Reducing a cortical network to a Potts model yields storage capacity estimates

Reducing a cortical network to a Potts model yields storage capacity estimates

Michelangelo Naim[1,2,*,¶], Vezha Boboeva[1,*], Chol Jun Kang[1,3,4], and Alessandro Treves[1,5]

[1]SISSA - International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, Italy

[2]La Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italy

[3]The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy

[4]Current address: Institute for Theoretical Physics, Department of Physics, Kim Il Sung University, Pyongyang, DPRK

[5]Kavli Institute for Systems Neuroscience/Centre for Neural Computation, Norwegian University of Science and Technology, Trondheim, Norway

Email: ale@sissa.it

*: These authors contributed equally to the work

¶: Current address: Weizmann Institute, Rehovot, Israel

Abstract

An autoassociative network of Potts units, coupled via tensor connections, has been proposed and analysed as an effective model of an extensive cortical network with distinct short- and long-range synaptic connections, but it has not been clarified in what sense it can be regarded as an effective model. We draw here the correspondence between the two, which indicates the need to introduce a local feedback term in the reduced model, i.e., in the Potts network. An effective model allows the study of phase transitions. As an example, we study the storage capacity of the Potts network with this additional term, the local feedback w, which contributes to drive the activity of the network towards one of the stored patterns. The storage capacity calculation, performed using replica tools, is limited to fully connected networks, for which a Hamiltonian can be defined. To extend the results to the case of intermediate partial connectivity, we also derive the self-consistent signal-to-noise analysis for the Potts network; and finally we discuss implications for semantic memory in humans.

Keywords: neural network, multi-modular network, Potts model, storage capacity

Contents

1 Introduction

Considerable research efforts in recent years have been driven by the ambition to reconstruct and simulate in microscopic detail the structure of the human brain, possibly at the 1:1 scale, with outcomes that have been questioned [1]. A complementary perspective is that put forward by the late neuroanatomist Valentino von Braitenberg, who in many publications argued for the need to understand overarching principles of mammalian brain organisation, even by recourse to dramatic simplification [2]. In this spirit, over 40 years ago Braitenberg proposed the notion of the *skeleton* cortex, that is comprised solely of its $\mathcal{N}$ pyramidal cells [3]. On their apical dendrites they receive predominantly synapses from axons that originate in the pyramidal cells of other cortical areas and travel through the white matter, while on their basal dendrites they receive mainly synapses from local axon collaterals, and the two systems, A(pical) and B(asal), can be estimated to include similar numbers of synapses C_A and C_B per receiving cell. Braitenberg then detailed what could have later been called a *small world* scheme [4]. In such a scheme, the $\mathcal{N}$ pyramidal cells are allocated to $N = \sqrt{\mathcal{N}}$ modules, each including N cells, fully connected with each other – so that $C_B = N - 1$. Each cell would further receive, on the A system, $N - 1$ connections from one cell drawn at random in each of the other modules, so that also $C_A = N - 1$. Therefore each cell gets $2(N - 1)$ connections from other pyramidal cells, the A and B systems are perfectly balanced, and the average minimal path length between any cell pair is just below 2. Of course, the modules are largely a fictional construct, apart from special cases, or at least their generality and character are quite controversial [5], [6], [7], but the distinction between long-range and local connections is real, and the simple model recapitulates a rough square-root scaling of both systems, with $N \sim 10^3 \div 10^5$, in skeleton cortices which in mammals range from ca. $\mathcal{N} \sim 10^6$ to ca. $\mathcal{N} \sim 10^{10}$.

The functional counterpart to the neuroanatomical scheme is the notion of Hebbian associative plasticity [8], considered as the key mechanism that modulates both long- and short-range connections between pyramidal cells. In such a view, autoassociative memory storage and retrieval are universal processes through which both local and global networks operate [2]. Cortical areas across species would then share these

universal processes, whereas the information they express would be specific to the constellation of inputs each area receives, which the simplified skeleton model does not attempt to describe. Underlying the diversity of higher-order processes of which cortical cognition is comprised, there would be the common associative operation of multi-modular autoassociative memory.

At a more abstract mathematical level, the Hopfield model of a simple autoassociative memory network [9] has opened the path to a quantitative statistical understanding of how memory can be implemented in a network of model neurons, through thorough analyses of attractor neural networks. Crucially, it has allowed to sketch a phase diagram, and to approach the nature of the *phase transitions* an associative memory network may demonstrate, what is beyond the reach of non quantitative models. The initial analyses, with networks of binary units, then shifted towards networks with more of the properties seen in the cortex [10], [11].

As for modelling cortical connectivity, attempts to reproduce quantitative observations [12], given the apparent lack of specificity at the single cell level [13], in some cases have led to models in which, even without modules, the probability of pyramidal-to-pyramidal connections depends on the distance between neurons, rapidly decreasing beyond a distance that conceptually corresponds to the radius of a module [14]. In other models, the basis is a strict parcellation into modules, but either with the specific assumptions of binary synapses [15] or with the network interactions across modules different in nature from those within a module (which itself can be structured in sub-modules, or *mini-columns*, with winner-take-all competition among them, and synergy among fully equivalent "clone" units within them [16, 17]), thereby departing from the Braitenberg assumption that associative Hebbian plasticity governs both intra- and inter-module interactions.

But has Braitenberg's suggested simplification, the skeleton of units with their A and B system, both associative, enabled the use of the powerful statistical-physics-derived analyses that had been successfully applied to the Hopfield model? Has it allowed an understanding of phase transitions? Only up to a point. Studies of multi-modular network models including full connectivity within individual modules and sparse connectivity with other modules could only be approached in their most basic formulation, in which all modules participate in every memory, and their sparse

connectivity is random [18], [19]; and attempts to articulate them further have led to analytical complexity [20] [21], [22], [23] or to the recourse to very sparse, effectively local coding schemes [15], without yielding a plausible quantification of storage capacity. The Potts associativ e network, in contrast, has been, from the early study by Ido Kanter [24] fully analysed in its original and sparsely coded versions [25], [26], [27], [28], [29] and it has been argued to offer an ever further simplification of a cortical network than Braitenberg's [30], amenable to study also its latching dynamics [31]. The correspondence between Braitenberg's notion and the Potts model has not, however, been discussed. We do it here, with the aim of establishing a clearer rationale for using the Potts model to study cortical processes.

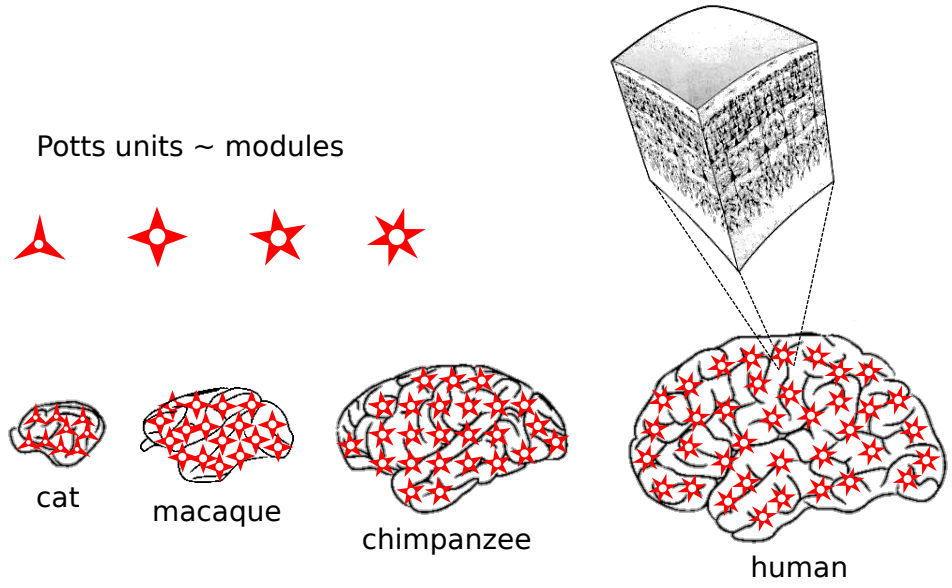


Figure 1: The Braitenberg model regards a skeleton cortex of $\mathcal{N}$ pyramidal cells as comprised of $\sqrt{\mathcal{N}}$ modules of $\sqrt{\mathcal{N}}$ cells each. The Potts model then reduces each module to a multi-state unit, where a state corresponds to a dynamical attractor of the local cortical module. How should the number of states per module, S, be thought to scale with $\mathcal{N}$?

2 The Potts network

The Potts neural network, studied by [24] [25], [26] and [27], is a network of units which can be in more than two states, generalizing Hopfield's binary network, [9], in which units are either active or quiescent. A Potts unit, introduced in statistical physics in 1952 [32] can be regarded in our neuroscience context as representing a local subnetwork or cortical patch of real neurons, endowed with its set of dynamical attractors, which span different directions in activity space, and are therefore converted to the states of the Potts unit (which is defined precisely as having states pointing each along a different dimension of a simplex), as schematically illustrated in Fig. 1. Whatever the interpretation, however, one can define the model as an autoassociative network of N_m Potts units interacting through tensor connections. The memories are

stored in the weight matrix of the network and they are fixed, reflecting an earlier learning phase [9]: each memory μ is a vector or list of the states taken in the overall activity configuration by each unit i: ξ_i^μ. We take each Potts unit to have S possible active states, labelled e.g. by the index k, as well as one quiescent state, $k = 0$, when the unit does not participate in the activity configuration of the memory. Therefore $k = 0, ..., S$, and each ξ_i^μ can take values in the same categorical set. The tensor weights read [24]

$$c_{ij} J_{ij}^{kl} = \frac{c_{ij}}{c_m a(1 - \frac{a}{S})} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \frac{a}{S}\right) \left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right) (1 - \delta_{k0})(1 - \delta_{l0}) \,, \tag{1}$$

where i, j denote units, k, l denote states, a is the fraction of units active in each memory, $c_{ij} = 1$ or 0 if unit j gives input or not to unit i, c_m is the number of input connections per unit, and the δ's are Kronecker symbols. The subtraction of the mean activity per state a/S ensures a higher storage capacity [24]. The units of the network are updated in the following way:

$$\sigma_i^k = \frac{\exp\left(\beta r_i^k\right)}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta(\theta_i^0 + U_i)\right]} \tag{2}$$

and

$$\sigma_i^0 = \frac{\exp\left[\beta(\theta_i^0 + U_i)\right]}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta(\theta_i^0 + U_i)\right]} \,, \tag{3}$$

where r_i^k is the variable representing the input to unit i in state k within a time scale τ_1 and U_i is effectively a threshold. From Eqs. (2) and (3), we see that $\sum_{k=0}^{S} \sigma_i^k \equiv 1$, and note also that σ_i^k takes *continuous* values in the (0,1) range for each k, whereas the memories, for simplicity, are assumed discrete, implying that perfect retrieval is approached when $\sigma_i^k \simeq 1$ for $k = \xi_i^\mu$ and $\simeq 0$ otherwise.

2.1 Potts model dynamics

When the Potts model is studied as a model of cortical *dynamics*, U_i is often written as $U + \theta_i^0$, where U is a common threshold acting on all units, and θ_i^0 is the threshold component specific to unit i, but acting on *all* its active states, and varying in time with time constant τ_3. This threshold is intended to describe local inhibitory

effects, which in the cortex are relayed by $\mathrm{GABA_A}$ and $\mathrm{GABA_B}$ receptors, with widely different time courses, from very short to very long. As discussed elsewhere [33], also the dynamical behaviour of the Potts model is much more interesting if both fast and slow inhibition is included. Here, however, we do not treat dynamics beyond this sketch, and stay with a single τ_3 time constant for the sake of simplicity.

The time evolution of the network is governed by the equations

$$\begin{aligned} \tau_1 \frac{dr_i^k(t)}{dt} &= h_i^k(t) - \theta_i^k(t) - r_i^k(t) \\ \tau_2 \frac{d\theta_i^k(t)}{dt} &= \sigma_i^k(t) - \theta_i^k(t) \\ \tau_3 \frac{d\theta_i^0(t)}{dt} &= \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^0(t) \end{aligned} \tag{4}$$

where the variable θ_i^k is a specific threshold for unit i in state k, varying with time constant τ_2, and intended to model adaptation, i.e. synaptic or neural fatigue specific to the neurons active in state k; and the field that the unit i in state k experiences reads

$$h_i^k = \sum_{j \neq i}^{N_m} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l + w \left(\sigma_i^k - \frac{1}{S} \sum_{l=1}^{S} \sigma_i^l \right). \tag{5}$$

Note that w is another parameter, the "local feedback term", first introduced in [31], aimed at modelling the stability of local attractors in the full model. It helps the network converge towards an attractor, by giving more weight to the most active states, and thus it effectively deepens the attractors.

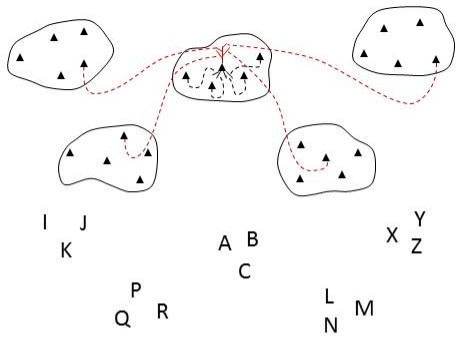
3 From a multi-modular Hopfield network to a Potts network

We do not review here the Hopfield network [9] not its implementation with threshold-linear units [10] but briefly recapitulate, in order to draw the correspondence with the Potts network, the multi-modular version of the threshold-linear Hopfield network, as considered earlier without [18] and with globally sparse coding [21].

