configuration of the visible layer, and conversely; hence the following Gibbs sampling procedure, schematized in Fig. 25:

- Compute hidden units inputs $I_\mu^H = \sum_i w_{i\mu} v_i$
- Sample each hidden unit independently $P(h_\mu | I_\mu^H) \propto \exp[-\mathcal{U}_\mu(h_\mu) + h_\mu I_\mu^H]$
- Compute the visible layer inputs $I_i^V = \sum_\mu w_{i\mu} h_\mu$
- Sample each visible unit independently $P(v_i | I_i^V) \propto \exp[(g_i + I_i^V)v_i]$

The first two steps can be seen as a stochastic feature extraction from configuration \mathbf{v} , whereas the last two steps are a stochastic reconstruction of \mathbf{v} from the features \mathbf{h} . One can define in particular a data representation as the most likely hidden layer configuration given a visible layer configuration, that is, through the set of

$$h_\mu^*(\mathbf{v}) = \arg \max P(h_\mu | \mathbf{v}) = \arg \max P(h_\mu | \mathbf{v}) = \Phi_\mu(I_\mu^H(\mathbf{v})) , \quad (58)$$

where $\Phi_\mu = (\mathcal{U}'_\mu)^{-1}$ is the transfer function, see Fig. 26.

6.3 Phenomenology of RBM trained on data

Once maximum-likelihood training is completed, RBM can be very good generative models for complex, multimodal distributions. In the following, we describe the phenomenology of RBM, with various kind of hidden-unit potentials, trained on MNIST, a dataset of 60,000 28×28 grayscale images of handwritten digits. Each image can be flattened, binarized by thresholding the grayscale level, into a 784-dimensional binary vector. The following observations can be made:

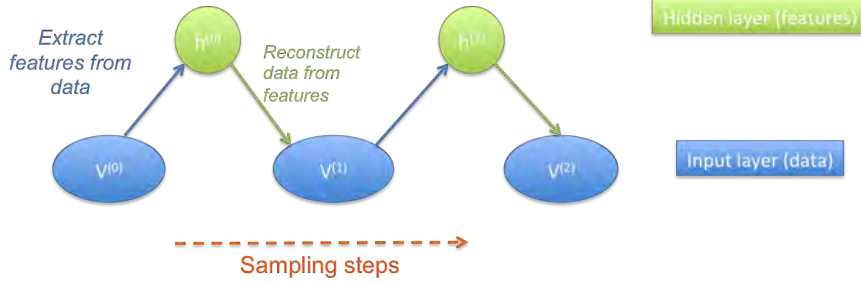


Figure 25: Back-and-forth sampling procedure in RBM. Hidden configurations \mathbf{h} are sampled from visible configurations \mathbf{v} , and, in turn, define the distribution of visible configurations at the next sampling step.

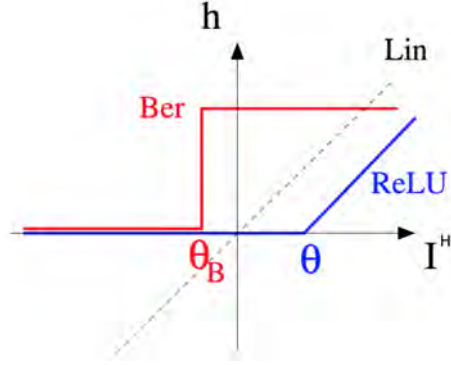


Figure 26: Transfer functions Φ for various hidden units potentials \mathcal{U}_μ , see main text. The transfer function gives the most likely value of the hidden unit, h^* , as a function of the input I^H received from the visible units.

- After training, samples drawn from the equilibrium distribution of Bernoulli or ReLU RBM look like real digits, suggesting that it is a good fit for the data distribution, see Fig. 27(b). On the contrary, Gaussian RBM, *i.e.* pairwise Hamiltonians, do not fit the data distribution as well.
- Each hidden unit is activated selectively by the presence of a specific feature of the data: this is seen by visualizing the columns of the weight matrix $w_{i\mu}$, see Fig. 27(a). The features are strokes, that is, small part of digits. The weight matrix is therefore essentially sparse, with a fraction of nonzero weights $p \sim 0.1$, see Fig. 27(d).
- For ReLU hidden units, each data image strongly activates around ~ 20 hidden units, whereas most hidden units are silent or weakly activated, see Fig. 27(c). For a precise definition of how the number strongly hidden units is estimated, see eqn (62) and [86].
- The learnt probability distribution is very rough, with many local maxima of probability (much larger than the values of N or M), as seen in Fig. 27(e). Remarkably, after training, each data sample is within few pixels of a local maximum of probability. This shows the combinatorial nature of RBM, capable of generating a very large number of configurations after training.