Let us consider an underlying network of N_m modules ([18], [19], [21], [23]), each comprised of N_u neurons, each of which is connected to all $N_u - 1$ other neurons within the same module, and to C_A other neurons distributed randomly throughout all the other modules (in earlier papers the notation $L \equiv C_A$ has been used). We make the critical "Hopfield" assumption [9] that both short- and long-range synaptic connections are symmetric. The activity V_i of each neuron is a threshold-linear function of its summed input, as in [10]. The modularity finds expression in the articulation of the global activity patterns that comprise the attractor states of the network. Each module can retrieve one of S local activity patterns, or *features*, that are learned with the corresponding short range connections. We index it with $\xi = 1, \ldots, S$. Furthermore, p global activity patterns, each consisting of combinations of aN_m features, are stored on the dilute long-range connections, as illustrated in Fig. 2. The total number of connections to a neuron is given by $C = C_A + N_u - 1$ and we define the fraction of long range connections as $\gamma = C_A/C$. The model therefore partially incorporates Braitenberg's assumptions, by setting $C_B = N_u - 1$; the implementation would be complete if also $N_m = N_u$, $C_A = N_m - 1$ and therefore $\gamma = 1/2$, but this is not necessary for the analytical treatment.

We make here the simplifying assumption that the firing rates, η, that represent a local pattern ξ within a module m, are identically and independently distributed across units, given by the distribution $P_\eta\left(\eta_{i_m}^{\xi}\right)$. A global pattern, $\mu = 1, \ldots, p$, is a random combination $\{\xi_1^\mu, \ldots, \xi_m^\mu, \ldots, \xi_{N_m}^\mu\}$, with the constraint that only aN_m of the k's are non-zero. We denote as $\zeta \equiv pa/S$ the average number of global patterns represented by a specific local pattern, given global sparsity a, and assume it for simplicity to be an integer number. We also impose, as in [34], that P_η satisfies $\langle\eta\rangle = \langle\eta^2\rangle = a_u$, such

that local representations are also sparse, with sparsity parameter a_u distinct from the global one a, both measures parametrizing, at different scales, sparse coding.



memory\module	1		m		N_m
1		P		M	X
	J	Q		M	
μ	K		B		Y
	I	Q	A		
		R		L	Z
p	K		C	N	

Figure 2: In a cortex comprised of modules, with pyramidal cells receiving their sparse inputs from other modules on the apical dendrites (in color; top panel), memory patterns can be thought of as comprised of features, whose values are coded in the local attractors of each module (middle panel, which reproduces the layout of the modules in the top panel). Features have to be bound together by the tensor connections, in the Potts model. Sparse coding means that not all features pertain to every memory; the rest of the Potts units are in their quiescent state, as in the toy example at the bottom, where $N_m = 5, p = 6, S = 3, a = 0.6$.

Using Hebbian covariance rules [35] in the multi-modular network, we have

$$J^{\text{short}}_{i_m,j_m} = \rho_s \frac{1}{C} \sum_{\mu=1}^{p} \left(\frac{\eta^{\xi^\mu_m}_{i_m}}{a_u} - 1 \right) \left(\frac{\eta^{\xi^\mu_m}_{j_m}}{a_u} - 1 \right) \tag{6}$$

$$J^{\text{long}}_{i_m,j_n} = \rho_l \frac{c_{i_m,j_n}}{C} \sum_{\mu=1}^{p} \left(\frac{\eta^{\mu}_{i_m}}{a_u} - 1 \right) \left(\frac{\eta^{\mu}_{j_n}}{a_u} - 1 \right) \tag{7}$$

where ρ_s and ρ_l are parameters that adjust the dimensions of short- and long-range connections, and can regulate their relative strength. Note that we adopt the more complex index ξ_m^μ in Eq. (6) to emphasize that the summation over μ implies repeatedly using the same local pattern k, for all global patterns that have $\xi_m^\mu = k$. The variable c_{i_m,j_n} is a binary variable

$$c_{i_m,j_n} = \begin{cases} 1 & \text{with probability} \quad \epsilon \\ 0 & \text{with probability} \quad (1-\epsilon) \end{cases} \tag{8}$$

where $\epsilon = C_A/N_u(N_m - 1)$.

In those cases in which an energy function can be defined, i.e., essentially, if $c_{i_m,j_n} = c_{j_n,i_m}$ the attractor states of the system, [36], correspond to the minima of a "free energy". The "Hamiltonian" of the multi-modular network, which is proportional to $N_u \times N_m$, is in those cases given by

$$\begin{aligned} \mathcal{H} &= -\frac{1}{2}\sum_m \sum_{i_m,j_m \neq i_m} J^{\text{short}}_{i_m,j_m} V_{i_m} V_{j_m} - \frac{1}{2}\sum_{m,n\neq m}\sum_{i_m,j_n} J^{\text{long}}_{i_m,j_n} V_{i_m} V_{j_n} \\ &= \mathcal{H}_s + \mathcal{H}_l. \end{aligned} \tag{9}$$

3.1 Thermodynamic correspondence

Estimating c_{i_m,j_n} with its mean ϵ, we can rewrite the second term above as

$$\begin{aligned} \mathcal{H}_l &= -\sum_{m,n>m}\sum_{i_m,j_n} J^{\text{long}}_{i_m,j_m} V_{i_m} V_{j_n} \\ &= -\rho_l \sum_{m,n>m}\sum_{i_m,j_n} \frac{c_{i_m,j_n}}{C} \sum_{\mu=1}^{p} \left(\frac{\eta^\mu_{i_m}}{a_u} - 1\right)\left(\frac{\eta^\mu_{j_n}}{a_u} - 1\right) V_{i_m} V_{j_n} \\ &\simeq -\rho_l \frac{\epsilon}{C} \sum_{m,n>m}\sum_{\mu}\sum_{i_m,j_n} \left(\frac{\eta^\mu_{i_m}}{a_u} - 1\right)\left(\frac{\eta^\mu_{j_n}}{a_u} - 1\right) V_{i_m} V_{j_n}. \end{aligned}$$

We note that for a given pattern μ the only contribution to $\eta^\mu_{i_m}$ is $\eta^{\xi^\mu_m}_{i_m}$. We now define the local correlation of the state of the network with each local memory pattern as

$$\sigma^{\xi^\mu_m}_m = \frac{1}{N_u}\sum_{i_m}\left(\frac{\eta^{\xi^\mu_m}_{i_m}}{a_u} - 1\right) V_{i_m} \tag{10}$$

where to avoid introducing additional dimensional parameters, we assume that the activity V_i of each model neuron is measured in such units, and suitably regulated by inhibition, that the local correlations are automatically normalized to reach a maximum value of 1. We then obtain

$$\begin{aligned}
\mathcal{H}_l &= -\rho_l \frac{\epsilon N_u^2}{C} \sum_{m,n>m} \sum_{\mu} \sigma_m^{\xi_m^\mu} \sigma_n^{\xi_n^\mu} \\
&= -\rho_l \frac{\epsilon N_u^2}{C} \sum_{m,n>m} \sum_{\mu} \sum_{k} \sum_{l} \delta_{\xi_m^\mu k} \delta_{\xi_n^\mu l} \sigma_m^k \sigma_n^l \\
&= -N_u \sum_{m,n>m} \sum_{k,l} J_{mn}^{kl} \sigma_m^k \sigma_n^l \,, \qquad (11)
\end{aligned}$$

where we have introduced

$$J_{mn}^{kl} = \rho_l \frac{\epsilon N_u}{C} \sum_{\mu} \delta_{\xi_m^\mu k} \delta_{\xi_n^\mu l} = \rho_l \frac{\gamma}{N_m - 1} \sum_{\mu} \delta_{\xi_m^\mu k} \delta_{\xi_n^\mu l}. \qquad (12)$$

On the other hand, using Eq. (10), the first term can be rewritten as

$$\begin{aligned}
\mathcal{H}_s &= -\sum_m \sum_{i_m, j_m > i_m} J_{i_m, j_m}^S V_{i_m} V_{j_m} \\
&\simeq -\rho_s \frac{\zeta}{C} \sum_m \sum_{i_m, j_m > i_m} \sum_{\xi=1}^{S} \left(\frac{\eta_{i_m}^\xi}{a_u} - 1 \right) \left(\frac{\eta_{j_m}^\xi}{a_u} - 1 \right) V_{i_m} V_{j_m} \\
&= -\rho_s \frac{\zeta}{C} \sum_m \sum_{\xi=1}^{S} \left\{ \sum_{i_m, j_m} \left(\frac{\eta_{i_m}^\xi}{a_u} - 1 \right) \left(\frac{\eta_{j_m}^\xi}{a_u} - 1 \right) V_{i_m} V_{j_m} - \sum_{i_m} \left[\left(\frac{\eta_{i_m}^\xi}{a_u} - 1 \right) V_{i_m} \right]^2 \right\} \\
&\simeq -\rho_s \frac{\zeta}{C} \sum_m \left\{ N_u^2 \sum_k \left(\sigma_m^k \right)^2 - \frac{S(1-a_u)}{a_u} \sum_{i_m} \left[V_{i_m} \right]^2 \right\}. \qquad (13)
\end{aligned}$$

where we have noted the absence of self-interactions, and estimated with its mean $\zeta \equiv pa/S$ the number of contributions to the encoding of each local attractor state. Putting together Eqs. (11) and (13), where we neglect the last term in the $N_u \to \infty$ limit, and noting that $N_u/C \simeq C_B/C = 1 - \gamma$, we have

$$\mathcal{H} \simeq -N_u \sum_{m,n>m} \sum_{k,l} J_{mn}^{kl} \sigma_m^k \sigma_n^l - N_u \rho_s \zeta (1-\gamma) \sum_m \sum_k \left(\sigma_m^k \right)^2 . \qquad (14)$$

We have therefore expressed the Hamiltonian of a multi-modular Hopfield network in terms of *mesoscopic* parameters, the σ_m^k's, characterizing the state of each module

in terms of its correlation with locally stored patterns. This could be regarded as (proportional to) the effective Hamiltonian of a reduced Potts model, if due attention is paid to entropy and temperature.

First, let us consider the temperature. Since the σ_m^k's are infinite (in the $N_m \to \infty$ limit) but infinitely fewer than the V_i's (in the $N_u \to \infty$ limit), the correct Potts Hamiltonian is akin to a free-energy for the full multimodular model, it should scale with N_m and not with $N_m \times N_u$, and it should include the proper entropy terms. One can write

$$\exp -\beta_{Potts}\mathcal{H}_{Potts}(\{\sigma_m^k\}) = \sum_{\{V_i\}} \exp -\beta\mathcal{H}(\{V_i\}|\{\sigma_m^k\}). \tag{15}$$

The correct scaling of the Potts Hamiltonian implies that an extra N_u factor present in the original Hamiltonian has to be reabsorbed in the effective inverse Potts temperature β_{Potts}, which then diverges in the thermodynamic limit. This means that the Potts network can be taken to operate at zero temperature, in relation to its interactions between modules. Within modules, however, the effects of a non-zero noise level in the underlying multi-modular network persist in the entropy terms.

Let us now turn, then, to the entropy. Here, delineating the correspondence requires suitable assumptions on the distribution of microscopic configurations that dominate the thermodynamic (mesoscopic) state of each module, which are expressed as entropy terms of the effective Potts model. One such assumption is that a module is mostly in states fragmented into competing *domains* of $n_0, n_1, \ldots, n_k, \ldots, n_S$ units, fully correlated with the corresponding local patterns, except for the first n_0, which are at a spontaneous activity level. This would imply that, dropping the module index m, $\sigma^k = n_k/N_u$, and the constraint $\sum_{k=0}^S \sigma^k = 1$ is automatically satisfied. The number of microscopic states characterized by the same $S+1$-plet $n_0, \ldots, n_k, \ldots, n_S$ is $N_u!/\prod_{k=0}^S n_k!$. The log of this number, which can be estimated as $-N_u \sum_{k=0}^S \sigma^k \ln \sigma^k$, has to be divided by β and then subtracted for each module from the original Hamiltonian, as the entropy term that comes from the microscopic free-energy. This becomes the effective Hamiltonian of the Potts network by further dividing by N_u, because a factor N_u has to be reabsorbed into β. Therefore one finds the additional

entropy term in the reduced Hamiltonian

$$\beta\mathcal{H}_{Potts}^{\text{entropy}}(\{\sigma_m^k\}) = \sum_m \sum_{k=0}^{S} \sigma_m^k \ln \sigma_m^k. \tag{16}$$

The above shows that the original inverse temperature β retains its significance as a local parameter, that modulates the stiffness of each module or Potts units, even though the effective noise level in the long-range interactions between modules vanishes. The precise entropy formula depends also on the assumptions that all microscopic states be dynamically accessible from each other, which would have to be validated depending on the dynamics assumed to hold within each module. An alternative assumption is that individual units can in practice only be exchanged between a fragment correlated with local pattern k and the pool n_0 of uncorrelated units. Under that assumption the entropy can be estimated from the log of the number $\prod_{k=1}^{S}(N_u!/n_0!n_k!)$, which yields

$$\beta\mathcal{H}_{Potts}^{\prime\text{entropy}}(\{\sigma_m^k\}) = \sum_m \sum_{k=1}^{S} \left\{ \sigma_m^k \ln \frac{\sigma_m^k}{\sigma_m^k + \sigma_m^0} + \sigma_m^0 \ln \frac{\sigma_m^0}{\sigma_m^k + \sigma_m^0} \right\} \tag{17}$$

as in [31], Eq. (11).

Note that, in Eq. (14), the sparse connectivity between modules of the multi-modular network does not translate into a diluted Potts connectivity: each module, or Potts unit, receives inputs from each of the other $N_m - 1$ modules, or Potts units. One can consider cases in which, instead, there are only c_m connections per Potts unit, e.g. the *highly diluted* and *intermediate connectivity* considered in the storage capacity analysis below.

3.2 Where it gets vague: inhibition and dynamics

These arguments indicate how the local attractors of each module can be reinterpreted as dynamical variables of a system of interacting Potts units. The correspondence cannot be worked out completely, however (and Eq. (14) is not fully equivalent to the Hamiltonian defined in [31]), if anything because the effects of inhibition cannot be included, given the inherent asymmetry of the interactions, in a Hamiltonian formulation. In the body of work on neural networks stimulated by the Hopfield

model, some of the effects ascribed to inhibition have been regarded as incapsulated in the peculiar *Hebbian* learning rule that determines the contribution of each stored pattern to the synaptic matrix, with its subtractive terms. Similar subtractive terms can be argued on the same basis to take into account inhibitory effects at the *module level*, and they lead to replace the interaction

$$J_{mn}^{kl} = \rho_l \frac{\gamma}{N_m - 1} \sum_{\mu} \delta_{\xi_m^\mu k} \delta_{\xi_n^\mu l} \tag{18}$$

with

$$J_{mn}^{\prime kl} = \rho_l \frac{\gamma}{N_m - 1} \sum_{\mu} (\delta_{\xi_m^\mu k} - a/S)(\delta_{\xi_n^\mu l} - a/S), \tag{19}$$

the form which appears in [31]. The local feedback term there, parametrized by w, can be made to roughly correspond to the second term in Eq. (14) by imposing that $\rho_s \zeta (1-\gamma)/\rho_l \gamma = w/2$.

To extend further the approximate correspondence, beyond thermodynamics and into dynamics, we may assume that underlying the Potts network there is in fact a network of $N_m \times N_u$ integrate-and-fire model neurons, emulating the dynamical behaviour of pyramidal cells in the cortex, as considered by [37] and [38]. The simple assumptions concerning the connectivity and the synaptic efficacies are reflected in the fact that the inputs to any model neuron in the extended network are determined by globally defined quantities, namely the mean fields, which are weighted averages of quantities that measure, as a function of time, the effective fraction of synaptic conductances (g, in suitable units normalized to Δg) open on the membrane of any cell of a given class, or cluster (G) by the action of all presynaptic cells of another given class, or cluster (F)

$$z_G^F(t) = \frac{1}{N_{local,F}} \sum_{\alpha \in F} \frac{g^\alpha(t)}{\Delta g_G^F}, \tag{20}$$

where g^α is the conductance of a specific synaptic input. The point is that among the clusters that have to be defined in the framework of Ref.[37], many cluster pairs (F,G), those that comprise pyramidal cells, share the same or a similar biophysical time constant, describing their conductance dynamics [37], i.e.

$$\frac{dz_G^F(t)}{dt} = -\frac{1}{\tau_G^F} z_G^F(t) + \nu_F(t - \Delta t), \tag{21}$$

where $\nu_F(t)$ is the firing rate. If τ_G^F is the same across distinct values for F and G, one can compare the equation for any such cluster pair to the first equation of Eq. (4), namely

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t)\,.$$

Since r_i^k is the temporally integrated variable representing the activity of unit i in state k varying with the time scale of τ_1, it can be taken to correspond to the (integrated) activation of pyramidal cells in a module. One can conclude that τ_1 summarizes the time course of the conductances opened on pyramidal cells by the inputs from other pyramidal cells. It represents the inactivation of synaptic conductance and, like the firing rates are a function of the z, our overlap is a function of the r. Neglecting adaptation (θ_i^k), we can think of the correspondence as

$$h_i^k \sim \sum_{\alpha \in F} \nu^\alpha \rightarrow r_i^k \sim \sum_{\alpha \in F} z^\alpha \tag{22}$$

therefore r_i^k represents the state of the inputs to the integrate-and-fire neurons within a module, i.e., a Potts unit, and we can identify the constant τ_1 with the inactivation time constant for the synapses between pyramidal cells, τ_E^E, whereas inhibitory and adaptation effects will be represented by τ_2 and τ_3 in the Potts model.

The considerations in this subsection appear to be particularly *ad hoc*, as they are bound to be, since we are drawing a possible correspondence which is not a one to one mapping, but rather a reduction to a system with a very large number of variables from another system with a yet much larger number, which itself is intended to represent in simplified form the extreme complexity of the cortex. Still, the correspondence, even though approximate, helps in interpreting the result of the mathematical analysis of the "thermodynamics" of the reduced model, to which we turn next.

4 Storage capacity of the Potts network

In the previous section, we have expressed the approximate equivalence between the Hamiltonian of a multi-modular Hopfield network and that of the Potts network. This means that we can study the retrieval properties of the Potts network, as an effective model of the full multi-modular network.

4.1 Fully connected network

In this subsection, we study the storage capacity of the Potts network with full connectivity using the classic replica method. We quantify the storage load with $\alpha \equiv p/c_m$ or, in the case of full connectivity, $\alpha \equiv p/N$. Taking inspiration from [29] and [31], let us consider the Hamiltonian which is defined as:

$$\mathcal{H} = -\frac{1}{2}\sum_{i,j\neq i}^{N}\sum_{k,l=0}^{S} J_{ij}^{kl}\delta_{\sigma_i k}\delta_{\sigma_j l} + U\sum_{i}^{N}\left(1-\delta_{\sigma_i 0}\right) - \frac{w}{2}\sum_{i}^{N}\left[\sum_{k>0}\delta_{\sigma_i k}^2 - \frac{1}{S}\left(1-\delta_{\sigma_i 0}\right)^2\right]. \tag{23}$$

The coupling between the state k in unit i and the state l in unit j is a Hebbian rule ([29], [9], [27], [31], [39])

$$\begin{cases} J_{ij}^{kl} = \frac{1}{Na(1-\tilde{a})}\sum_{\mu=1}^{p} v_{\xi_i^\mu k} v_{\xi_j^\nu l} \\ \\ v_{\xi_i^\mu k} = \left(\delta_{\xi_i^\mu k} - \tilde{a}\right)\left(1-\delta_{k0}\right) \end{cases} \tag{24}$$

where N is the total number of units in our Potts network (for clarity we drop henceforth the subscript N_m, except when discussing parameters in Sect.6.1), p is the number of stored random patterns, a is their sparsity, i.e., the fraction of active Potts units in each, and $\tilde{a} = a/S$. As mentioned above, U is the time-independent threshold acting on all units in the network, as in [29]. The main difference with the analysis in [29] is that here we have included the term proportional to w in Eq. (23). This self-reinforcement term pushes each unit into the more active of its states, thus providing positive feedback.

The patterns to be learned are drawn from the following probability distribution

([29], [31], [39])

$$\begin{cases} P\left(\xi_i^\mu = 0\right) = 1 - a \\ P_k \equiv P\left(\xi_i^\mu = k\right) = \tilde{a} \equiv a/S\,. \end{cases} \tag{25}$$

Using the trivial property that $\delta_{i,j}^2 = \delta_{i,j}$ we can rewrite the Hamiltonian as

$$\begin{aligned} \mathcal{H} &= -\frac{1}{2Na\left(1-\tilde{a}\right)} \sum_{\mu=1}^{p} \left(\sum_i^N v_{\xi_i^\mu \sigma_i} \right)^2 + \frac{1}{2Na\left(1-\tilde{a}\right)} \sum_i^N \sum_{\mu=1}^{p} v_{\xi_i^\mu \sigma_i}^2 + \\ &+ \left(U - \frac{w\left(S-1\right)}{2S} \right) \sum_i^N \frac{v_{\xi_i^\mu \sigma_i}}{\delta_{\xi_i^\mu \sigma_i} - \tilde{a}}\,. \end{aligned}$$

In the following let us define

$$\tilde{U} = U - \frac{w\left(S-1\right)}{2S}\,. \tag{26}$$

We now apply the replica technique ([40, 41, 42]) to compute from $\mathcal{H}$ a free energy expressed in terms of *overlap* order parameters m^μ, following refs. [11, 24, 36, 43, 44]. The m's measure the correlation between the thermodynamic state of the network and each of the stored memory patterns, and we are interested here in the case where only one such order parameter (pertaining to the so-called *condensed* pattern) differs from zero. The free energy of N Potts units in replica theory reads

$$f = -\frac{1}{\beta} \lim_{n\to 0} \lim_{N\to\infty} \frac{\left\langle Z^n \right\rangle - 1}{Nn}\,, \tag{27}$$

where $\langle\cdot\rangle$ is an average over the quenched disorder (in this case represented by all the other, *uncondensed* patterns in our network), as in [36]. The quenched average requires introducing additional conjugate order parameters q, r, again as in [36], and their diagonal values $\tilde{q}, \tilde{r}$. In Appendix A we compute the replica symmetric free energy to be

$$\begin{aligned} f &= \frac{a\left(1-\tilde{a}\right)}{2} m^2 + \frac{\alpha}{2\beta} \left[\ln\left(a\left(1-\tilde{a}\right)\right) + \ln\left(1-\tilde{a}C\right) - \frac{\beta\tilde{a}q}{\left(1-\tilde{a}C\right)} \right] + \\ &+ \frac{\alpha\beta\tilde{a}^2}{2}\left(\tilde{r}\tilde{q} - rq\right) + \tilde{a}\tilde{q}\left[\frac{\alpha}{2} + S\tilde{U}\right] + \\ &- \frac{1}{\beta} \left\langle \int D\mathbf{z} \ln\left(1 + \sum_{\mathrm{l}\neq 0} \exp\left[\beta \mathcal{H}_l^\xi\right] \right) \right\rangle \end{aligned} \tag{28}$$

where

$$\int Dz = \int dz \frac{\exp\left(-z^2/2\right)}{\sqrt{2\pi}}\,, \tag{29}$$

$C = \beta\left(\tilde{q} - q\right)$ (note that for consistency with the notation in earlier studies we use the same symbol C to denote the –unrelated– total number of connections per unit in the underlying multi-modular model) and

$$\mathcal{H}_l^\xi = m v_{\xi l} - \frac{\alpha a \beta\left(r - \tilde{r}\right)}{2S^2}\left(1 - \delta_{l0}\right) + \sum_{k=1}^{S} \sqrt{\frac{\alpha r P_k}{S\left(1-\tilde{a}\right)}} z_k v_{kl}\,. \tag{30}$$

C and $\mathcal{H}_l^\xi$ are both quantities that are typical of a replica analysis. $\mathcal{H}_l^\xi$ is the mean field with which the network affects state l in a given unit if it is in the same state as condensed pattern ξ (note that $\mathcal{H}_0^\xi = 0$). No such interpretation can be given to C: it measures the difference between $\tilde{q}$, the mean square activity in a given replica, and q, the coactivation between two different replicas. Note that in the zero temperature limit ($\beta \to \infty$), this difference goes to 0, such that C is always of order 1. It will be clarified in section 4.3, through a separate analysis, that C is related to the derivative of the output of an average neuron with respect to variations in its mean field.