This phenomenology raises several questions. First, how can such simple networks generate a complex distribution with a large variety of local minima, matching the original data points? Secondly, why do some hidden unit potentials give good results, whereas others do not? Lastly, can we connect this behavior to the one of the Hopfield model, corresponding to the case of quadratic potential?

6.4 Statistical mechanics of RBM

It is hopeless to provide answers to these questions in full generality for a given RBM with parameters fitted from real data. However, statistical physics methods and concepts allow us to study the typical energy landscape and properties of RBM drawn from appropriate random ensembles. We follow this approach hereafter, using the replica method [86]. We define the Random-RBM ensemble model for ReLU hidden units as follows, see also drawing in Fig. 28,

- N binary visible units, M ReLU hidden units, with $N, M \rightarrow \infty$ and $\alpha = \frac{M}{N}$ is finite.
- uniform visible layer fields, *i.e.* $g_i = g, \forall i$.
- uniform hidden layer thresholds, *i.e.* $\theta_\mu = \theta, \forall \mu$.
- a random weight matrix $w_{i\mu} = \frac{\xi_{i\mu}}{\sqrt{N}}$, where each 'pattern' $\xi_{i\mu}$ is drawn independently, taking values $+1, -1$ with probabilities $\frac{p}{2}$ and 0 with probability $1 - p$. The *degree of sparsity* p is the fraction of non-zero weights.

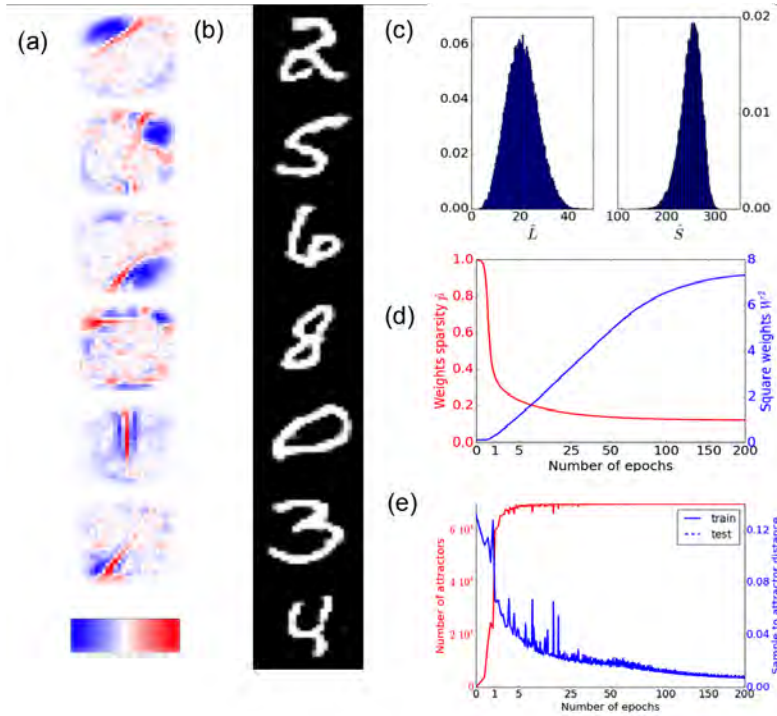


Figure 27: Training of RBM on MNIST. Data are composed of $N = 28 \times 28$ binarized images, and the RBM includes $M = 400$ hidden ReLU. (a) Set of weights \mathbf{w}_μ attached to four representative hidden units μ . (b) Averages of \mathbf{v} conditioned to five hidden-unit configurations \mathbf{h} sampled from the RBM at equilibrium. Black and white pixels correspond respectively to averages equal to 0 and 1; few intermediary values, indicated by grey levels, can be seen on the edges of digits. (c) Distributions of the numbers of very strongly activated hidden units, \hat{L} (left), and of silent hidden units, \hat{S} (right), at equilibrium. (d) Evolution of the weight sparsity \hat{p} (red) and the squared weight value W_2 (blue). The training time is measured in epochs (number of passes over the data set), and represented on a square-root scale. (e) Evolution of the number of distinct local maxima of $P(\mathbf{v})$ in eqn (56) (left scale) and distance to the original sample (right scale, for training and test sets). For each sample, the local minimum is obtained through $T = 0$ -sampling of the RBM, see Section 6.2.

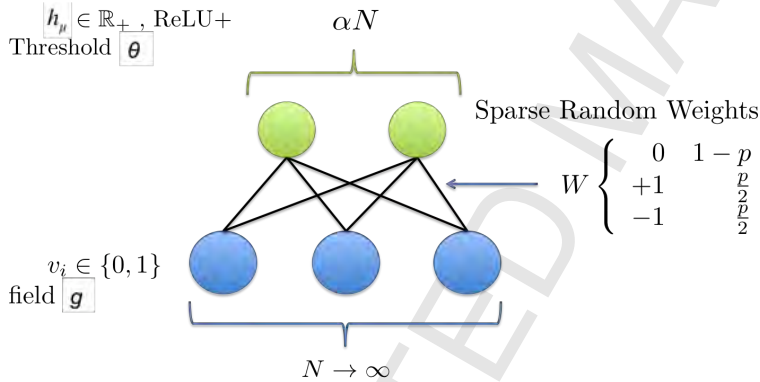


Figure 28: The Random-RBM ensemble, with its control parameters: threshold θ of hidden ReLU, ratio α of the sizes of the hidden and visible layers, field g on visible units, sparsity p of the weights (rescaled by $W = 1/\sqrt{N}$).

Hence, α , p , g and θ are the control parameters of our model. Several variants or special cases have already been addressed in the literature. Choosing Gaussian hidden units and ± 1 visible units leads back to the original Hopfield model, studied in [3]. The sparse weight distribution was previously introduced to study parallel storage of multiple sparse items in the Hopfield model [87, 88].

It is important to understand the magnitude of hidden-unit activations for a given a visible layer configuration \mathbf{v} . Two cases are encountered:

- let us call L the number of hidden units μ coding for features \mathbf{w}_μ present in \mathbf{v} . These hidden units will be strongly activated, as their inputs $I_\mu^H = \mathbf{w}_\mu \cdot \mathbf{v}$ will be strong and positive, comparable to the product of the norms of \mathbf{w}_μ , of the order of $\simeq \sqrt{p}$ for large N , and \mathbf{v} , of the order of \sqrt{pN} . Therefore, we expect I_μ^H to scale as $m\sqrt{N}$, where the prefactor m , called magnetization, is finite in the thermodynamical limit.
- The remaining $M - L$ hidden units μ' have, however, features $\mathbf{w}_{\mu'}$ essentially orthogonal to \mathbf{v} . Hence, the vast majority of hidden units receive random inputs $I_{\mu'}^H$ fluctuating around zero, with finite variances.

To answer the questions raised in Section 6.3, we are interested in computing the averages of m , L over the distribution and over the random weights. They can be obtained through a replica computation of the average free energy,

$$f(\alpha, p, g, \theta) \equiv \lim_{N \rightarrow \infty} -\frac{1}{\beta N} \overline{\log Z(\alpha, \beta, p, g, \theta, \{\xi_{i\mu}\})}, \quad (59)$$

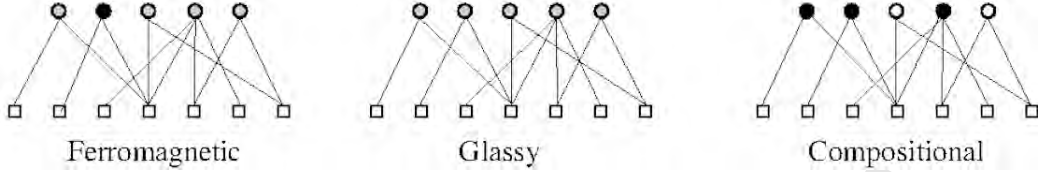


Figure 29: The three regimes of operation of Random RBM, see text. Black, grey and white hidden units symbolize, respectively, strong ($h \sim \sqrt{N}$), weak ($h \sim \pm 1$) and null ($h = 0$) activations.