The self-consistent mean field equations in the limit of $\beta \to \infty$ are obtained by taking the derivatives of f with respect to the three replica symmetric variational parameters, m, q, r

$$m = \frac{1}{a\left(1-\tilde{a}\right)} \left\langle \int Dz \sum_{l\neq 0} v_{\xi l} \left[\frac{1}{1 + \sum\limits_{n\neq l} \exp\left[\beta\left(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right)\right]} \right] \right\rangle$$

$$\to \frac{1}{a\left(1-\tilde{a}\right)} \sum_{l\neq 0} \left\langle \int Dz\, v_{\xi l} \prod_{n\neq l} \Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right] \right\rangle \tag{31}$$

$$q \to \tilde{q} = \frac{1}{a} \sum_{l\neq 0} \left\langle \int Dz \prod_{n\neq l} \Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right] \right\rangle \tag{32}$$

$$C = \frac{1}{\tilde{a}^2\sqrt{\alpha r}} \sum_{l\neq 0} \sum_k \left\langle \int Dz \sqrt{\frac{P_k}{S\left(1-\tilde{a}\right)}} v_{kl} z_k \prod_{n\neq l} \Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right] \right\rangle \tag{33}$$

$$\tilde{r} \to r = \frac{q}{\left(1 - \tilde{a}C\right)^2} \tag{34}$$

$$\beta\left(r-\tilde{r}\right)=2\left(\tilde{U}\frac{S^2}{a\alpha}-\frac{C}{1-\tilde{a}C}\right). \tag{35}$$

The Θ function gives non-vanishing contribution only for $\mathcal{H}_l^\xi-\mathcal{H}_n^\xi>0$, i.e.

$$\sum_{k>0}\left(v_{kl}-v_{kn}\right)z_k>-m\sqrt{\frac{S^2\left(1-\tilde{a}\right)}{\alpha a r}}\left(v_{\xi l}-v_{\xi n}\right)-\frac{\alpha a\beta\left(r-\tilde{r}\right)}{2S^2}\sqrt{\frac{S^2\left(1-\tilde{a}\right)}{\alpha a r}}\left(\delta_{n0}-\delta_{l0}\right)\,.$$

Moreover, it is convenient to introduce two combinations of order parameters,

$$\begin{aligned} x &= \frac{\alpha a\beta\left(r-\tilde{r}\right)}{2S^2}\sqrt{\frac{S^2\left(1-\tilde{a}\right)}{\alpha a r}}\,, \\ y &= m\sqrt{\frac{S^2\left(1-\tilde{a}\right)}{\alpha a r}}\,. \end{aligned}$$

At the saddle point, they become

$$\begin{aligned} x &= \frac{1}{\sqrt{q}+\tilde{a}C\sqrt{r}}\sqrt{\frac{1-\tilde{a}}{\tilde{\alpha}}}\left[\tilde{U}-\tilde{\alpha}\frac{C}{2}\sqrt{\frac{r}{q}}\right], \\ y &= \sqrt{\frac{1-\tilde{a}}{\tilde{\alpha}}}\left(\frac{m}{\sqrt{q}+\tilde{a}C\sqrt{r}}\right)\,, \end{aligned} \tag{36}$$

where $\tilde{\alpha}=\alpha a/S^2$. By computing the averages in Eqs. (31) and (35), we get three equations that close the self consistent loop with Eq. (36),

$$\begin{aligned} q &= \frac{1-a}{\tilde{a}}\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\phi\left(z\right)^{S-1} \\ &+ \int Dp\int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\phi\left(z+y\right)^{S-1} \\ &+ \left(S-1\right)\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\phi\left(z-y\right)\phi\left(z\right)^{S-2}\,, \end{aligned} \tag{37}$$

$$m = \frac{1}{1-\tilde{a}}\int Dp\int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\phi\left(z+y\right)^{S-1}-q\frac{\tilde{a}}{1-\tilde{a}}\,, \tag{38}$$

$$
\begin{aligned}
C\sqrt{r} &= \frac{1}{\sqrt{\tilde{\alpha}\left(1-\tilde{a}\right)}}\left\{\frac{1-a}{\tilde{a}}\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\left(z+i\sqrt{\tilde{a}}p\right)\phi\left(z\right)^{S-1}\right. \\
&+ \int Dp\int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\left(z+i\sqrt{\tilde{a}}p\right)\phi\left(z+y\right)^{S-1} \qquad (39)\\
&+ \left.(S-1)\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty}Dz\left(z+i\sqrt{\tilde{a}}p\right)\phi\left(z-y\right)\phi\left(z\right)^{S-2}\right\},
\end{aligned}
$$

where $\phi(z)=\left(1+\mathrm{erf}(z/\sqrt{2})\right)/2$. Eqs. (36)-(39) are complicated in their current form, such that it is useful to see their behavior in some limit cases. One such limit case is $\tilde{a}\ll 1$. Using the equalities

$$
\begin{aligned}
\int Dw &= \int\frac{dw}{\sqrt{2\pi}}\exp\left(-w^2/2\right)=1\\
d\phi &= Dz\\
1-\phi(x) &= \phi(-x)
\end{aligned}
$$

and considering that $\phi(x)\sim\Theta(x)$ (the Heaviside function) away from $x\sim 0$, we get to the following self-consistent equations

$$
\begin{aligned}
x &= \frac{1}{\sqrt{\tilde{\alpha}q}}\left(\tilde{U}-\frac{\tilde{\alpha}C}{2}\sqrt{\frac{r}{2}}\right) && (40)\\
y &= \frac{m}{\sqrt{\tilde{\alpha}q}} && (41)\\
m &= \phi(y-x) && (42)\\
q &= \frac{1-a}{\tilde{a}}\phi(-x)+\phi(y-x) && (43)\\
C\sqrt{r} &= \frac{1}{2\pi\tilde{a}}\left\{\frac{1-a}{\tilde{a}}\exp\left(-x^2/2\right)+\exp\left(-(y-x)^2/2\right)\right\}. && (44)
\end{aligned}
$$

4.2 Diluted networks and the highly diluted limit

A more biologically plausible case is that of *diluted* networks, where the number of connections per unit c_m is less than N. Specifically, we consider connections of the form $c_{ij}J_{ij}$, where J_{ij} is the usual symmetric matrix derived from Hebbian

learning. c_{ij} equals 0 or 1 according to a given probability distribution and we note $\lambda = \langle c_{ij} \rangle / N = c_m / N$ the dilution parameter. In general, c_{ij} is different from c_{ji}, leading to asymmetry in the connections between units.

When the connectivity is not full, the type of probability distribution assumed for the c_{ij} matters. We then consider three different distributions. The first is referred to as *random dilution* (RD), which is

$$P(c_{ij}, c_{ji}) = P(c_{ij}) P(c_{ji}) \tag{45}$$

with

$$P(c_{ij}) = \lambda \delta(c_{ij} - 1) + (1 - \lambda) \delta(c_{ij}) \,. \tag{46}$$

The second is the *symmetric dilution* (SD), defined by

$$P(c_{ij}, c_{ji}) = \lambda \delta(c_{ij} - 1) \delta(c_{ji} - 1) + (1 - \lambda) \delta(c_{ij}) \delta(c_{ji}) \,. \tag{47}$$

The third is what we call *state dependent random dilution* (SDRD) –specific to the Potts network– in which

$$P(c_{ij}^{kl}) = \lambda \delta(c_{ij}^{kl} - 1) + (1 - \lambda) \delta(c_{ij}^{kl}) \,; \tag{48}$$

note that in this case the connectivity coefficients are state-dependent.

We have performed simulations with all three types of connectivity, but will focus the analysis onto the RD type, which is the simplest to treat analytically. The storage capacity curve for all three models, estimated from simulations, will be shown later in Fig. 6. RD and SD are known in the literature as Erdos-Renyi graphs. Many properties are known about such random graph models [45], [46]. It is known that for λ below a critical value, essentially all connected components of the graph are trees, while for λ above this critical value, loops are present. In particular, a graph with $c_m < \log(N)$ will almost surely contain isolated vertices and be disconnected, while with $c_m > \log(N)$ it will almost surely be connected. $\log(N)$ is a threshold for the connectedness of the graph, distinguishing the highly diluted limit, for which a simplified analysis of the storage capacity is possible, as in [47], from the intermediate case of the next section, for which a complete analysis is necessary, following the approach by Shiino and Fukai [48].

With Random Dilution, the capacity cannot be analysed through the replica method, as the symmetry of the interactions is a necessary condition for the existence of an energy function, and hence for the application of the thermodynamic formalism. We therefore apply the signal to noise analysis. The local field of unit i in state k writes

$$h_i^k = \sum_j \sum_l c_{ij} J_{ij}^{kl} \sigma_j^l - \tilde{U}\,(1 - \delta_{k,0}) \tag{49}$$

where the coupling strength between two states of two different units is defined as

$$J_{ij}^{kl} = \frac{1}{c_m a(1-\tilde{a})} \sum_\mu v_{\xi_i^\mu k} v_{\xi_j^\mu l}\,. \tag{50}$$

In the highly diluted limit $c_m \sim \log(N)$ at most, the assumption is that the field can be written simply as the sum of two terms, signal and noise. While the signal is what pushes the activity of the unit such that the network configuration converges to an attractor, the noise, or the crosstalk from all of the other patterns, is what deflects the network away from the cued memory pattern. The noise term writes

$$n_i^k \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} \sum_l v_{\xi_i^\mu k} v_{\xi_j^\mu l}\, \sigma_j^l\,,$$

that is, the contribution to the weights J_{ij}^{kl} by all non-condensed patterns. By virtue of the subtraction of the mean activity in each state $\tilde{a}$, the noise has vanishing average:

$$\langle n_i^k \rangle_{P(\xi)} \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} \sum_l \langle v_{\xi_i^\mu,k} \rangle \langle v_{\xi_j^\mu,l} \sigma_j^l \rangle = 0\,.$$

Now let us examine the variance of the noise. This can be written in the following way:

$$\langle (n_i^k)^2 \rangle \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)=1}^{N} \sum_l \sum_{\mu'>1}^{p} \sum_{j'(\neq i)=1}^{N} \sum_{l'} \langle v_{\xi_i^\mu,k}\, v_{\xi_i^{\mu'},k} \rangle \langle v_{\xi_j^\mu,l}\, v_{\xi_{j'}^{\mu'},l'}\, \sigma_j^l \sigma_{j'}^{l'} \rangle\,,$$

where statistical independence between units has been used. For randomly correlated patterns, all terms but $\mu = \mu'$ vanish. Having identified the non-zero term, we can proceed with the capacity analysis. We can express the field using the overlap

parameter, and single out, without loss of generality, the first pattern as the one to be retrieved

$$h_i^k = v_{\xi_i^1 k} m_i^1 + \sum_{\mu>1} v_{\xi_i^\mu k} m_i^\mu - \tilde{U}(1-\delta_{k0}). \tag{51}$$

where we define the local overlap m_i as

$$m_i = \frac{1}{c_m a(1-\tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^1 l} \sigma_j \,. \tag{52}$$

We now write

$$\sum_{\mu>1} v_{\xi_i^\mu, k} m_i^\mu \equiv \sum_{n=1}^{S} v_{n,k}\, \rho^n\, z_i^n \tag{53}$$

where ρ is a positive constant and z_i^n is a standard Gaussian variable. Indeed in highly diluted networks the l.h.s., i.e. the contribution to the field from all of the non-condensed patterns $\mu > 1$, is approximately a normally distributed random variable, as it is the sum of a large number of uncorrelated quantities. ρ can be computed to find

$$\rho^n = \sqrt{\frac{\alpha P_n}{(1-\tilde{a})S} q} \tag{54}$$

where we have defined

$$q = \left\langle \frac{1}{Na} \sum_j \sum_l (\sigma_j^l)^2 \right\rangle. \tag{55}$$

The mean field then writes

$$h_i^k = v_{\xi_i^1 k} m + \sum_{n=1}^{S} v_{n,k} \sqrt{\frac{\alpha P_n}{(1-\tilde{a})S} q} z_n - \tilde{U}(1-\delta_{k0})\,. \tag{56}$$

Averaging m_i and q over the connectivity and the distribution of the Gaussian noise z, and taking the $\beta \to \infty$ we get to the mean field equations that characterize the fixed points of the dynamics, Eqs. (31) and (32). In the highly diluted limit however, we do not obtain the last equation of the fully connected replica analysis, Eq. (34).

The difference between fully connected and diluted cases must vanish in the $\tilde{a} \ll 1$ limit, as shown in ([29], [47]). In this limit we have $x = \tilde{U}/\sqrt{\tilde{\alpha} q}$, $y = m/\sqrt{\tilde{\alpha} q}$ while Eqs. 38 and (37) remain identical.

4.3 Network with partial connectivity

We now consider the more complex case of partial connectivity, *i.e.* $\log(N) < c_m < N$, which can be approached with the self-consistent signal to noise analysis (SCSNA) [48]. As in the previous section, we can express the field using the overlap parameter, and single out the contribution from the pattern to be retrieved, that we label as $\mu = 1$, as in Eq. (49). With high enough connectivity, however, one must revise Eq. (53): the mean field has to be computed in a more refined way, through the SCSNA method that we recapitulate here (see also [14], [49]).