where the overbar denotes the average over the $\{\xi_{i\mu}\}$ and the partition function reads

$$Z(\alpha, \beta, p, g, \theta, \{\xi_{i\mu}\}) = \sum_{\mathbf{v} \in \{0,1\}^N} \int \prod_{\mu=1}^M dh_{\mu} e^{-\beta E(\mathbf{v}, \mathbf{h})}. \quad (60)$$

After some algebra, see [86], we find that $f(\alpha, p, g, \theta)$ is obtained through optimizing a free-energy functional over the order parameters L, m, q, r, B, C :

- m and L are, respectively, the magnetization and the number of feature-encoding hidden units,
- r is the mean squared activity of the other hidden units,
- $q = \frac{1}{N} \sum_i \overline{v_i}$ is the mean activity of the visible layer in the GS,
- B, C are response functions, *i.e.* derivatives of the mean activity of, respectively, hidden and visible units with respect to their inputs.

For non-sparse weights ($p = 1$) and depending on the values of the other control parameters, the system can show one of two following qualitative different behaviors, as is found for the Hopfield model [3] and Fig. 29:

- A ferromagnetic phase, in which hidden configurations with $L = 1$ and $m > 0$ dominate. Visible configurations have strong overlap with one feature, say, $\mu = 1$. It is likely that $v_i = 1$ if $\xi_{i,1} = 1$ and $v_i = 0$ if $\xi_{i,1} = -1$. As the choice of μ is arbitrary, there are αN such 'basins' of visible configurations. Phases with $L > 1$, *i.e.* having strong overlap with several features exist and may be thermodynamically stable, but are unfavorable: their free energies increase with L .
- A spin-glass phase, in which configurations with $m = 0$ dominate. Most configurations have weak normalized overlap $\sim \frac{1}{\sqrt{N}}$ with all hidden-unit weight vectors.

The phase transition occurs out of the frustration in the system. Assume for instance that the system is in the ferromagnetic phase. The input received by a visible unit, say, i has a strong contribution (of the order of 1 as $N \rightarrow \infty$) from the strongly magnetized unit, say, $\mu = 1$, and a lot (of the order of αN) of weak inputs (of the order of $\pm 1/\sqrt{N}$) from the other hidden units. As the ratio α increases, these numerous, weak noisy contributions win over the unique, strong signal contribution, and the systems enters the glassy phase. The transition takes place at a well defined value of α , which depends on θ and g [86].

For small p , a new intermediate qualitative behavior emerges:

- The compositional phase, in which visible configurations have strong overlap with L features, where $1 \ll L \ll M$, see Fig. 29. As observed for RBM trained on real data in Fig. 27(e), random RBM may generate a combinatorial diversity of low-energy visible configurations, corresponding to different choices of the subset $\{\mu_1, \dots, \mu_L\}$ of strongly activated hidden units. This new phase is found in the low p limit, and for appropriate values of the threshold θ (large enough to silence a large number of hidden units and suppress interference, see Fig. 27(c)), and of the field g (to reproduce the average activity of the data in the visible layer).

6.5 Validation on data

One of the outcomes of our statistical physics analysis is that, in the compositional phase, the number L of strongly activated hidden units scales as the inverse of the degree of sparsity, p . More precisely, $L \sim \frac{\ell^*}{p}$, when $p \rightarrow 0$, where ℓ^* is determined by minimizing the free energy of the Random-RBM model. The minimum ℓ^* of the free energy is found at $\ell^* > 0$ in the compositional phase, contrary to the ferromagnetic phase, where $\ell^* = 0$.

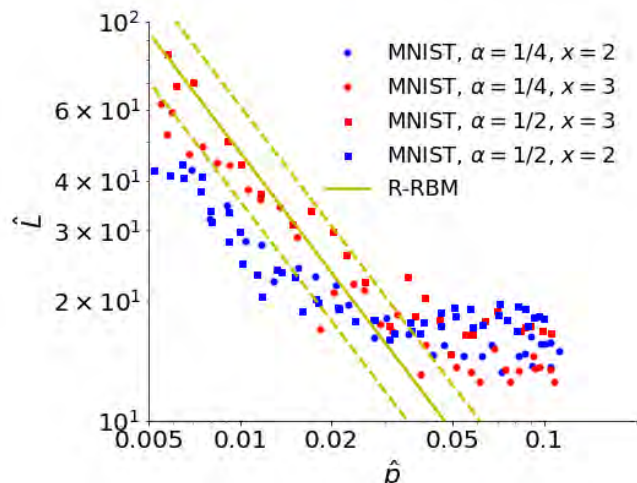


Figure 30: Average number L of active hidden units vs. degree p of sparsity of the weights, for RBM trained on MNIST data. Values of the ratio α and of the exponent x in the regularization term in eqn (61) are reported in the figure. The figure also shows the theoretical curve obtained for Random RBM model of Fig. 28; Dashed lines show one standard deviations away from the mean value of L due to finite-size fluctuations, see [86] for details. Results for a more realistic statistical ensemble of random weights, in which the degree of sparsity p fluctuates with the visible sites i , can be found in [86].

This prediction can be tested in RBM trained on real data, e.g. MNIST, see Fig. 27(b). The value of p at the end of training without any regularization was found to be ~ 0.1 , see Fig. 27(d). However, higher sparsities, i.e. lower values of p , can be imposed through regularization of the weights. To do so, we add to the log-likelihood the penalty term

$$C(\{w_{i\mu}\}) = - \sum_{\mu} \left(\sum_i |w_{i\mu}| \right)^x, \quad (61)$$

where $x \geq 0$. The case $x = 1$ gives standard L_1 regularization, while, for $x > 1$, the effective penalty strength, $\propto \left(\sum_i |w_{i\mu}| \right)^{x-1}$, increases with the weights, hence promoting homogeneity among hidden units. After training we generate Monte Carlo samples of each RBM at equilibrium, and monitor the average number of active hidden units, L , estimated through the participation ratio

$$L = \frac{(\sum_{\mu} h_{\mu}^2)^2}{\sum_{\mu} h_{\mu}^4}. \quad (62)$$

By changing the value of x , we obtain, at the end of training, RBM with higher sparsities. Figure 30 shows that the theoretical scaling law $L \sim \ell^*/p$ is well reproduced over one decade of variation of p . In addition, the product $L \times p$ is in good agreement with the theoretical prediction ℓ^* [86].

Acknowledgements. This work benefited from the financial support of the Human Frontier Science Program through the RGP0057/2016 project.

References

- [1] S. Kirkpatrick, C. J. Gelatt, and M. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, pp. 671–80, 1983.
- [2] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.
- [3] D. J. Amit, H. Gutfreund, and H. Sompolinsky, “Statistical mechanics of neural networks near saturation,” *Annals of Physics*, vol. 173, no. 1, pp. 30–67, 1987.
- [4] F. Rieke, D. Warland, R. De Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the neural code*. The MIT press, Cambridge, USA, 1997.
- [5] H. Sompolinsky, “Computational neuroscience: beyond the local circuit,” *Current opinion in neurobiology*, vol. 25, pp. xiii–xviii, 2014.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, may 2015.
- [7] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013.