The noise term is assumed to be a sum of two terms

$$\sum_{\mu>1} v_{\xi_i^\mu,k} m_i^\mu = \gamma_i^k \sigma_i^k + \sum_{n=1}^{S} v_{n,k}\, \rho_i^n z_i^n \tag{57}$$

where z_i^n are standard Gaussian variables, and γ_i^k and ρ_i^n are positive constants to be determined self-consistently. The first term, proportional to σ_i^k, represents the noise resulting from the activity of unit i on itself, after having reverberated in the loops of the network; the second term contains the noise which propagates from units other than i. The activation function writes

$$\sigma_i^k = \frac{e^{\beta h_i^k}}{\sum\limits_l e^{\beta h_i^l}} \equiv F^k\Big(\Big\{y_i^l + \gamma_i^l \sigma_i^l\Big\}_l\Big)\,. \tag{58}$$

where $y_i^l = v_{\xi_i^1,l} m_i^1 + \sum_n v_{n,l} \rho_i^n z_i^n - U(1-\delta_{l,0})$. One would need to find σ_i^k as

$$\sigma_i^k = G^k\Big(\Big\{y_i^l\Big\}_l\Big)\,, \tag{59}$$

where G^k are functions solving Eq. (58) for σ_i^k. However, Eq. (58) cannot be solved explicitly. Instead we make the assumption that $\{\sigma_i^l\}$ enters the fields $\{h_i^l\}$ only through their mean value $\langle\sigma_i^l\rangle$, so that we write

$$G^k\Big(\Big\{y_i^l\Big\}_l\Big) \simeq F^k\Big(\Big\{y_i^l + \gamma_i^l \langle\sigma_i^l\rangle\Big\}_l\Big)\,. \tag{60}$$

We report to Appendix B the details of the calculation that yield $\gamma_i^k = \gamma$ and $\rho_i^k = \rho^k$.

$$\gamma = \frac{\alpha}{S}\lambda \frac{\Omega/S}{1-\Omega/S} \tag{61}$$

where $\alpha = p/c_m$, $\langle\cdot\rangle$ indicates the average over all patterns and where we have defined

$$\Omega = \left\langle \frac{1}{N} \sum_{j_1} \sum_{l_1} \frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} \right\rangle . \tag{62}$$

From the variance of the noise term one reads

$$(\rho^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q \Big\{ 1 + 2\lambda\Psi + \lambda\Psi^2 \Big\} , \tag{63}$$

where we have defined

$$q = \left\langle \frac{1}{Na} \sum_{j,l} (G_j^l)^2 \right\rangle \tag{64}$$

and

$$\Psi = \frac{\Omega/S}{1-\Omega/S} \ . \tag{65}$$

The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k} m + \frac{\alpha}{S} \lambda\Psi(1-\delta_{k,0}) + \sum_{n=1}^{S} v_{n,k} z^n \sqrt{\frac{\alpha P_n}{S(1-\tilde{a})} q \Big\{ 1 + 2\lambda\Psi + \lambda\Psi^2 \Big\}} - \tilde{U}(1-\delta_{k,0}) . \tag{66}$$

Taking the average over the non-condensed patterns (the average over the Gaussian noise z), followed by the average over the condensed pattern $\mu = 1$ (denoted by $\langle\cdot\rangle_\xi$), in the limit $\beta \to \infty$, we get the self-consistent equations satisfied by the order parameters

$$m = \frac{1}{a(1-\tilde{a})} \left\langle \int D^S z \sum_{l(\neq 0)} v_{\xi,l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi , \tag{67}$$

$$q = \frac{1}{a} \left\langle \int D^S z \sum_{l(\neq 0)} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi , \tag{68}$$

$$\Omega = \left\langle \int D^S z \sum_{l(\neq 0)} \sum_k z^k \frac{\partial z^k}{\partial y^l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi . \tag{69}$$

where in the last equation for Ω, integration by parts has been used. Note the similarities to Eqs. (31)-(33), obtained through the replica method for the fully

connected case. The equations just found constitute their generalization to $\lambda < 1$. In particular, in the highly diluted limit $\lambda \to 0$, we get $\gamma \to 0$ and $(\rho^n)^2 \to \alpha P_n q/(1-\tilde{a})S$, which are the results obtained in the previous section; in the fully connected case, $\lambda = 1$, the correspondence between the m and q variables is obvious, while for Ω it can be shown with some algebraic manipulation. Indeed, from the following identity,

$$\rho^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q(1+\Psi)^2 \,, \tag{70}$$

by using the replica variable $r = q/(1-\tilde{a}C)^2$ we get

$$\rho^2 = \frac{\alpha P_n}{S(1-\tilde{a})} r(1-\tilde{a}C)^2(1+\Psi)^2 \,. \tag{71}$$

By comparing this with Eq. (30), the mean field, we get an equivalent expression for Ψ,

$$\Psi = \frac{\tilde{a}C}{1-\tilde{a}C} \,. \tag{72}$$

From the original definition of Ψ in Eq. (65), it follows that the order parameter C, obtained through the replica method, is equivalent to Ω, up to a multiplicative constant:

$$C = \Omega/a \,. \tag{73}$$

We can show that Eq. (69) coincides with Eq. (33). Moreover, by comparing the SCSNA result for γ to the replica one, we must have

$$\frac{\alpha}{S}\Psi - \tilde{U} = -\frac{\alpha a \beta (r - \tilde{r})}{2S^2} \tag{74}$$

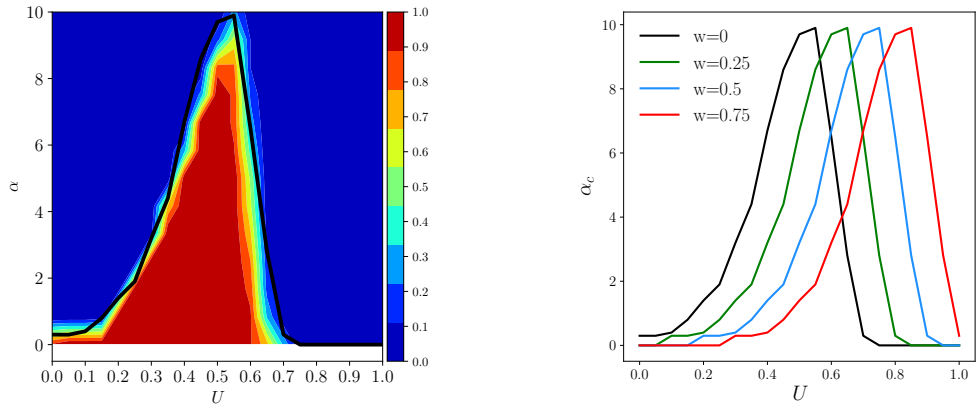
from which

$$\beta(r-\tilde{r}) = 2\left(\tilde{U}\frac{S^2}{\alpha a} - \frac{C}{1-\tilde{a}C}\right) \,, \tag{75}$$

identical to Eq. (35).

5 Simulation results

Do computer simulations confirm the analyses above? Starting with the effect of setting the overall threshold, we show, in Fig. 3(a), retrieval performance as a function of the threshold for $w = 0.0$, both through simulations and by solving Eqs. (36).



(a) (b)

Figure 3: (a) How often a fully connected Potts network retrieves memories, as a function of the threshold U and the number of stored memories p, with $N = 1000$, $S = 7$, $a = 0.25$, $\beta = 200$ and $w = 0.0$. Color represents the fraction of simulations in which the overlap between the activity state of network and a stored pattern is ≥ 0.9. The solid line is obtained by numerical solution of Eqs.(36)-(39). (b) The dependence of α_c on U for different values of w. If the threshold U is already set to its optimal value, subtracting a non-zero w is detrimental to the capacity, but if it can be adjusted *after* considering w, it can lead to an optimal effective threshold $\tilde{U}$, maximizing capacity.

It is clear that the simulations agree very well with numerical results. The maximum storage capacity α_c (where $\alpha \equiv p/c_m$, or $\alpha \equiv p/N$ for a fully connected Potts network) is found at approximately $U = 0.5$, as can also be shown through a simple signal to noise analysis. It is possible to compute approximately the standard deviation γ_i^k of the field, Eq. (49) with respect to the distribution of all the patterns,

as well as the connectivity c_{ij}, by making the assumption that all units are aligned with a specific pattern to be retrieved $\sigma_j^l = \xi_j^1$. We further discriminate units that are in active states $\xi_i^1 \neq 0$ from those that are in the quiescent states $\xi_i^1 = 0$ in the retrieved pattern $\mu = 1$.

$$\gamma_i^k \equiv \sqrt{\langle (h_i^k)^2 \rangle - \langle h_i^k \rangle^2} = \sqrt{\frac{(p-1)a}{c_m S^2} + (\delta_{\xi_i^1,k} - \tilde{a})^2 \left(\frac{1}{c_m a} - \frac{1}{N} \right)}. \tag{76}$$

The optimal threshold U_0 is one that separates the two distributions, optimally, such that a minimum number of units in either distribution reach the threshold to go in the wrong state

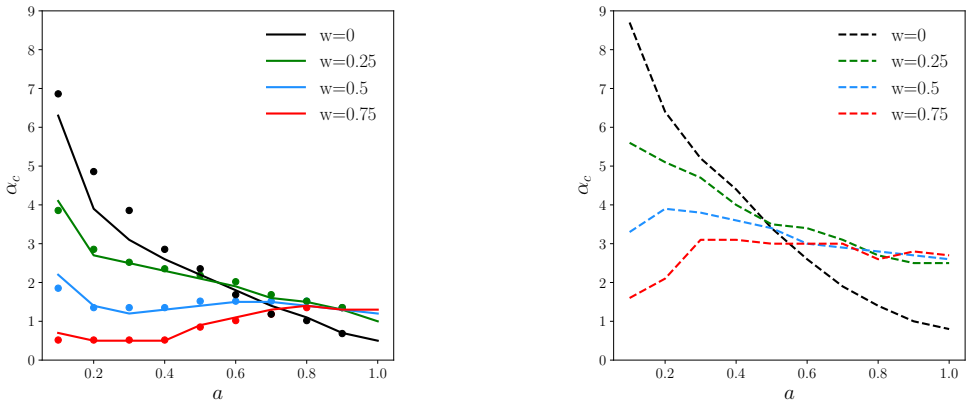
$$\frac{U_0 - \langle h_i^k |_{\xi_i^1=0} \rangle}{\gamma_i^k |_{\xi_i^1=0}} = - \frac{U_0 - \langle h_i^k |_{\xi_i^1 \neq 0} \rangle}{\gamma_i^k |_{\xi_i^1 \neq 0}}$$

$$U_0 = \frac{\gamma_i^k |_{\xi_i^1=0}}{\gamma_i^k |_{\xi_i^1=0} + \gamma_i^k |_{\xi_i^1 \neq 0}} - \frac{a}{S}. \tag{77}$$

We can see that $U_0 \longrightarrow 1/2 - \tilde{a}$ for $\gamma_i^k |_{\xi_i^1=0} \sim \gamma_i^k |_{\xi_i^1 \neq 0}$, roughly consistent with the replica analysis and simulations in Fig. 3(a) (in fact the variance $\gamma_i^k |_{\xi_i^1=0}$ is larger than $\gamma_i^k |_{\xi_i^1 \neq 0}$, especially for low a, hence U_0 is slightly larger and has a more complex dependence on the sparsity). Given such an optimal value for U, Fig. 3(b) shows that the effect of the feedback term w on the storage capacity, purely subtractive, is just to shift to the right the optimal value.

5.1 The effect of network parameters

Fig. 4 illustrates the same effect of the feedback term, by setting $U = 0.5$ and charting the storage capacity as a function of the sparsity a for different values of w, for both fully connected (a) and highly diluted networks (b). In both cases, α_c decreases monotonically with increasing w, for low a, when $U = 0.5$ is close to optimal. Increasing a, one reaches a region where $U = 0.5$ is set too high, and therefore α_c benefits from a non-zero w, even though its exact value is not critical. For very high sparsity parameter (non-sparse coding) all curves except $w = 0$ seem to coalesce. The envelope of the different curves represents optimal threshold setting that takes

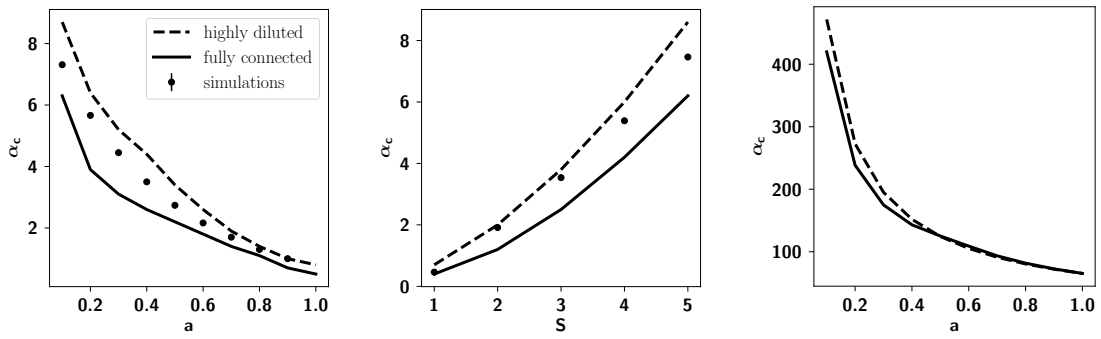


(a) Fully connected network (b) Diluted network

Figure 4: Storage capacity α_c as a function of sparsity a for different values of w for both fully connected (a) and highly (RD) diluted networks (b) as obtained by numerical solution of Eqs. (36)-(39). (a) also includes points from simulations. The parameters are $S = 5$, $U = 0.5$, $\beta = 200$.

feedback into account, and as a function of a it shows, both for fully connected and diluted networks the decreasing trend familiar from the analysis of simpler memory networks [50].

The two connectivity limit cases are illustrated in Fig. 5, which shows, in (a), the dependence of the storage capacity α on the sparsity a in the fully connected and diluted networks with $U = 0.5$, $w = 0$ and $S = 5$. In Fig. 5 (b) instead, S is varied and in Fig. 5 (c) $S = 50$, corresponding to the highly sparse limit $\tilde{a} \ll 1$. While for $S = 5$ the two curves are distinct, for the highly sparse network with $S = 50$ the two curves coalesce. The curves are obtained by numerically solving Eqs. (36)-(39). Moreover, the storage capacity curve for the fully connected case in (a) matches very well with Fig. 2 of [29]. Diluted curves are always above the fully connected ones in both (a) and (b), as found in [29].



(a) (b) (c)

Figure 5: (a) Storage capacity α_c as a function of the sparsity a. Dots correspond to simulations of a network with $N = 2000$, $c_m/N = 0.1$, $S = 5$, and $\beta = 200$ while curves are obtained by numerical solution of Eqs. (36)-(39). (b) Storage capacity as a function of S with the same parameters as in (a) and with $a = 0.1$. (c) $S = 50$, illustrating the $\tilde{a} \ll 1$ limit case.

5.2 The effect of the different connectivity models

In Fig. 6 we show simulation results for the storage capacity of all three connectivity models introduced earlier. The RD and SDRD networks seem to have almost identical capacity. All models have the same capacity in the fully connected case, as they should. Note in particular the very limited decrease of $\alpha_c = p/c_m$ with c_m/N increasing up to almost full connectivity, with all three models. In particular with the RD model, as already shown analytically, the degree of dilution has almost no effect, because already for moderate values of S the network is effectively in the sparse coding regime, $a/S \ll 1$, where c_m/N becomes irrelevant. The apparent decrease in capacity for very low c_m/N values is likely an artefact of c_m being very small.

Our results can be contrasted to the storage capacity with the same connectivity models (RD and SD; SDRD is not relevant) of the Hopfield model. For the Hopfield model, the effects of SD were investigated and it was found that the capacity decreases monotonically from the value $\simeq 0.4$ for highly diluted to the well-known value of $\alpha_c \simeq 0.14$ for the fully connected network [51]. In [47], instead, the highly diluted limit

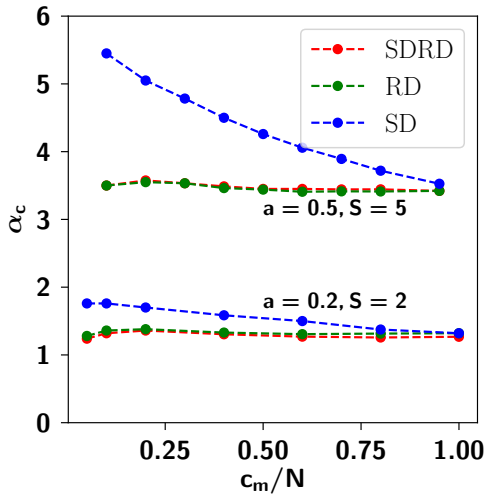


Figure 6: Storage capacity curves, obtained through simulations, as a function of the mean connectivity per unit c_m/N, for three different types of connectivity, namely the random dilution (RD), symmetric dilution (SD) and state-dependent random dilution (SDRD). We find that SD has higher capacity than RD. The capacity for all three models coalesces at the fully connected limit, as the models become equivalent. Simulations carried out for two sets of parameters: ($N = 5000$, $S = 2$, $a = 0.2$) and ($N = 2000$, $S = 5$, $a = 0.5$). $U = 0.5$ and $\beta = 200$.

of RD was studied and a value of $\alpha_c = 2/\pi \simeq 0.64$ was found. If we plausibly assume that the intermediate RD values interpolate those of the highly diluted $\alpha_c \simeq 0.64$ and fully connected $\alpha_c \simeq 0.14$ limit cases, the Hopfield network seems to have higher capacity for RD than for SD.

However, it is important to note that the overlap with which the network retrieves at α_c, m_c, is not the same in the two models (RD and SD). In the highly diluted RD model [47], the authors find that at zero temperature (which is the only case we consider), m_c undergoes a second order phase transition with control parameter α, such that $m_c \simeq \sqrt{3(\alpha_c - \alpha)}$: close to α_c, m_c is small, and smaller than the values of m_c for the highly diluted SD model [51] that we report in the left y-axis of Fig. 7: at $c_m/N \simeq 0.0024$, $m_c \simeq 0.64$. If we require the same precision of retrieval from the RD model, the above equation yielding the m_c gives us a value $\alpha \simeq 0.5$, still higher than

the analytic SD value of 0.4. However, through the simulations shown in Fig. 7 we have found that the SD network has a higher capacity (> 0.6) than the one predicted analytically (0.4).

When taking into consideration, for the Hopfield model, the increased capacity of the SD model with respect to what is predicted analytically, as well as the precision of retrieval, we find that the two models behave similarly. We clarify this in the next section by making the Potts-Hopfield correspondence exact, in a different sense than when considering the multi-modular Hopfield model.

5.3 The Hopfield model as a special case of the Potts model, for $S = 1$

We can rewrite the Potts Hamiltonian, Eq. (23) with $S = 1$, $a = 0.5$, $U = w = 0.0$ such that:

$$H = -\frac{1}{2}\sum_{i,j\neq i}^{N} J_{ij}\sigma_i\sigma_j\,, \tag{78}$$

$$J_{ij} = \frac{4}{c_m}\sum_{\mu=1}^{p}\left(\xi_i^\mu - \frac{1}{2}\right)\left(\xi_j^\mu - \frac{1}{2}\right)\,. \tag{79}$$

where σ and ξ take the values $\{0, 1\}$. We can rewrite the latter quantities using the spin formulation $\{-1, +1\}$ using the transformation $2\sigma_i = s_i + 1$

$$\tilde{H} = -\frac{1}{8}\sum_{i,j\neq i}^{N} c_{ij}\tilde{J}_{ij}s_i s_j - \frac{1}{8}\sum_{i,j\neq i}^{N} c_{ij}\tilde{J}_{ij}(s_i + s_j) - \frac{1}{8}\sum_{i,j\neq i}^{N} c_{ij}\tilde{J}_{ij}\,, \tag{80}$$

$$\tilde{J}_{ij} = \frac{1}{c_m}\sum_{\mu=1}^{p}\eta_i^\mu\eta_j^\mu\,. \tag{81}$$

We note now that the first term in Eq. (80) is the Hopfield Hamiltonian for storing unbiased patterns, modulo a multiplicative term 1/4 [43]; at zero-temperature, however, an overall rescaling of the energies leaves the statistics of the system unchanged, so that we can consider the first term in Eq. (80) as exactly the Hopfield Hamiltonian. The last term is an additive constant that can be neglected, while the second term

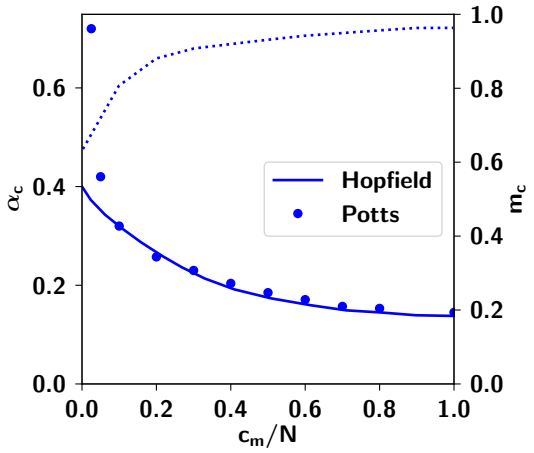


Figure 7: Setting $S = 1$, $a = 0.5$ and the threshold to be unit dependent ($U = U_i$) the Hamiltonians of the two models become equivalent. Dots correspond to simulations of the Potts network with the latter parameters, while the uninterrupted line corresponds to analytical results obtained by Sompolinsky. The dashed line, to be read with the right y-axis, corresponds to the overlap at the critical capacity. For intermediate values of the connectivity, up to $c_m/N = 0.1$, our simulation results fit the analytical curve well, and we find, in particular, the well-known value of $\simeq 0.14$ for the fully connected network. For higher levels of dilution, we find a greater capacity than predicted analytically. Simulations performed with a network of $N = 2000$ units.

can be made to vanish by the addition of a unit dependent threshold term to Eq. (80)

$$\tilde{U}_i = \frac{1}{8} \sum_{j(\neq i)}^{N} (c_{ij} + c_{ji}) \tilde{J}_{ij} \tag{82}$$

or equivalently, to Eq. (78) using the binary formulation

$$H = -\frac{1}{2} \sum_{i,j \neq i}^{N} J_{ij} \sigma_i \sigma_j + \sum_{i}^{N} \left(\frac{1}{4} \sum_{j(\neq i)}^{N} (c_{ij} + c_{ji}) J_{ij} \right) \sigma_i \tag{83}$$

Considering c_{ij} to be of the SD type such that $c_{ij} = c_{ji}$, this is the Hamiltonian considered by Sompolinsky [51]. The system with Hamiltonian given by Eq. (83) can be simulated by setting the parameters of the Potts network to $S = 1$, $a = 0.5$ and $U = U_i$ and the results compared to the analytical results derived in the latter study.

We have carried out the simulations to reproduce these results, that we report in Fig. 7. Instead, considering c_{ij} to be of the RD type yields the model studied by [47] at the highly diluted limit.

The unit-dependent threshold that correlates with the learned patterns (and in our case with the diluted connectivity), for the equivalence of the two formulations of the Hamiltonians with spin and binary variables, was first found to be significant when the storage of biased patterns was considered [35].

6 Discussion

In this paper we elaborate on the correspondence between a multi-modular neural network and a coarse grained Potts network, by grounding the Hamiltonian of the Potts model in the multi-modular one. Units are taken to be threshold-linear, in the multi-modular model, and they are fully connected within a module, with Hebbian synaptic weights. Sparse connectivity links units that belong to different modules, via synapses that in the cortex impinge primarily on the apical dendrites, after their axons have travelled through the white matter.

We relate Potts states to the overlap or correlation between the activity state in a module and the local memory patterns, i.e., to weighted combinations of the activity of its threshold-linear units. The long range interactions between the modules then roughly correspond, after suitable assumptions about inhibition, to the tensorial couplings between Potts units in the Potts Hamiltonian. It becomes apparent how the w-term, which was initially introduced by [31] to model positive state-specific feedback on Potts units, arises from the short range interactions of the multi-modular Hamiltonian.

Keeping the w-term in the Potts Hamiltonian, we apply the replica method to derive analytically the storage capacity for the fully connected Potts model. A simplified derivation is applied also to the highly diluted connectivity network, while the case with intermediate connectivity is studied by a self-consistent signal-to-noise analysis. The intermediate results smoothly interpolate the limit cases of fully and high diluted networks, but the two limit cases themselves are in fact very similar in capacity, if measured by $\alpha \equiv p/c_m$, in the sparse coding limit $a \to 0$, a limit which is approached very rapidly in the Potts model, because the relevant parameter is in fact $\tilde{a} \equiv a/S$. The effect of w term is effectively, in the vicinity of the memory states, reduced to altering the threshold, which leads to the storage capacity being suppressed by this term, if the threshold was originally close to its optimal value. If one assumes that the threshold is set close to its optimal value *after* taking the feedback term into account, the value w becomes irrelevant for the storage capacity, while it still affects network dynamics [33].

6.1 The storage capacity parameters

In the end, the storage capacity of the Potts network is primarily a function of a few parameters, c_m, S and a, that suffice to broadly characterize the model, with minor adjustments due to other factors. How can these parameters be considered to reflect cortically relevant quantities? This a critical issue, if we are to make cortical sense of the distinct thermodynamic phases that can be analysed with the Potts model, and to develop informed conjectures about cortical phase transitions [30].

The Potts network, if there are N_m Potts variables, requires, in the fully connected case, $N_m \cdot (N_m - 1) \cdot S^2/2$ connection variables (since weights are taken to be symmetric we have to divide by 2). In the diluted case, we would have $N_m \cdot c_m \cdot S^2$ variables (the factor 2 is no longer relevant, at least for $c_m \to 0$). The multi-modular Hopfield network, as shown in Sect.3, has only $N_m \cdot N_u \cdot C_A$ long-range synaptic weights. This diluted connectivity between modules is summarily represented in the Potts network by the tensorial weights. Therefore, the number of Potts weights cannot be larger than the total number of underlying synaptic weights it represents. Then $c_m \cdot S^2$ cannot be larger than $C_A \cdot N_u$.

In the simple Braitenberg model of mammalian cortical connectivity [4], which motivated the multi-modular network model [18], $N_u \simeq N_m \sim 10^3 - 10^5$, as the total number of pyramidal cells ranges from $\sim 10^6$ in a small mammalian brain to $\sim 10^{10}$ in a large one. In a large, e.g. human cortex, a module may be taken to correspond to roughly $1\,\mathrm{mm}^2$ of cortical surface, also estimated to include $N_u \sim 10^5$ pyramidal cells [52]. A module, however, cannot be plausibly considered to be fully connected; the available measures suggest that, even at the shortest distance, the connection probability between pyramidal cells is at most of order 1/10. Therefore we can write, departing from the assumption $C_B = N_u - 1$ in the simplest version of Braitenberg's model, that $C_B \simeq 0.1 N_u$. If we were to keep the approximate equivalence $C_A \simeq C_B$, that would imply also $C_A \simeq 0.1 N_u$. Inserting this into the inequality above, $c_m \cdot S^2 < C_A \cdot N_u$, yields the constraint $S < N_u \sqrt{0.1/c_m}$.

One can argue, however, for another constraint that limits the value S, given the Potts connectivity c_m. The number S of local patterns on N_u neurons receiving C_B connections from each other can at most be, given associative storage of patterns

with sparsity a_u, of order C_B/a_u. If we assume that local storage tends to saturate this capacity bound, *and* we take $C_A \simeq C_B$, we have $S \cdot a_u \simeq C_B \simeq C_A$, but again we have, above, $c_m \cdot S^2 < C_A \cdot N_u$, hence

$$S < N_u \cdot a_u / c_m.$$

This more stringent upper bound is compatible with S small and c_m scaling linearly with N_u, as well as with c_m small and S scaling with N_u, and all intermediate regimes. If we take S and c_m to be both proportional to $\sqrt{N_u}$, and $a_u \sim 0.1$, it would lead to c_m and S to be at most of order $10^1 - 10^2$ over mammalian cortices of different scale, essentially scaling like the fourth root of the total number of pyramidal cells, which appears like a plausible, if rough, modelling assumption.