- [8] P. Coveney, E. Dougherty, and R. Highfield, “Big data need big theory too,” *Phil. Trans. R. Soc. A*, vol. 374, p. 20160153, 2016.
- [9] T. Poggio and F. Anselmi, *Visual cortex and deep networks: learning invariant representations*. MIT Press, Cambridge, 2016.
- [10] S. Ganguli and H. Sompolinsky, “Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis,” *Annu. Rev. Neurosci.*, vol. 35, pp. 485–508, 2012.
- [11] D. O. Hebb, *The organization of behavior; a neuropsychological theory*. John Wiley & Sons, 1949.
- [12] D. J. Amit, H. Gutfreund, and H. Sompolinsky, “Storing infinite numbers of patterns in a spin-glass model of neural networks,” *Physical Review Letters*, vol. 55, no. 14, pp. 1530–1533, 1985.
- [13] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond*, vol. 9. World Scientific, 1987.
- [14] D. J. Amit, *Modeling brain function. The world of attractor neural networks*. Cambridge University Press, Cambridge, 1989.
- [15] E. Moser, E. Kripff, and M. Moser, “Place cells, grid cells, and the brain’s spatial representation system,” *Annu. Rev. Neurosci.*, vol. 31, pp. 69–89, 2008.
- [16] J. O’Keefe and J. Dostrovsky, “The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat.,” *Brain Res*, vol. 34, no. 1, pp. 171–175, 1971.
- [17] K. Jezek *et al.*, “Theta-paced flickering between place-cell maps in the hippocampus,” *Nature*, vol. 478, no. 7368, pp. 246–249, 2011.
- [18] A. Treves and E. T. Rolls, “Computational analysis of the role of the hippocampus in memory,” *Hippocampus*, vol. 4, no. 3, pp. 374–391, 1994.
- [19] M. Tsodyks, “Attractor neural network models of spatial maps in hippocampus,” *Hippocampus*, vol. 9, no. 4, pp. 481–489, 1999.
- [20] D. M. Smith and S. J. Mizumori, “Hippocampal place cells, context, and episodic memory,” *Hippocampus*, vol. 16, no. 9, pp. 716–729, 2006.
- [21] F. P. Battaglia and A. Treves, “Attractor neural networks storing multiple space representations: a model for hippocampal place fields,” *Physical Review E*, vol. 58, no. 6, p. 7738, 1998.
- [22] A. Samsonovich and B. L. McNaughton, “Path integration and cognitive mapping in a continuous attractor neural network model,” *The Journal of Neuroscience*, vol. 17, no. 15, pp. 5900–5920, 1997.
- [23] R. Monasson and S. Rosay, “Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram,” *Phys. Rev. E*, vol. 87, p. 062813, 2013.
- [24] J. L. Lebowitz and O. Penrose, “Rigorous treatment of the van der waals-maxwell theory of the liquid-vapor transition,” *Journal of Mathematical Physics*, vol. 7, no. 1, pp. 98–113, 1966.
- [25] S. F. Edwards and P. W. Anderson, “Theory of spin glasses,” *Journal of Physics F: Metal Physics*, vol. 5, no. 5, p. 965, 1975.
- [26] J. R. L. De Almeida and D. J. Thouless, “Stability of the sherrington-kirkpatrick solution of a spin glass model,” *Journal of Physics A: Mathematical and General*, vol. 11, no. 5, p. 983, 1978.
- [27] S. Rosay, “A statistical mechanics approach to the modelling and analysis of place-cell activity,” *PhD thesis*, 2014.
- [28] R. Monasson and S. Rosay, “Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Collective motion of the activity,” *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 89, no. 3-1, pp. 032803–032803, 2014.
- [29] R. Monasson and S. Rosay, “Transitions between spatial attractors in place-cell models,” *Physical review letters*, vol. 115, no. 9, p. 098101, 2015.
- [30] J. Langer, “Statistical theory of the decay of metastable states,” *Annals of Physics*, vol. 54, pp. 258–75, 1969.