We could take these range of values, together with the approximate formula (see [29] and Fig. 5b)

$$p_c \sim 0.15 \frac{c_m S^2}{a \ln(S/a)} \tag{84}$$

to yield estimates of the actual capacity the cortex of a given species. The major factor that such estimates do not take into account, however, is the correlation among the memory patterns. All the analyses reported here apply to randomly assigned memory patterns. The case of correlations will be treated elsewhere (Boboeva, PhD Thesis, SISSA, 2018; and Boboeva *et al.*, in preparation).

The above considerations may sound rather vague. They neglect, *inter alia*, the large variability in the number of spines, hence probably in synapses, among cortical areas within the same species [55]. They capture, however, the quantitative change of perspective afforded by the coarse graining inherent in the Potts model. We can simplify the argument by neglecting sparse coding as well as the exact value of the numerical pre-factor k (which is around 0.15 in Eq. (84). The Potts model uses $N_m c_m S^2$ weights to store up to $k c_m S^2 / \ln S$ memory patterns, each containing of order $N_m \ln S$ bits of information, therefore storing up to k bits per weight. In this respect, and in keeping with the Frolov conjecture [53], it is not different from any other associative memory network based on Hebb-like plasticity, including the multi-modular model which it effectively represents. In the multi-modular model, however, (in its simplest version) the $2kN_u^2 N_m$ bits available are allocated to memory

patterns that are specified in single-neuron detail, and hence contain of order $N_u N_m$ bits of information each. The network can store and retrieve up to a number p_c of them, which has been argued in [19] to be limited by the *memory glass* problem to be of the same order of magnitude as the number S of local attractors, itself limited to be (at most) of order N_u or perhaps, as argued above, $\sqrt{N_u}$. By glossing over the single-neuron resolution, the Potts model forfeits the locally extensive character of the information contained in each pattern, losing a factor $N_u/\ln S$, but it gains the factor $c_m S^2/(2N_u \ln S)$ in the number of patterns. Whether S scales with N_u or with $\sqrt{N_u}$ or in between, the upshot is more, but less informative, memories. Therefore, by focusing on long range interactions the Potts model misses out in information, but effectively circumvents the memory glass issue, which had plagued the earlier incarnation of the Braitenberg idea [52], and stores more patterns. How is that possible, if the Potts model is a reduced description of the underlying multi-modular model? The trick is likely in the Hebbian form of the tensor interactions, Eq. (19), which is *not* a straightforward reduction – it implies a fine inhibitory regulation that the multi-modular model had not attempted to achieve. This argument can be expanded and made more precise by considering, again, a more plausible scenario with correlated memories.

Finally, separate studies are needed also to assess the dynamical properties of Potts network, which also reflect the strength of the w-term, as we have begun to undertake in an earlier paper elsewhere [33]. It is such an analysis of the dynamics that may reveal the unique statistical properties of large cortices, as expressed in latching dynamics [30].

Acknowledgements

Work supported by the Human Frontier collaboration with the groups of Naama Friedmann and Rémi Monasson on analog computations underlying language mechanisms, HFSP RGP0057/2016.

A Calculation of replica symmetric free energy

The partition function Z^n of n replicas can be written as

$$\begin{aligned}\left\langle Z^n \right\rangle &= \left\langle \mathrm{Tr}_{\{\sigma^\gamma\}} \exp\left[-\beta \sum_\gamma^n H^\gamma\right] \right\rangle \qquad (85)\\ &= \left\langle \mathrm{Tr}_{\{\sigma^\gamma\}} \exp\left[\frac{\beta}{2Na\left(1-\tilde{a}\right)} \sum_{\mu\gamma} \left(\sum_i^N v_{\xi_i^\mu \sigma_i^\gamma}\right)^2 - \frac{\beta}{2Na\left(1-\tilde{a}\right)} \sum_i^N \sum_{\mu\gamma} v^2_{\xi_i^\mu \sigma_i^\gamma}\right.\right.\\ &\quad \left.\left. -\beta\tilde{U} \sum_{i\gamma} \frac{v_{\xi_i^\mu \sigma_i^\gamma}}{\delta_{\xi_i^\mu \sigma_i^\gamma} - \tilde{a}}\right]\right\rangle.\end{aligned}$$

Using the Hubbard-Stratonovich transformation

$$\exp\left[\lambda a^2\right] = \int \frac{dx}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2} + \sqrt{2\lambda} a x\right],$$

the first term in Eq. (86) can be written as

$$\exp\left[\frac{\beta}{2Na\left(1-\tilde{a}\right)} \left(\sum_i^N v_{\xi_i^\mu \sigma_i^\gamma}\right)^2\right] = \int \frac{dm_\mu^\gamma}{\sqrt{2\pi}} \exp\left[-\frac{\left(m_\mu^\gamma\right)^2}{2} + \sqrt{\frac{\beta}{Na\left(1-\tilde{a}\right)}} m_\mu^\gamma \sum_i^N v_{\xi_i^\mu \sigma_i^\gamma}\right].$$

The change of variable $m_\mu^\gamma \to m_\mu^\gamma \sqrt{\beta N a \left(1-\tilde{a}\right)}$, and neglecting the sub-leading terms in the $N \to \infty$ limit, gives us

$$\begin{aligned}\left\langle Z^n \right\rangle &= \left\langle \mathrm{Tr}_{\{\sigma^\gamma\}} \int \prod_{\mu\gamma} dm_\mu^\gamma \cdot \right.\\ &\quad \cdot \exp \beta N \left[\frac{a\left(1-\tilde{a}\right)}{2} \sum_{\mu\gamma} \left(m_\mu^\gamma\right)^2 + \sum_{\mu\gamma} \frac{m_\mu^\gamma}{N} \sum_i^N v_{\xi_i^\mu \sigma_i^\gamma} - \frac{1}{2N^2 a\left(1-\tilde{a}\right)} \sum_i^N \sum_{\mu\gamma} v^2_{\xi_i^\mu \sigma_i^\gamma}\right.\\ &\quad \left.\left. -\frac{1}{N}\tilde{U} \sum_{i\gamma} \frac{v_{\xi_i^\mu \sigma_i^\gamma}}{\delta_{\xi_i^\mu \sigma_i^\gamma - \tilde{a}}}\right]\right\rangle. \qquad (86)\end{aligned}$$

Discriminating the condensed patterns (ν) from non condensed ones (μ) in the limit $p \to \infty$ and $N \to \infty$ with the fixed ratio $\alpha = p/N$,

$$\left\langle Z^n \right\rangle = \mathrm{Tr}_{\{\sigma^\gamma\}} \int \prod_{\mu\gamma} dm_\mu^\gamma \int \prod_{\lambda\gamma} dq_{\gamma\lambda} dr_{\gamma\lambda} \cdot \exp\left\{ -\frac{\beta N}{2} \sum_{\mu>s} \left[a\left(1-\tilde{a}\right) \sum_\gamma \left(m_\mu^\gamma\right)^2 \right.\right.$$
$$\left. -a\left(1-\tilde{a}\right)\beta\tilde{a} \sum_{\gamma\lambda} m_\mu^\gamma m_\mu^\lambda q_{\gamma\lambda} \right] - \frac{\alpha\beta\tilde{a}N}{2} \sum_{\gamma\gamma} q_{\gamma\gamma} - \beta N a \tilde{U} \sum_{\gamma\gamma} q_{\gamma\gamma}$$
$$\left. -\frac{N\alpha\beta^2}{2} \sum_{\gamma\lambda} r_{\gamma\lambda} \left(\tilde{a}^2 q_{\gamma\lambda} - \frac{1}{NS\left(1-\tilde{a}\right)} \sum_{ik} P_k v_{k\sigma_i^\gamma} v_{k\sigma_i^\lambda} \right) \right\} \cdot \left\langle \exp \beta N \left[\frac{a(1-\tilde{a})}{2} \right.\right.$$
$$\left.\left. \sum_{\nu\gamma}^{\nu\leq s} \left(m_\nu^\gamma\right)^2 + \sum_{\nu\gamma}^{\nu\leq s} \frac{m_\nu^\gamma}{N} \sum_i^N v_{\xi_i^\nu \sigma_i^\gamma} - \frac{1}{2N^2 a(1-\tilde{a})} \sum_i^N \sum_{\nu\gamma}^{\nu\leq s} v^2_{\xi_i^\nu \sigma_i^\gamma} \right] \right\rangle \tag{87}$$

where we introduced $q_{\gamma\lambda}$, the overlap between different replicas, analogous to the Edwards-Anderson order parameter [54],

$$q_{\gamma\lambda} = \frac{1}{Na\tilde{a}\left(1-\tilde{a}\right)} \sum_{ik} P_k v_{k\sigma_i^\gamma} v_{k\sigma_i^\lambda} \,. \tag{88}$$

The saddle point equations are

$$\frac{\partial}{\partial m_\nu^\gamma} = 0 \longrightarrow m_\nu^\gamma = \left\langle \frac{1}{Na\left(1-\tilde{a}\right)} \sum_i \left\langle v_{\xi_i^\nu \sigma_i^\gamma} \right\rangle \right\rangle, \tag{89}$$

$$\frac{\partial}{\partial r_{\gamma\lambda}} = 0 \longrightarrow q_{\gamma\lambda} = \frac{1}{Na\tilde{a}\left(1-\tilde{a}\right)} \sum_i^N \left\langle \sum_k P_k \left\langle v_{k\sigma_i^\gamma} v_{k\sigma_i^\lambda} \right\rangle \right\rangle, \tag{90}$$

$$\frac{\partial}{\partial q_{\gamma\lambda}} = 0 \longrightarrow r_{\gamma\lambda} = \frac{S\left(1-\tilde{a}\right)}{\alpha} \sum_\mu \left\langle m_\mu^\gamma m_\nu^\lambda \right\rangle - \left[\frac{2S}{\alpha} \tilde{U} + 1 \right] \frac{\delta_{\gamma\lambda}}{\beta\tilde{a}} \,. \tag{91}$$

After performing the multidimensional Gaussian integrals over fluctuating (non condensed) patterns we have

$$\left\langle Z^n \right\rangle = \int \prod_{\nu\gamma}^{\nu\in[1,\ldots,s]} dm_\nu^\gamma \int \prod_{\lambda\gamma} dq_{\gamma\lambda} dr_{\gamma\lambda} \cdot$$
$$\cdot \exp N \left\{ -\beta \frac{a\left(1-\tilde{a}\right)}{2} \sum_{\nu\gamma} \left(m_\nu^\gamma\right)^2 - \frac{\alpha}{2} \mathrm{Tr} \ln\left[a\left(1-\tilde{a}\right)\left(1-\beta\tilde{a}\mathbf{q}\right)\right] - \right. \tag{92}$$
$$\left. \frac{\alpha\beta^2\tilde{a}^2}{2} \sum_{\gamma\lambda} r_{\gamma\lambda} q_{\gamma\lambda} - \beta\tilde{a} \left[\frac{\alpha}{2} + S\tilde{U} \right] \sum_{\gamma\gamma} q_{\gamma\gamma} + \left\langle \ln \mathrm{Tr}_{\{\sigma^\gamma\}} \exp\left[\beta \mathcal{H}_\sigma^\xi\right] \right\rangle_{\xi^v} \right\},$$

where

$$\mathcal{H}_{\sigma}^{\xi} = \sum_{\nu\gamma} m_{\nu}^{\gamma} v_{\xi^{\nu}\sigma^{\gamma}} + \frac{\alpha\beta}{2S\left(1-\tilde{a}\right)} \sum_{\gamma\lambda} r_{\gamma\lambda} \sum_{k} P_k v_{k\sigma^{\gamma}} v_{k\sigma^{\lambda}} \,. \tag{93}$$

We can now compute the free energy Eq. (27)

$$\begin{aligned} f &= \lim_{n\to 0} f_n = \lim_{n\to 0} \left\{ \frac{a\left(1-\tilde{a}\right)}{2n} \sum_{\nu\gamma} \left(m_{\nu}^{\gamma}\right)^2 + \right. \\ &+ \frac{\alpha}{2n\beta} \mathrm{Tr} \ln \left[a\left(1-\tilde{a}\right)\left(1-\beta\tilde{a}\mathbf{q}\right)\right] + \frac{\alpha\beta\tilde{a}^2}{2n} \sum_{\gamma\lambda} r_{\gamma\lambda} q_{\gamma\lambda} \\ &+ \left. \frac{\tilde{a}}{n}\left[\frac{\alpha}{2} + S\tilde{U}\right] \sum_{\gamma\gamma} q_{\gamma\gamma} - \frac{1}{n\beta} \left\langle \ln \mathrm{Tr}_{\{\sigma^{\gamma}\}} \exp\left[\beta\mathcal{H}_{\xi}\right] \right\rangle_{\xi^v} \right\}. \end{aligned} \tag{94}$$

Imposing the replica symmetry condition [40],

$$\begin{aligned} m_{\gamma}^{\nu} &= m \\ q_{\gamma\lambda} &= \begin{cases} q & \text{for} \quad \gamma \neq \lambda \\ \tilde{q} & \text{for} \quad \gamma = \lambda \end{cases} \\ r_{\gamma\lambda} &= \begin{cases} r & \text{for} \quad \gamma \neq \lambda \\ \tilde{r} & \text{for} \quad \gamma = \lambda \end{cases} \end{aligned}$$

we finally obtain the replica symmetric free energy Eq. (28).