- [31] S. Cocco, R. Monasson, L. Posani, and G. Tavoni, “Functional networks from inverse modeling of neural population activity,” *Current Opinion in Systems Biology*, vol. 3, pp. 103–110, 2017.
- [32] L. Posani, S. Cocco, K. Jezek, and R. Monasson, “Functional connectivity models for decoding of spatial representations from hippocampal ca1 recordings,” *Journal of Computational Neuroscience*, vol. 43, pp. 17–33, Aug 2017.
- [33] E. T. Jaynes, “Information theory and statistical mechanics,” *Physical review*, vol. 106, no. 4, p. 620, 1957.
- [34] M. J. Wainwright, M. I. Jordan, *et al.*, “Graphical models, exponential families, and variational inference,” *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.
- [35] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- [36] G. Tavoni, U. Ferrari, F. P. Battaglia, S. Cocco, and R. Monasson, “Functional coupling networks inferred from prefrontal cortex activity show experience-related effective plasticity,” *Network Neuroscience*, p. doi: 10.1162/NETN_a_00014, 2017.
- [37] G. Tavoni, S. Cocco, and R. Monasson, “Neural assemblies revealed by inferred connectivity-based models of prefrontal cortex recordings,” *Journal of Computational Neuroscience*, vol. 41, pp. 269–293, 2016.
- [38] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, “Weak pairwise correlations imply strongly correlated network states in a neural population,” *Nature*, vol. 440, no. 7087, pp. 1007–1012, 2006.
- [39] I. H. Stevenson, J. M. Rebesco, L. E. Miller, and K. P. Körding, “Inferring functional connections between neurons,” *Current opinion in neurobiology*, vol. 18, no. 6, pp. 582–588, 2008.
- [40] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, “A learning algorithm for boltzmann machines,” *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.
- [41] E. Aurell and M. Ekeberg, “Inverse ising inference using all the data,” *Physical Review Letters*, vol. 108, no. 9, p. 090201, 2012.
- [42] S. Cocco and R. Monasson, “Adaptive cluster expansion for inferring boltzmann machines with noisy data,” *Physical review letters*, vol. 106, no. 9, p. 090601, 2011.
- [43] B. L. McNaughton, J. O’Keefe, and C. A. Barnes, “The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records,” *Journal of neuroscience methods*, vol. 8, no. 4, pp. 391–397, 1983.
- [44] M. Meister, J. Pine, and D. A. Baylor, “Multi-neuronal signals from the retina: acquisition and analysis,” *Journal of neuroscience methods*, vol. 51, no. 1, pp. 95–106, 1994.
- [45] A. Peyrache, M. Khamassi, K. Benchenane, S. I. Wiener, and F. P. Battaglia, “Replay of rule-learning related neural patterns in the prefrontal cortex during sleep,” *Nature neuroscience*, vol. 12, no. 7, pp. 919–926, 2009.
- [46] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.
- [47] R. D. Finn, P. Coghill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, *et al.*, “The pfam protein families database: towards a more sustainable future,” *Nucleic acids research*, vol. 44, no. D1, pp. D279–D285, 2016.
- [48] D. De Juan, F. Pazos, and A. Valencia, “Emerging methods in protein co-evolution,” *Nature Reviews Genetics*, vol. 14, no. 4, pp. 249–261, 2013.
- [49] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, “Direct-coupling analysis of residue coevolution captures native contacts across many protein families,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.
- [50] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead, “Learning generative models for protein fold families,” *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 4, pp. 1061–1078, 2011.

- [51] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, “New method for parameter estimation in probabilistic models: minimum probability flow,” *Physical review letters*, vol. 107, no. 22, p. 220601, 2011.
- [52] M. Opper and D. Saad, *Advanced Mean Field Methods. Theory and Practice*, MIT Press, Feb. 2001.
- [53] V. Sessak and R. Monasson, “Small-correlation expansions for the inverse ising problem,” *Journal of Physics A: Mathematical and Theoretical*, vol. 42, no. 5, p. 055001, 2009.
- [54] S. Cocco and R. Monasson, “Adaptive cluster expansion for the inverse ising problem: convergence, algorithm and tests,” *J. Stat. Phys.*, vol. 147, no. 2, pp. 252–314, 2012.
- [55] K. J. Friston, “Functional and effective connectivity: a review,” *Brain connectivity*, vol. 1, no. 1, pp. 13–36, 2011.
- [56] R. Monasson and S. Rosay, “Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram,” *Physical Review E*, vol. 87, no. 6, p. 062813, 2013.
- [57] R. Monasson and S. Rosay, “Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Collective motion of the activity,” *Physical Review E*, vol. 89, no. 3, p. 032803, 2014.
- [58] M. Tsodyks and T. Sejnowski, “Associative memory and hippocampal place cells,” *International journal of neural systems*, vol. 6, pp. 81–86, 1995.
- [59] J. O’Keefe and J. Dostrovsky, “The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat,” *Brain research*, vol. 34, no. 1, pp. 171–175, 1971.
- [60] J. O’keefe and D. Conway, “Hippocampal place units in the freely moving rat: why they fire where they fire,” *Experimental Brain Research*, vol. 31, no. 4, pp. 573–590, 1978.
- [61] M. Fyhn, T. Hafting, A. Treves, M.-B. Moser, and E. I. Moser, “Hippocampal remapping and grid realignment in entorhinal cortex,” *Nature*, vol. 446, no. 7132, p. 190, 2007.
- [62] L. Posani, S. Cocco, K. Jezek, and R. Monasson, “Position is coherently represented during flickering instabilities of place-cell cognitive maps in the hippocampus,” *26th Annual Computational Neuroscience Meeting (CNS*2017), BMC Neuroscience 2017, 18(Suppl 1):O21*, 2017.
- [63] S. Mallat, “Understanding deep convolutional networks,” *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, p. 20150203, 2016.
- [64] Y. LeCun, “L’apprentissage profond.” Lectures at Collège de France, 2016.
- [65] S. Mueller Stein, A. Loccisano, S. Firestine, and J. Evanseck, “Principal components analysis: A review of its application on molecular dynamics data,” *Annual Reports in Computational Chemistry*, vol. 2, pp. 233–66, 2006.
- [66] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR’91., IEEE Computer Society Conference on*, pp. 586–591, IEEE, 1991.
- [67] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, vol. 46. John Wiley & Sons, 2004.
- [68] A. Ng, “Sparse autoencoder,” *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.
- [69] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, p. 607, 1996.
- [70] D. L. Ringach, R. M. Shapley, and M. J. Hawken, “Orientation selectivity in macaque v1: diversity and laminar dependence,” *Journal of Neuroscience*, vol. 22, no. 13, pp. 5639–5651, 2002.
- [71] J. Zylberberg, J. T. Murphy, and M. R. DeWeese, “A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields,” *PLoS computational biology*, vol. 7, no. 10, p. e1002250, 2011.
- [72] V. A. Marcenko and L. A. Pastur, “Distribution of eigenvalues for some sets of random matrices,” *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.

- [73] J. Baik, G. B. Arous, S. Péché, *et al.*, “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,” *The Annals of Probability*, vol. 33, no. 5, pp. 1643–1697, 2005.
- [74] D. Hoyle and M. Rattay, “Principal-component-analysis eigenvalue spectra from data with symmetry breaking structure,” *Physical Review E*, vol. 69, p. 026214, 2004.
- [75] T. Watkin and J.-P. Nadal, “Optimal unsupervised learning,” *Journal of Physics A: Mathematical and General*, vol. 27, no. 6, p. 1899, 1994.
- [76] L. Badea, D. Tilivea, L. Badea, and D. Tilivea, “Sparse factorizations of gene expression data guided by binding data,” in *Pacific Symposium of Biocomputing*, 2005.
- [77] R. Zass and A. Shashua, “Nonnegative sparse pca,” in *Advances in Neural Information Processing Systems*, pp. 1561–1568, 2007.
- [78] E. Richard and A. Montanari, “Non-negative principal component analysis: Message passing algorithms and sharp asymptotics,” *IEEE Transactions on Information Theory*, pp. 1458 – 1484, 2016.
- [79] R. Monasson and D. Villamaina, “Estimating the principal components of correlation matrices from all their empirical eigenvectors,” *EPL (Europhysics Letters)*, vol. 112, no. 5, p. 50001, 2015.
- [80] S. Cocco, R. Monasson, and M. Weigt, “From principal component to direct coupling analysis of co-evolution in proteins: Low-eigenvalue modes are needed for structure prediction,” *PLoS computational biology*, vol. 9, no. 8, p. e1003176, 2013.
- [81] R. Monasson, “Inference of principal components of noisy correlation matrices with prior information,” in *50th Asilomar Conference on Signals, Systems and Computers*, pp. 95–99, IEEE, 2016.
- [82] M. Mézard, G. Parisi, and M. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, vol. 9. World Scientific Publishing Co Inc, 1987.
- [83] S. Cocco, R. Monasson, and V. Sessak, “High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections,” *Physical Review E*, vol. 83, no. 5, p. 051123, 2011.
- [84] A. Fischer and C. Igel, “Training restricted boltzmann machines: An introduction,” *Pattern Recognition*, vol. 47, no. 1, pp. 25–39, 2014.
- [85] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, “On the equivalence of hopfield networks and boltzmann machines,” *Neural Networks*, vol. 34, pp. 1–9, 2012.
- [86] J. Tubiana and R. Monasson, “Emergence of compositional representations in restricted boltzmann machines,” *Physical Review Letters*, vol. 118, no. 13, p. 138301, 2017.
- [87] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, and F. Moauro, “Multitasking associative networks,” *Physical review letters*, vol. 109, no. 26, p. 268101, 2012.
- [88] E. Agliari, A. Annibale, A. Barra, A. Coolen, and D. Tantari, “Immune networks: multitasking capabilities near saturation,” *Journal of Physics A: Mathematical and Theoretical*, vol. 46, no. 41, p. 415003, 2013.