B Self consistent signal to noise analysis

Since the l.h.s. of Eq. (57) includes $p-1 \gg 1$ terms, the ansatz is still valid also when singling out one of these many contributions, so that we can equivalently write it as

$$\sum_{\nu>1} v_{\xi_i^\nu,k} m_i^\nu = v_{\xi_i^\mu,k} m_i^\mu + \sum_{\nu\neq 1,\mu} v_{\xi_i^\nu,k} m_i^\nu = v_{\xi_i^\mu,k} m_i^\mu + \gamma_i^k \langle \sigma_i^k \rangle + \sum_n v_{n,k}\, \rho_i^n z_i^n \ , \qquad (95)$$

where γ_i^k and ρ_i^n are independent of μ. The contribution from the non-condensed pattern $\mu \neq 1$ is assumed to be small, so that we can expand G_i^k to first order in $v_{\xi_i^\mu,k} m_i^\mu$:

$$\begin{aligned} \sigma_j^l &= G^l\left[\left\{ v_{\xi_j^1,k} m_j^1 + \sum_n v_{n,k} \rho_j^n z_j^n - U(1-\delta_{k,0}) \right\}_{k=0}^{S}\right] \\ &\quad + \sum_n v_{\xi_j^\mu,n}\, m_j^\mu \, \frac{\partial G^l}{\partial y^n}\left[\left\{ v_{\xi_j^1,k} m_i^1 + \sum_n v_{n,k} \rho_j^n z_j^n - U(1-\delta_{k,0}) \right\}\right] . \qquad (96) \end{aligned}$$

Reinserting the expansion into the r.h.s of Eq. (52) we recognize a relation of the form

$$m_i^\mu = L_i^\mu + \sum_j K_{ij}^\mu m_j^\mu \qquad (97)$$

where

$$\begin{aligned} K_{ij}^\mu &\equiv \frac{1}{c_m a(1-\tilde{a})} \sum_{l,n} c_{ij} v_{\xi_j^\mu,l} v_{\xi_j^\mu,n} \frac{\partial G_j^l}{\partial y^n} \ , \\ L_i^\mu &\equiv \frac{1}{c_m a(1-\tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^\mu,l} G_j^l \ . \end{aligned}$$

The overlap m_i^μ can be found by iterating Eq. (97),

$$m_i^\mu = L_i^\mu + \sum_{j_1} L_{j_1}^\mu \left\{ K_{ij_1}^\mu + \sum_{j_2} K_{ij_2}^\mu K_{j_2 j_1}^\mu + \sum_{j_2}\sum_{j_3} K_{ij_2}^\mu K_{j_2 j_3}^\mu K_{j_3 j_1}^\mu + ... \right\} . \qquad (98)$$

Therefore, the noise term can be written explicitly as

$$\begin{aligned} \sum_{\mu>1} v_{\xi_i^\mu,k} m_i^\mu &= \sum_n v_{n,k} \sum_{\mu>1} \left\{ \sum_j \sum_l \frac{1}{c_m a(1-\tilde{a})}\, c_{ij} \delta_{\xi_i^\mu,n} v_{\xi_j^\mu,l} G_j^l + \right. \\ + \quad & \left. \sum_{j_1}\sum_j\sum_l \frac{1}{c_m a(1-\tilde{a})}\, c_{j_1 j} \delta_{\xi_i^\mu,n} v_{\xi_j^\mu,l} G_j^l \left(\sum_{l_1,n_1} \frac{1}{c_m a(1-\tilde{a})}\, c_{ij_1} v_{\xi_{j_1}^\mu,l_1} v_{\xi_{j_1}^\mu,n_1} \frac{\partial G_{j_1}^{l_1}}{\partial y^{n_1}} + ... \right) \right\} . \end{aligned}$$

In order to obtain the expression for γ_i^k, in Eq. (57) we consider only the terms with $j = i$ and $l = k$, and take the average over the connectivity and the patterns:

$$\begin{aligned} \gamma_i^k &= \frac{\alpha}{S}\lambda\left\langle \frac{1}{S}\frac{1}{N}\sum_{j_1}\sum_{l_1}\frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} + ... \right\rangle \\ &= \frac{\alpha}{S}\lambda\left\{ \Omega/S + (\Omega/S)^2 + ... \right\} \\ &= \frac{\alpha}{S}\lambda\frac{\Omega/S}{1-\Omega/S} \end{aligned} \tag{99}$$

where we use the fact that $c_{ii} = 0$, $\alpha = p/c_m$, $\langle\cdot\rangle$ indicates the average over all patterns and where we have defined

$$\Omega = \left\langle \frac{1}{N}\sum_{j_1}\sum_{l_1}\frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} \right\rangle . \tag{100}$$

By virtue of the statistical independence of units, the average over the non-condensed patterns for the $i \neq j$ terms vanishes. From the variance of the noise term one reads

$$(\rho_i^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q\left\{ 1 + 2\lambda\Psi + \lambda\Psi^2 \right\}, \tag{101}$$

where

$$q = \left\langle \frac{1}{Na}\sum_{j,l}(G_j^l)^2 \right\rangle \tag{102}$$

and

$$\Psi = \frac{\Omega/S}{1-\Omega/S} \ . \tag{103}$$

The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k}m + \frac{\alpha}{S}\lambda\Psi(1-\delta_{k,0}) + \sum_n v_{n,k}z^n\sqrt{\frac{\alpha P_n}{S(1-\tilde{a})}q\left\{1 + 2\lambda\Psi + \lambda\Psi^2\right\}} - \tilde{U}(1-\delta_{k,0}) . \tag{104}$$

References

[1] Leonid Schneider. Human Brain Project: bureaucratic success despite scientific failure, https://forbetterscience.com/2017/02/22/human-brain-project-bureaucratic-success-despite-scientific-failure/.

[2] Valentino Braitenberg and Almut Schüz. *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Science & Business Media, 2013.

[3] Valentino Braitenberg. Thoughts on the cerebral cortex. *Journal of theoretical biology*, 46(2):421–447, 1974.

[4] Valentino Braitenberg. Cortical architectonics: general and areal. mab brazier and h. petsche (eds), architectonics of the cerebral cortex (443–465), 1978.

[5] Vernon B Mountcastle. The columnar organization of the neocortex. *Brain: a journal of neurology*, 120(4):701–722, 1997.

[6] Pasko Rakic. Confusing cortical columns. *Proceedings of the National Academy of Sciences*, 105(34):12099–12100, 2008.

[7] Jon H Kaas. Evolution of columns, modules, and domains in the neocortex of primates. *Proceedings of the National Academy of Sciences*, 109(Supplement 1):10655–10660, 2012.

[8] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[9] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[10] Alessandro Treves. Graded-response neurons and information encodings in autoassociative memories. *Physical Review A*, 42(4):2418, 1990.

[11] Daniel J Amit and Nicolas Brunel. Learning internal representations in an attractor neural network with analogue neurons. *Network: Computation in Neural Systems*, 6(3):359–388, 1995.

[12] Bernhard Hellwig. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological cybernetics*, 82(2):111–121, 2000.

[13] Nir Kalisman, Gilad Silberberg, and Henry Markram. The neocortical microcircuit as a tabula rasa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):880–885, 2005.

[14] Yasser Roudi and Alessandro Treves. An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(07):P07010, 2004.

[15] Alexis M Dubreuil and Nicolas Brunel. Storing structured sparse memories in a multi-modular cortical network model. *Journal of computational neuroscience*, 40(2):157–175, 2016.

[16] Christopher Johansson and Anders Lansner. Imposing Biological Constraints onto an Abstract Neocortical Attractor Network Model. *Neural Computation*, 19: 1871–1896, 2007.

[17] Mikael Lundqvist, Albert Compte and Anders Lansner. Bistable, Irregular Firing and Population Oscillations in a Modular Attractor Memory Network. *PLoS Computational Biology*, 6(6): e1000803, 2010.

[18] Dominic O'Kane and Alessandro Treves. Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General*, 25(19):5055, 1992.

[19] Dominic O'kane and Alessandro Treves. Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Computation in Neural Systems*, 3(4):379–384, 1992.

[20] R Lauro-Grotto, S Reich, and Miguel A Virasoro. The computational role of conscious processing in a model of semantic memory. 1997.

[21] Carlo Fulvi Mari and Alessandro Treves. Modeling neocortical areas with a modular neural network. *Biosystems*, 48(1):47–55, 1998.

[22] Nir Levy, David Horn, and Eytan Ruppin. Associative memory in a multimodular network. *Neural Computation*, 11(7):1717–1737, 1999.

[23] Carlo Fulvi Mari. Extremely dilute modular neuronal networks: Neocortical memory retrieval dynamics. *Journal of Computational Neuroscience*, 17(1):57–79, 2004.

[24] Ido Kanter. Potts-glass models of neural networks. *Physical Review A*, 37(7):2739, 1988.

[25] Désiré Bollé, Patrick Dupont, and Jort van Mourik. Stability properties of Potts neural networks with biased patterns and low loading. *Journal of Physics A: Mathematical and General*, 24(5):1065, 1991.

[26] Désiré Bollé, Patrick Dupont, and J Huyghebaert. Thermodynamic properties of the q-state Potts-glass neural network. *Physical Review A*, 45(6):4194, 1992.

[27] Désiré Bollé, Roland Cools, Patrick Dupont, and J Huyghebaert. Mean-field theory for the q-state Potts-glass neural network with biased patterns. *Journal of Physics A: Mathematical and General*, 26(3):549, 1993.

[28] Désiré Bollé, B Vinck, and VA Zagrebnov. On the parallel dynamics of the q-state Potts and q-Ising neural networks. *Journal of statistical physics*, 70(5):1099–1119, 1993.

[29] Emilio Kropff and Alessandro Treves. The storage capacity of Potts models for semantic memory retrieval. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(08):P08010, 2005.

[30] Alessandro Treves. Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive Neuropsychology*, 22(3-4):276–291, 2005.

[31] Eleonora Russo and Alessandro Treves. Cortical free-association dynamics: Distinct phases of a latching network. *Physical Review E*, 85(5):051920, 2012.

[32] Renfrey Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 5(5):965, 1975.

[33] Chol Jun Kang, Michelangelo Naim, Vezha Boboeva, and Alessandro Treves. Life on the edge: Latching dynamics in a Potts neural network. *Entropy*, 19(9):468, 2017.

[34] Alessandro Treves and Edmund T Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, 1991.

[35] MV Tsodyks and MV Feigel'Man. The enhanced storage capacity in neural networks with low activity level. *EPL (Europhysics Letters)*, 6(2):101, 1988.

[36] Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge University Press, 1992.

[37] Alessandro Treves. Mean-field analysis of neuronal spike dynamics. *Network: Computation in Neural Systems*, 4(3):259–284, 1993.

[38] Francesco P Battaglia and Alessandro Treves. Stable and rapid recurrent processing in realistic autoassociative memories. *Neural Computation*, 10(2):431–450, 1998.

[39] Emilio Kropff and Alessandro Treves. The complexity of latching transitions in large scale cortical networks. *Natural Computing*, 6(2):169–185, 2007.

[40] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[41] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Co Inc, 1987.

[42] Tamás Geszti. *Physical models of neural networks*. World Scientific, 1990.

[43] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[44] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of physics*, 173(1):30–67, 1987.

[45] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[46] Andreas Engel, Rémi Monasson, and Alexander K Hartmann. On large deviation properties of erdös–rényi random graphs. *Journal of Statistical Physics*, 117(3-4):387–426, 2004.

[47] Bernard Derrida, Elizabeth Gardner, and Anne Zippelius. An exactly solvable asymmetric neural network model. *EPL (Europhysics Letters)*, 4(2):167, 1987.

[48] Masatoshi Shiino and Tomoki Fukai. Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Physical Review E*, 48(2):867, 1993.

[49] Emilio Kropff. Full solution for the storage of correlated memories in an autoassociative memory. *Computational Modelling in Behavioural Neuroscience: Closing the Gap Between Neurophysiology and Behaviour*, 2:225, 2009.

[50] Alessandro Treves and Edmund T Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, 1991.

[51] Haim Sompolinsky. Neural networks with nonlinear synapses and a static noise. *Physical Review A*, 34(3):2571–2574, 1986.

[52] Valentino Braitenberg. Cell assemblies in the cerebral cortex. In *Theoretical approaches to complex systems*, pages 171–188. Springer, 1978.

[53] Alexander A Frolov and Dusan Husek and Igor P Muraviev. Informational capacity and recall quality in sparsely encoded Hopfield-like neural network: analytical approaches and computer simulation. *Neural Networks*, 10(5):845–855, 1997.

[54] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[55] Guy Elston. Pyramidal cells of the frontal lobe: all the more spinous to think with. *Journal of Neuroscience* 20:1-4, 2000.